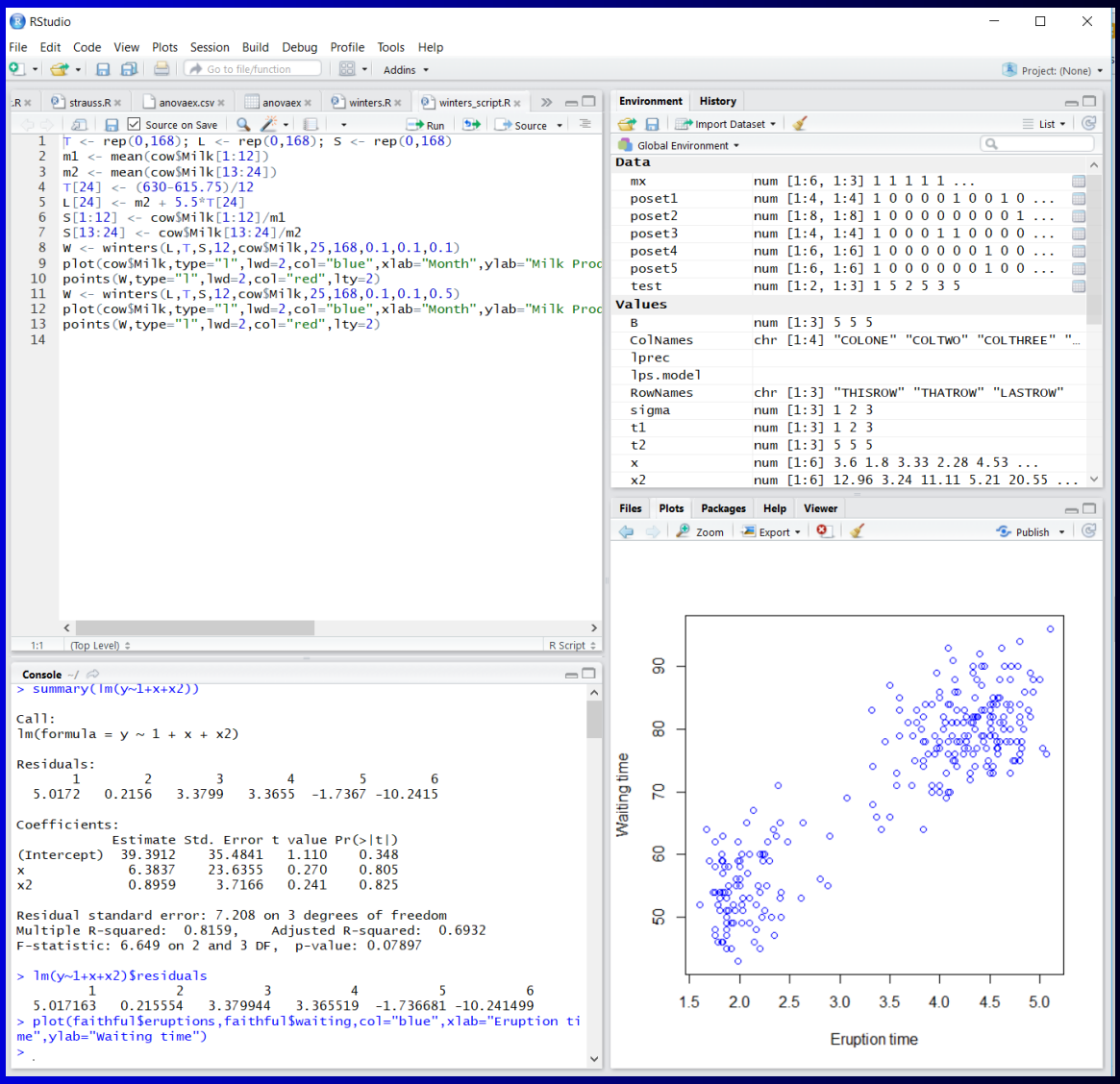
Statistical Inference
Theory and Labs



Mark Huber

About this book

Purpose This book was created to teach a one semester undergraduate course in the theory of statistics for students who have already completed a one semester course in calculus based probability. The course is designed to be 2/3 traditional lecture, and 1/3 inquiry based learning, where students complete labs primarily on their own to learn the practical side of analyzing data with statistical tools.

Organization Part I of the text reviews ideas that are covered in the prerequisites of the course. In part II, the new ideas about statistics and the theory behind it are given. In part III, hands on lab exercises are presented that teach a student how to conduct statistical analyses in R. Finally, Part IV gives solutions to many of the exercises at the end of each chapter.

My approach When I teach the course, I leave everything in Part I to be read as needed by students, and start lecturing immediately with Part II of the text. But whether you start with Chapter 1 or a later chapter, Part I should serve as a valuable reference for students as they delve into Part II.

The course as I teach it has three meeting sessions per week. I alternate between lectures on Monday and Friday, followed by a lab where students learn to actually ran statistical analyses on data sets on Wednesday. These lab exercises are collected in Part III of the text, and are implemented using R.

The classroom I teach in has a computer available for every student, but most students prefer to bring their own laptop. The labs (except the first) are structured as a main lab followed by an extended lab. Students who finish the main lab before the class session is up are required in my course to then complete the extended lab, as the time to finish varies considerably betweeen students based on their familiarity with computers.

The book covers both frequentist and Bayesian, parametric and nonparametric methods whenever possible.

Why are all the numerical answers in problems and examples given to 4 significant digits? In my homework assignments to students I require that all noninteger answers be presented to four significant digits. There are several reasons why I do this.

The first is that it makes answers uniform. I do not have to worry if $1/(3+\sqrt{2}) = \frac{3-\sqrt{2}}{5}$ or not if the answer given is 0.2265. The second is that it emphasizes to students that in most cases data is uncertain leading to uncertainty of results. The number 1/3 is specific and exact, but not actually encountered outside of toy problems. Third, it builds numerical literacy (or numeracy as it is sometimes called.) Seeing that $\exp(-2) \approx 13.53\%$ is a useful thing, as it gives that much desired reality check on the answers provided.

Contents

Part I

What you need to know before learning statistics

Chapter 1

Notation

Question of the Day What is the mathematics that I need to know before starting a statistics course?

Statistics is a set of concepts, ideas, and methods that uses Mathematics extensively. In that respect, it has a relationship to Mathematics that is similar to the relationship betwween Physics and Mathematics. Physics has its own concepts such as Mass and Energy that do not arise from Mathematics, but Physics uses Mathematics a lot, using differential equations and partial differential equations to model how mass and energy interact. In a similar fashion, Statistics utilizes probability to model how data is collected, and that allows us to prove theorems about the properties of Statistical estimators, but at the end of the day Statistical concepts lie outside of mathematics.

So like Physics, to be a Statistician requires knowledge of several areas of mathematics. In particular, you need to know:

- Various important mathematical notation. This includes summation notation, product notation, and indicator function notation.
- Random variables.
- Probability distributions such as Bernoulli, Binomial, Geometric, Uniform, Beta, Gamma, and Normal (Gaussian).
- How to work with probability densities, especially joint densities of random variables.
- The rules of logarithms.
- How to optimize functions, and the different between finding $\max f(x)$ and $\arg\max f(x)$.

The descriptions and facts in this part of the text are meant to be review that assumes you have had a calculus based Probability courses. Hence these chapters do not contain much in the way of proof or examples. If an idea is these first few chapters has not been seen before, please talk to your instructor to get more resources and information.

1.1. Summation notation

Often in statistics we are adding or multiplying several numbers together, and it is useful to have notation that deals with that. First, we have *summation notation*, which uses a capital Greek letter Sigma ($\sum$) since "Sigma" and "sum" both start with the letter S.

Notation 1 (Summation notation)

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n.$$

For example, if $(x_1, x_2, x_3) = (2.3, 1.1, -2.0)$, then $\sum_{i=1}^{3} = 2.3 + 1.1 - 2.0 = 1.4$.

1.2. Product notation

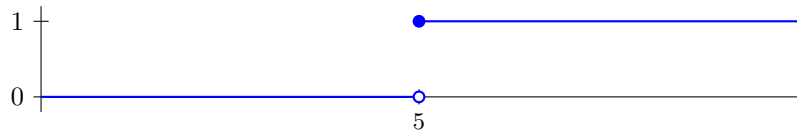
Product notation is similar. Since the word "Product" starts with P, we use a capital Greek letter pi ($\prod$) to represent the product of the numbers.

Notation 2 (Product notation)

$$\prod_{i=1}^{n} x_i = x_1 x_2 \cdots x_n.$$

1.3. Indicator functions

The indicator function $\mathbb{1}$ is a very useful function that takes an argument that is either true or false. If the argument is true, then it returns a 1, but if the argument is false, it returns a 0. For example, let $f(x) = \mathbb{1}(x > 5)$. Then $f(6) = \mathbb{1}(6 > 5) = 1$, while $f(4.5) = \mathbb{1}(4.5 > 5) = 0$. If we graph $f(x)$, it looks like this:



Definition 1
The **indicator function** $\mathbb{1} : \{\text{TRUE}, \text{FALSE}\} \to \{0, 1\}$ is defined as

$$\mathbb{1}(\text{TRUE}) = 1, \ \mathbb{1}(\text{FALSE}) = 0.$$

Since the indicator function always integrates to 0 whenever the argument is false, it can be used to change the limits of integration within a function, a useful trick to know.

Fact 1
For a real valued function f and $a < b$,

$$\int_a^b f(t)\, dt = \int_{-\infty}^{\infty} f(t)\mathbb{1}(t \in [a, b])\, dt.$$

1.4. Circle constants

There are two circle constants currently in widespread use. The constant $\pi = 3.141\ldots$ represents the arclength of the upper half of a circle of radius 1, and will be referred to here as the *half-circle constant*. This is 180 degrees.

The constant $\tau = 6.283\ldots$ is the arclength of the entire circle of radius 1, and will be referred to here as the *full-circle constant*. This is 360 degrees.

They are related as $\pi = \tau/2$ (since π represents half of a circle) or equivalently $\tau = 2\pi$.

Why prefer τ over π? There are many reasons, but the simplest is that if I am interested in the angle spanning 1/4 of the circle, that is $\tau/4$. If I am interested in the angle spanning 1/3 of the circle, that is $\tau/3$. There is no extra conversion factor of 1/2 that comes into play, and so less of a need to memorize the angles of the unit circle.

1.5. Significant Digits

You will notice that most of the answers to problems in the book are given to four significant digits. Significant digits are a shorthand way of indicating how much precision is in an answer. For instance, if I say $\tau = 6$, that is one significant digit (or sig fig for short), $\tau = 6.2$ is two sig figs, $\tau = 6.383$ is four sig figs, and so on.

Note that 6.383 is given to three decimal places of accuracy, but has four sig figs. The only time the number of sig figs equals the number of decimal places is when the first sig fig appears right after the decimal point (for example 0.3007.)

Presenting an answer to four sig figs means that it is relatively accurate to about 0.5%. Therefore, usually no rounding is necessary. The exception is when dealing with upper and lower bounds. For instance, if $x < \exp(-1)$, then it would be correct to say (given $\exp(-1) = 0.36787944117$) that $x < 0.3679$) to 4 sig figs.

Similarly, if $y > \exp(-1)$, then it would be correct to say $y > 0.3678$ to 4 sig figs. If $z \in [1/3, 2/3]$, then to 4 sig figs, $z \in [0.3333, 0.6667]$. But outside of lower and upper bounds, rounding after 4 sig figs is typically not necessary.

Problems

1.1: Go to the website `www.wolframalpha.com` and type in

```
sum(1/2)^i for i from 1 to infinity
```

What is $\sum_{i=1}^{\infty}(1/2)^i$?

1.2: Graph $f(s) = \mathbb{1}(s \geq 0)$

1.3: Solve $\int_{-\infty}^{\infty} 2s\mathbb{1}(s \in [0,1])\, ds$

1.4: What is $\sqrt{\tau}$?

Chapter 2

Probability

Question of the Day How do I represent partial information about something?

Probability is the mathematics of partial information.

2.1. Random variables

Variables are mathematical symbols that stand for something where the true value is completely unknown. For instance, we might say $x \in \mathbb{R}$ to indicate that x is any unknown real number, while we might write $i \in \{1, 2, 3\}$ to indicate that i is either 1, 2, or 3.

Random variables are similar to variables, but we have partial information about their value. Often (but not always) capital letters are used to denote random variables. The partial information is given through the use of *probabilities*, usually denoted using $\mathbb{P}$.

For example, we might say that the random variable I is in the set $\{1, 2, 3\}$, or more compactly $I \in \{1, 2, 3\}$. But we also have additional information about how likely each of the three values is. We might know that $\mathbb{P}(I = 1) = 0.8$, $\mathbb{P}(I = 2) = \mathbb{P}(I = 3) = 0.1$. So while we do not know the value of I exactly, we know much more about it that we do about the variable i. For instance, we can say that there are 4 to 1 odds that $I = 1$ versus $I \neq 1$.

For simple sets like $\{1, 2, 3\}$, every subset we can assign a probability to that makes sense, but this becomes difficult as we start to deal with larger sets like the real numbers. So we only worry about finding $\mathbb{P}(I \in A)$ for some sets A. The sets we care about we call *measurable*.

Intuition 1
For a random variable X and set A, if it is possible to measure the probability that X falls into A, we call the set A **measurable**.

For instance, if $\mathbb{P}(X \in \{1, 3, 7\}) = 0.3$, then $\{1, 3, 7\}$ is a measurable set. A nice fact is that all countable sets are measurable.

Definition 2
A set of random variables $X_1, \ldots, X_n$ are **independent** if for all measure $A_1, \ldots, A_n$,

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i).$$

That definition can be extended to an arbitrary size collection of random variables.

Definition 3
A collection of random variables $\{X_\alpha\}$ for $\alpha \in \Omega$ is **independent** if for all finite subsets $\{\alpha_1, \ldots, \alpha_n\} \subseteq \Omega$, $X_{\alpha_1}, \ldots, X_{\alpha_n}$ are independent. The $\{X_\alpha\}$ are **identically distributed** if for all α_1 and α_2 in Ω, and measurable A, $\mathbb{P}(X_{\alpha_1} \in A) = \mathbb{P}(X_{\alpha_2} \in A)$. A collection that is both independent and identically distributed is **iid**.

Typically we do not use this definition to prove that a collection of random variables are iid, instead we just assume that the variables are iid unless there is a special reason to believe otherwise.

2.2. Densities

For random variables over large (or infinite) state spaces, it is too cumbersome to write down all the probabilities exactly. So instead, we write a function that encodes all the probabilities. For this course, we consider two types of function.

Definition 4
If there is a finite or countably infinite set $\{x_1, x_2, \ldots\}$ such that $\mathbb{P}(X \in \{x_1, x_2, \ldots\}) = 1$, call X a **discrete** random variable, and let $f_X(x_i) = \mathbb{P}(X = x_i)$ be the **density** (aka **probability density function** aka **pdf** of X with respect to counting measure.

To find the probability that a *discrete* random variable X is in some finite set A, we sum the density of X over the elements of A.

Fact 2
If X has density f_X with respect to counting measure, then

$$\mathbb{P}(X \in A) = \sum_{a \in A} f_X(a).$$

Two measures are of particular importance, counting measure and Lebesgue measure.

Definition 5
Counting measure is the measure that returns the number of elements of a set.

For example, the counting measure of $\{1, 3, 7\}$ is 3, while the counting measure of $\emptyset$ is 0.

Counting measure is associated with discrete random variables. On the other hand, if X is a *continuous* random variable, we integrate the density of X over the set A using Lebesgue measure.

Definition 6
Say that X is a **continuous** random variable if there exists a **density** (aka **probability density function** aka **pdf**) f_X such that for all (measurable) A,

$$\mathbb{P}(X \in A) = \int_{x \in A} f_X(x)\, dx.$$

Intuition 2
Lebesgue measure is length in one dimension, the area of a set in two dimensions, volume in three, and so on.

For example, the Lebesgue measure of $[3, 10]$ is 7, while the Lebesgue measure of the interior of the triangle connecting $(0, 0)$, $(4, 0)$ and $(5, 0)$ is 10.

If ν is Lebesgue measure, then

$$\int_{a \in A} f_X(a)\, d\nu = \int_{a \in A} f_X(a)\, da,$$

and if ν is counting measure, then

$$\int_{a\in A} f_X(a)\ d\nu = \sum_{a\in A} f_X(a).$$

Therefore, we can use say

Fact 3
If X has density f_X with respect to measure ν, then

$$\mathbb{P}(X \in A) = \int_{a\in A} f_X(a)\ d\nu.$$

If X is a discrete random variable then ν is counting measure, and if X is a continuous random variable then ν is Lebesgue measure.

Now suppose that X and Y are two random variables. Then they have a joint density $f_{(X,Y)}$ if integration can also be used to find the probabilities associated with X and Y.

Definition 7
Random variables X and Y have joint density $f_{(X,Y)}$ with respect to ν if for all (measurable) A and B,

$$\mathbb{P}(X \in A, Y \in B) = \int_{(x,y)\in A\times B} f_{(X,Y)}(x,y)\ d\nu.$$

An important fact about joint densities is that for independent random variables, the joint density is the product of the individual densities.

Fact 4
Let $X_1, \ldots, X_n$ have densitys f_{X_1} through f_{X_n} with respect to a common measure ν. Then

$$f_{(X_1,\ldots,X_N)}(x_1,\ldots,x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n).$$

2.3. Mean of a random variable

Definition 8
The **expected value**, **mean**, **average**, or **expectation** of a random variable $X \in \mathbb{R}$ with density f_X with respect to measure ν is

$$\mathbb{E}[X] = \int_{a\in\mathbb{R}} a f_X(a)\ d\nu$$

when this integral exists and is finite. When this happens, say that X is **integrable**.

The first important fact about the mean operator is that it is *linear*.

Fact 5
Let a and b be real numbers, and X and Y be integrable random variables (that might be dependent or independent). Then

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

The second important fact

This can be extended to functions of a random variable X.

Fact 6
The mean of $g(X)$ is

$$\mathbb{E}[g(X)] = \int_{a\in\mathbb{R}} g(a) f_X(a)\ d\nu.$$

if that integral is finite.

This is often used is calculating the variance of a random variable.

Definition 9
The **variance** of an integrable random variable X is $\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$. If $\mathbb{V}(X) = \infty$, then we can say that the random variable has infinite variance, or that the variance does not exist.

For example, say $\mathbb{P}(X = 1) = 0.3$, $\mathbb{P}(X = 2) = 0.3$, $\mathbb{P}(X = 3) = 0.4$. Then

$$\mathbb{E}[X] = 0.3(1) + 0.3(2) + 0.4(3) = 2.1,$$

and

$$\mathbb{V}[X] = 0.3(1 - 2.1)^2 + 0.3(2 - 2.1)^2 + 0.4(3 - 2.1)^2 = 0.69.$$

Definition 10
For a random variable with finite variance, the **standard deviation** of the random variable is the nonnegative square root of the variance.

Fact 7
If X is a random variable with finite standard deviation $\mathrm{SD}(X)$, for all $c \in \mathbb{R}$, $\mathrm{SD}(cX) = |c|\,\mathrm{SD}(X)$ and $\mathbb{V}(cX) = c^2\mathbb{V}(X)$.

Definition 11
If $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y)$, say that X and Y are **uncorrelated**.

Fact 8
Independent random variables are uncorrelated.

Another useful formula for the variance is the following.

Fact 9
If X has finite variance, then

$$\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Following our earlier example,

$$\mathbb{V}(X) = 0.3(1)^2 + 0.3(2)^2 + 0.4(3)^2 - 2.1^2 = 0.69.$$

Problems

2.1: Let X have density $f_X(1) = 0.2$, $f_X(5) = 0.7$, and $f_X(6) = 0.1$.

(a) What is $\mathbb{P}(X = 5)$?
(b) What is $\mathbb{P}(X = 2)$?
(c) What is $\mathbb{E}[X]$?
(d) What is $\mathbb{V}(X)$?

2.2: Let X have density $f_X(i) = (2/3)^{i-1}(1/3)\mathbb{1}(i \in \{1, 2, \ldots\})$ with respect to counting measure.

(a) Find $\mathbb{P}(X \in \{1, 2, 3\})$.

(b) Find $\mathbb{E}(X)$.

2.3: Let T have density $f_T(s) = 2\exp(-2s)\mathbb{1}(s \geq 0)$.

(a) Find $\mathbb{P}(X \in [1, 3])$.

(b) Find $\mathbb{E}[X]$.

Chapter 3

Distributions

Question of the Day How do I calculate mean and variance for common random variables without going back to the original formulas?

We have shorthand that describes the probabilities associated with random variables, and that is the *distribution* of a random variable.

Definition 12
The **distribution** $\mathbb{P}_X$ of a random variable X is a function of measurable sets defined as

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A).$$

Notation 3
If X and Y have the same distribution, write

$$X \sim Y.$$

It can be cumbersome to write out $\mathbb{P}_X(A)$ for all measurable sets A. It turns out that for real valued random variables it suffices to know the distribution for sets A of the form $(-\infty, a]$.

Definition 13
The **cumulative distribution function** or **cdf** of a random variable X is

$$F_X(a) = \mathbb{P}_X((-\infty, a]) = \mathbb{P}(X \leq a).$$

Fact 10
If $F_X = F_Y$, then $X \sim Y$.

3.1. Names of distributions

The most common probability distributions all have names.

Uniform Say that $X \sim \mathsf{Unif}(A)$ if for all measurable $B \subset A$,

$$\mathbb{P}(X \in B) = \frac{\nu(A)}{\nu(B)}.$$

If ν is counting measure, then X has density with respect to counting measure

$$f_X(a) = \frac{1}{\#(B)} \cdot \mathbb{1}a \in B.$$

If ν is Lebesgue measure, then X has density with respect to Lebesgue measure

$$f_X(a) = \frac{1}{\nu(B)} \cdot \mathbb{1}a \in B.$$

and in the specific case that A is a one dimensional interval $[a, b]$, then

$$f_X(s) = \frac{1}{b-a} \cdot \mathbb{1}(s \in [a, b]).$$

Bernoulli The simplest nontrivial random variable takes on two values, 0 or 1. We can these *Bernoulli* random variables, and they have exactly one parameter: the probability that the random variable is 1. Write $X \sim \mathsf{Bern}(p)$ if $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$. A Bernoulli random variable has density with respect to counting measure

$$f_X(i) = p\mathbb{1}(i = 1) + (1 - p)\mathbb{1}(i = 0).$$

Another way to describe this distribution, is that X is Bernoulli with probability p if it counts the number of successes on a single experiment that is a success with probability p and a failure otherwise.

Binomial Let $X_1, \ldots, X_n$ be Bernoulli random variables that are independent and all have parameter p. Then $S = X_1 + \cdots + X_n$ is a Binomial random variable with parameters n and p. Write $S \sim \mathsf{Bin}(n, p)$. The density with respect to counting measure is

$$f_S(i) = \binom{n}{i} p^i (1 - p)^{n-i} \mathbb{1}(i \in \{0, 1, \ldots, n\}).$$

Another way to describe this distribution, is that S is Binomial with parameters n and p if it counts the number of successes on n independent experiments that each attain success with probability p.

Geometric Let $X_1, X_2, \ldots$ be an iid sequence of Bernoulli random variables with parameter p. Let $G = \inf\{t : X_t = 1\}$. Then we say G is a Geometric random variable with parameter p. The density of G with respect to counting measure is

$$f_G(i) = (1 - p)^{n-i} p\mathbb{1}(i \in \{1, 2, \ldots\}).$$

Negative Binomial Let $G_1, G_2, \ldots, G_r$ be an iid sequence of Geometric random variables with parameter p. Then $R = G_1 + G_2 + \cdots + G_r$ has a Negative Binomial distribution with parameters r and p. The density of R with respect to counting measure is

$$f_R(i) = \binom{i-1}{r-1} p^r (1 - p)^{i-r} \mathbb{1}(\{i \in \{r, r + 1, \ldots\})$$

The continuous analogue to the geometric random variable is the exponential, and the negative binomial distribution is the gamma distribution.

Exponential Say that X has an exponential distribution with rate parameter λ if it has density with respect to Lebesgue measure of

$$f_X(s) = \lambda \exp(-\lambda s)\mathbb{1}(s \geq 0).$$

Gamma Say that X has a gamma distribution with shape parameter n and rate parameter λ if it has density with repsect to Lebesgue measure of

$$f_X(s) = \lambda^n s^{n-1} \exp(-\lambda s)\mathbb{1}(s \geq 0)/\Gamma(n).$$

Here $\Gamma(n)$ is called the *Gamma function*, and when n is an integer, $\Gamma(n) = (n - 1)!$.

Poisson Let $A_1, A_2, \ldots$ be iid $\mathsf{Exp}(1)$, and $N = \#\{n : A_1 + \cdots + A_n \leq \mu\}$. Then we say N has a Poisson distribution with parameter μ. The density of N with respect to counting measure is

$$f_N(i) = \exp(-\mu)\frac{\mu^i}{i!}\mathbb{1}(i \in \{0, 1, 2, \ldots\}).$$

Beta This distribution comes from several different sources, including ratios of gammas and order statistics of uniforms over $[0,1]$. Say that $X \sim \mathsf{Beta}(a,b)$ if X has density with respect to Lebesgue measure of

$$f_X(s) = s^{a-1}(1-s)^{b-1}\mathbb{1}(s \in [0,1])/B(a,b),$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the Beta function.

The Central Limit Theorem leads to the importance of the normal distribution (aka the Gaussian aka the bell-shaped curve.)

Normal Say that Z has a standard normal distribution (write $Z \sim \mathsf{N}(0,1)$) if it has density with respect to Lebesgue measure

$$f_Z(x) = \frac{1}{\sqrt{\tau}} \exp(-x^2/2).$$

Write $\mu + \sigma Z \sim \mathsf{N}(\mu, \sigma^2)$, which has density

$$f_{\mu+\sigma Z}(x) = \frac{1}{\sqrt{\tau}\sigma} \exp(-(x-\mu)^2/(2\sigma^2)).$$

A canonical example of a heavy-tailed distribution is the Cauchy defined as follows.

Cauchy Say that X is a standard Cauchy if it has density

$$f_X(s) = \frac{2}{\tau} \cdot \frac{1}{1+s^2}.$$

Then $\mu + \sigma X$ is a Cauchy distribution with location parameter μ and scale parameter σ. Write $\mu + \sigma X \sim \mathsf{Cauchy}(\mu, \sigma^2)$. This has density (with respect to Lebesgue measure) of

$$f_{\mu+\sigma X}(s) = \frac{2}{\tau\sigma} \cdot \frac{1}{1+(s-\mu)^2/\sigma^2}.$$

This list is not meant to be exhaustive, in particular later in the course we will study the chi-squared, the t and the F distributions.

3.2. Means and Variances

One of the advantages of using named distributions with parameters is that we can compute the mean and variances once, and from then on just use formulas.

Fact 11
The named distributions have the following means and variances.

Distribution	Mean	Variance
$\mathsf{Unif}([a,b])$	$(b+a)/2$	$(b-a)^2/12$
$\mathsf{Bern}(p)$	p	$p(1-p)$
$\mathsf{Bin}(p)$	np	$np(1-p)$
$\mathsf{Geo}(p)$	$1/p$	$(1/p^2) - (1/p)$
$\mathsf{NegBin}(np)$	n/p	$(n/p^2) - (n/p)$
$\mathsf{Exp}(\lambda)$	$1/\lambda$	$1/\lambda^2$
$\mathsf{Gamma}(n,\lambda)$	n/λ	n/λ^2
$\mathsf{Pois}(\mu)$	μ	μ
$\mathsf{Beta}(a,b)$	$a/(a+b)$	$ab/[(a+b)^2(a+b+1)]$
$\mathsf{N}(\mu,\sigma^2)$	μ	σ^2
$\mathsf{Cauchy}(\mu,\sigma)$	DNE	DNE

Here DNE means does not exist.

3.3. Special cases

Note that by these definitions, a Bernoulli with parameter p is also a Binomial with parameter 1 and p, and an Exponential with rate parameter λ is a Gamma with parameter 1 and λ. That is

Notation 4
The following distributin names are equivalent:

$$\begin{aligned} \mathsf{Bern}(p) &\sim \mathsf{Bin}(1,p) \\ \mathsf{Geo}(p) &\sim \mathsf{NegBin}(1,p) \\ \mathsf{Exp}(\lambda) &\sim \mathsf{Gamma}(1,\lambda) \\ \mathsf{Unif}(\{0,1\}) &\sim \mathsf{Bern}(1/2). \end{aligned}$$

This is helpful to know to understand why R does not have separate commands for the Bernoulli distribution.

3.4. Adding distributions

When you add independent random variables together:

- The sum of binomials with the same p is a binomial.
- The sum of negative binomials with the same p is a negative binomial.
- The sum of gammas with the same λ is also a gamma.
- The sum of Poisson random variables is also a Poisson.
- The sum of normals is also a normal.

Remember that Bernoulli is a special case of binomial, geometric is a special case of negative binomial, and exponential is a special case of gamma, which gives

- The sum of Bernoullis with parameter p is a binomial.
- The sum of geometrics with parameter p is a negative binomial.
- The sum of exponentials with the same λ is also a gamma.

How do you find the new parameters of the sum? Remember that the mean of the sum of random variables is the sum of the mean, and the variance of the sum of independent random variables is the sum of the variances. Then the eight rules above give.

Fact 12
Let $X_1, X_2, \ldots, X_n$ be independent random variables, and $S = X_1 + \cdots + X_n$.

- If $X_i \sim \mathsf{Bern}(p)$, then $S \sim \mathsf{Bin}(n,p)$.
- If $X_i \sim \mathsf{Bin}(n_i,p)$, then $S \sim \mathsf{Bin}(\sum n_i, p)$.
- If $X_i \sim \mathsf{Geo}(p)$, then $S \sim \mathsf{NegBin}(n,p)$.
- If $X_i \sim \mathsf{Exp}(\lambda)$, then $S \sim \mathsf{Gamma}(n,\lambda)$.
- If $X_i \sim \mathsf{Gamma}(n_i,\lambda)$, then $S \sim \mathsf{Gamma}(\sum n_i, \lambda)$.
- If $X_i \sim \mathsf{Pois}(\mu_i)$, then $S \sim \mathsf{Pois}(\sum \mu_i)$.
- If $X_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$ then $S \sim \mathsf{N}(\sum \mu_i, \sum \sigma_i^2)$.

Problems

3.1: For $X \sim \mathsf{Unif}([3, 4])$ find

(a) $\mathbb{E}[X]$.

(b) $\mathbb{V}(X)$.

3.2: Suppose that I have 10 subjects in an experiment. For each subject, either a drug is effective in lowering blood sugar or it is not. Assuming that the probability the drug is effective is 0.3, and that each subject behaves independently from the rest, what is the distribution of N, the number of subjects where the drug was effective?

Chapter 4

Conditioning

Question of the Day Suppose that $X \sim \mathsf{Unif}(\{1,2,3,4,5,6\})$, so X is a roll of a fair six sided die. Now suppose that we have extra information about X, namely, that X is at most 4. We use a vertical bar followed by the information to describe this. So we started with X, now we have $[X|X \leq 4]$. How can we incorporate what we have learned about X into our distribution of X? That is, what is the distribution of $[X|X \leq 4]$?

We say that we are interested in X *conditioned on* or *given* the information that $X \leq 4$. In this section we will cover the rules needed to work with conditional probabilities and expectations.

4.1. Conditional Probability

Start with the most important rule.

Fact 13
Suppose that we begin with random variable X and then have information encoded as $Y \in A$ such that $\mathbb{P}(Y \in A) > 0$. Then

$$\mathbb{P}(X \in B | Y \in A) = \frac{\mathbb{P}(X \in B, Y \in A)}{\mathbb{P}(Y \in A)}.$$

From our earlier example

$$\begin{aligned}\mathbb{P}(X = 3 | X \leq 4) &= \frac{\mathbb{P}(X = 3, X \leq 4)}{\mathbb{P}(X \leq 4)} \\ &= \frac{\mathbb{P}(X = 3)}{\mathbb{P}(X \leq 4)} \\ &= \frac{1/6}{4/6} = 1/4 = \boxed{0.2500}.\end{aligned}$$

By rearranging this formula we get Bayes Rule

Fact 14 (Bayes Rule)
Suppose $\mathbb{P}(X \in B)$ and $\mathbb{P}(Y \in A)$ are nonnegative. Then

$$\mathbb{P}(X \in B | Y \in A) = \frac{\mathbb{P}(Y \in A | X \in B)\mathbb{P}(X \in B)}{\mathbb{P}(Y \in A)}$$

There are analogues of these rules for densities as well.

Fact 15
Suppose that X and Y are random variables with densities $f_X(x)$ and $f_Y(y)$, and joint density $f_{(X,Y)}(x,y)$. Then

$$f_{X|Y=y}(x) = \frac{f_{(X,Y)}(x,y)}{f_Y(y)}.$$

and

$$f_{X|Y=y}(x) = \frac{f_{Y|X=x}(y)f_X(x)}{f_Y(y)}.$$

4.2. Conditional expectation

If the probabilities of a random variable X depend upon a second random variable Y, then we can talk about $\mathbb{E}[X|Y]$ (read the mean of X given Y or the mean of X conditioned on Y.) In this case, treat Y as a constant. The final expectation will be a function of Y.

For example, suppose Y is equally likely to be one of $\{1,2,3,4\}$. Given Y, X is the sum of the rolls of Y independent, fair, six-sided dice. The mean of one such die roll is 3.5. So the mean of the sum of 3 such rolls would be $3.5+3.5+3.5 = 3(3.5)$ by linearity. That means the sum of Y such rolls would be $Y(3.5)$. Hence

$$\mathbb{E}[X|Y] = 3.5Y,$$

which is a function of Y.

Another useful fact about conditional expectation is known as the *Law of Total Expectation* or the *Fundamental Theorem of Probability*.

Theorem 1
Suppose X and $[X|Y]$ are integrable random variables. Then

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X].$$

Continuing our earlier example. Suppose we wanted to know $\mathbb{E}[X]$. Then using the FTP:

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[3.5Y] = 3.5[(1/4)1 + (1/4)(2) + (1/4)(3) + (1/4)(4)] = 8.75.$$

Problems

4.1: Suppose Y is equally likely to be 1, 2, or 3. Let X_1, X_2, X_3 be independent draws of a random variable with density $f(1) = 0.3$, $f(2) = 0.3$, $f(3) = 0.4$ with respect to counting measure.

(a) What is $\mathbb{E}[X_i]$?

(b) What is

$$\mathbb{E}\left[\sum_{i=1}^{Y} X_i\right]?$$

Chapter 5

Optimization and Logarithms

Question of the Day Find the minimum value of $(x-2)^2$.

A problem that arises frequently in statistics is optimization, specifically, finding the inputs to a function that leads to either the largest or smallest value of the function. Let A be the set of inputs to a function f and say the function has output in $\mathbb{R}$. Write $f : A \to \mathbb{R}$.

Definition 14
The maximum of f over A, written $\max_{x\in A} f(x)$, is a value M such that there exists an $x^* \in A$ such that $f(x^*) = M$ and for all $x \in A$, $f(x) \leq f(x^*)$.

For example, $\max_{x\in[-4,4]} -(x-2)^2 = 0$ since any quantity squared is at least 0, so $(x-2)^2 \geq 0$ and $(-(x-2)^2 \leq 0$ for all x.

Note that the maximum of the function refers to the *output* of the function. If we are interested in the input to the function that reaches the maximum output, that is the argument maximizer, or arg max for short.

Definition 15
The argument maximizer of f over A, written $\arg\max_{x\in A} f(x)$ is x^* if for all x in A, $f(x) \leq f(x^*)$.

Continuing our earlier example, $\arg\max_{x\in[-4,4]} -(x-2)^2 = 2$, since $-(2-2)^2 = 0$, which is the maximum value of the function.

Note that not every function has such a maximum value. For instance, $\max_{x\in(0,1)} 1/x = \infty$, so we say the maximum does not exist. A helpful fact is that if A is a *compact* set (so closed and bounded) and f is a continuous function, then the maximum must exist.

One useful fact in finding maxima is that if we can break the inputs into several regions and find the maximum value on each, then we can take the maximum of the maximum to find the overall maximum.

Fact 16
Suppose that $\max_{x\in A} f(x)$ and $\max_{x\in B} f(x)$ exist and are finite. Then

$$\max_{x\in A\cup B} f(x) = \max\left\{\max_{x\in A} f(x), \max_{x\in B} f(x)\right\}.$$

Finding the optimal values is greatly aided if the function has continuous derivatives. If the function has a continuous first derivative say that $f \in C^1$, for a continuous second derivative $f \in C^2$, and so on.

The following fact is useful when dealing with compact sets, or sets where the derivative is nonpositive or nonnegative.

Fact 17
Let $f \in C^1$. Then the following holds.

- If $(\forall x \le a)(f'(x) \ge 0)$ then $\max_{x \le a} f(x) = f(a)$ and $\arg\max_{x \le a} f(x) = a$.
- If $(\forall x \le a)(f'(x) \le 0)$ then $\max_{x \ge a} f(x) = f(a)$ and $\arg\max_{x \ge a} f(x) = a$.
- $\arg\max_{x \in [a,b]} f(x) \subseteq \{a, b\} \cup \{c : f'(c) = 0\}$.

These facts combined can allow us to solve problems over unbounded sets. For example, consider finding $\arg\max_{x \ge 0} f(x)$ where $f(x) = x^2 \exp(-3x)$. Then $f'(x) = 2x\exp(-3x) - 6x^2\exp(-3x) = 2x\exp(-3x)(2 - 3x)$. Since $\exp(-3x) > 0$ always, and $2x > 0$ for $x > 0$, the first derivative is nonnegative when $2 - 3x \ge 0$ (so $x \le 2/3$) and nonpositive when $2 - 3x \le 0$ (so $x \ge 2/3$).

Hence $\max_{x \in [0,2/3]} f(x) = f(2/3)$ and $\max_{x \ge 2/3} f(x) = f(2/3)$, so

$$\max_{x \in [0,\infty)} f(x) = \max_{x \in [0,2/3] \cup [2/3,\infty)} f(x) = \max(f(2/3), f(2/3)) = f(2/3).$$

Therefore, $\arg\max_{x \ge 0} f(x) = 2/3 = \boxed{0.6666\ldots}$.

5.1. Logarithms

Logarithms were originally developed as a computational tool into order to make multiplications and raising numbers to powers easier. Over the centuries, mathematicians learned even more about this function, and it turned out to show up in unexpected places. Here are the most important facts about logarithms.

Fact 18
The **natural logarithm** function $\ln(x)$ (aka **natural log** aka ell-en) maps positive numbers to real numbers, and has the following properties.

1: The logarithm is a strictly increasing function: $(\forall a < b)(\ln(a) < \ln(b))$.

2: Natural log is the inverse of the exponential function, so for any real x, $\ln(\exp(x)) = x$, and for any $w > 0$, $\exp(\ln(w)) = w$.

3: For positive a and b,
$$\ln(ab) = \ln(a) + \ln(b).$$

4: For positive a and arbitary b
$$\ln(a^b) = b\ln(a).$$

Of course, we can generalize the product/summation rule to arbitrary sets of numbers with our product and summation notations:

$$\ln\left(\prod_{i=1}^{n} a_i\right) = \sum_{i=1}^{n} \ln(a_i).$$

Because the natural logarithm is a strictly increasing function, it can be used to find the argument that maximizes or minimizes other functions.

Fact 19
Suppose that $f(x_1, \ldots, x_n) > 0$. Then

$$\arg\max f(x_1, \ldots, x_n) = \arg\max \ln(f(x_1, \ldots, x_n)).$$

Example Find $\arg\max_{x \in (0,1)} x^2(1 - x)$.

Note that

$$\begin{aligned}\arg\max_{x\in(0,1)} x^2(1-x) &= \arg\max_{x\in(0,1)} \ln(x^2(1-x)) \\ &= \arg\max_{x\in(0,1)} [2\ln(x) + \ln(1-x)].\end{aligned}$$

Now $[2\ln(x)+\ln(1-x)]' = 2/x - 1/(1-x)$ which is at least 0 for $x \leq 2/3$ and which is at most 0 for $x \geq 2/3$. Hence $\boxed{x = 2/3}$ is the argument that maximizes the original function.

Problems

5.1: True or false: If $\max_\theta f(\theta)$ exists for $f(\theta) \geq 0$, then $\max_\theta f(\theta) = \max_\theta \ln(f(\theta))$.

5.2: Find $\max_{[0,\infty)} x\exp(-2x)$.

5.3: Find $\arg\max \exp(-(x-4)^2/2)$.

5.4: Find $\arg\max_{\lambda>0} \lambda^3\exp(-2.1\lambda)$

Part II

Statistics

Chapter 6

Introduction to Statistics

Question of the Day How many chocolate chips are in my cookie?

Statistics is the science of collecting and analyzing data in order to make informed decisions. The term *statistics* was coined by a German political scientist Gottfried Achenwall in 1749. Originally the data to be considered came from the state, hence the term statistics. In England it became known as *political arithmetic* for similar reasons.

It was in the 1800's that statistics started to acquire the more general meaning involving data collected not just for and by the state, but from more general sources. Today statistics is a bedrock of all sciences and social sciences, and anywhere data is collected or analyzed, you will need tools from statistics.

This text is intended to form the basis for two-thirds of a one semester course in statistics. It covers the philosophy, theory, and mathematics behind the statistics that people use. The one-third that is missing is the applied part of the equation, which will be provided by your instructor with lab experiments where you get to actually use data and statistical software.

6.1. The cookie experiment

To begin, let us start with an experiment that will give you an idea of some of the issues facing any use of statistics to analyze data in the real world.

Suppose that you are a consultant for a cookie company. You take a package of the companies cookies, and pour the cookies out on a table. Your assistants then take each cookie, and count the number of chips that they find inside. The data is recorded.

Models

- What makes a good model for chocolate chip cookies?
- What decisions does the manufacturer need to make based on the data?
  - **1:** Could want average number of chips to be in a certain range
  - **2:** Could want there to be a low chance of no chips
  - **3:** Could want low spread in number of chips
- The point is: different goals require different statistical analyses.
- Since the number of chips in any particular cookie is unknown (and difficult to find exactly, it makes sense to use a *probabilistic model* for the number of chips in a cookie.
- Let N be the number of chips in a given cookie.
- Want to assign probabilities to $N = i$ for $i \in \{0, 1, 2, \ldots\}$.
- Two standard distributions assign positive probability to nonnegative integers: Geometric, and Poisson.

- Geometric: # of coin flips until you get a heads.
- Poisson: Number of events when expected number proportional to size. Bingo! Number of chips proportional to batter used per cookie.
- For $X \sim \mathsf{Pois}(\mu)$, $\mathbb{P}(X = i) = \exp(-\mu)\mu^i/i!$.
- Hence our statistical model becomes $X_1, \ldots, X_n \sim X$ are iid (independent, identically distributed):

$$\mathbb{P}(X = i|\mu) = \exp(-\mu)\frac{\mu^i}{i!}.$$

- $X_1, \ldots, X_n$ are data, μ is a parameter of the model.
- More sophisticated models might get rid of the independent, or identically distributed, or use a different distribution.
- The simpler the model, the easier to work with.

Now that we have a model now what?

- Could ask for a point estimate of μ. That means a single best guess for what the value of μ is.
- Could ask for an interval estimate of μ. That would be some interval $[a, b]$ that it is believed that μ falls into.
- Could ask to determine a property of μ. For instance, whether $\mu < 5$ or $\mu \geq 5$?
- These three equations are called point estimation, interval estimation, and hypothesis testing

6.2. Two main philosophies of statistics

There are two main philosophies of statistics that are in widespread use today. These are known as *Bayesian* and *Frequentist*.

The Bayesian approach works as follows. For a given quantity, such as a parameter of the model μ, use a probabilistic model to represent the partial information that we have about μ. Then once data $X_1, \ldots, X_n$ is obtained, use Bayes' Rule to find $\mathbb{P}(\mu \in A|X_1, \ldots, X_n)$ for a given set A. So the distribution of μ is a function of the data values $X_1, \ldots, X_n$ This approach is called *Bayesian statistics*

In the frequentist approach, again μ is unknown, but we do not try to model it ahead of time. Instead, we create functions of the data $f(X_1, \ldots, X_n)$ such that $\lim_{n\to\infty} f(X_1, \ldots, X_n) = \mu$ with probability 1. This is called *frequentist statistics*.

Bayesian advantages and disadvantages

- Ad: Very clear from a philosophical perspective, and can always apply at a theoretical level.
- Ad: Makes crystal clear starting position and ending position based on data.
- Ad: Allows knowledge of μ to be incorporated into estimate.
- Dis: impartiality brought into question.
- Dis: In practice, computations can become difficult (moreso than frequentist).
- Ad: Rise of computers has breathed new life into Bayesian statistics.

Frequentist advantages and disadvantages

- Ad: Fits scientific paradigm of statistician as impartial.
- Ad: Computations easier than Bayesian.
- Dis: takes advanced study to understand results (p-values, confidence intervals).
- Dis: does **not** return probabilities, which are much more easily understood.

Definition 16
Let $X_1, X_2, \ldots$ be a data stream. Then any function $f(X_1, X_2, \ldots)$ is a **statistic** of the data.

Example: Bayesian approach

- Suppose you know that $\mu \in \{4.1, 5.2\}$.
- Start with a *noninformative prior* (no knowledge):

$$\mathbb{P}(\mu = 4.1) = \mathbb{P}(\mu = 5.2) = 1/2.$$

- Suppose $X_1 = 3, X_2 = 4, X_3 = 7$. Let $\vec{X} = (X_1, X_2, X_3)$. Then

$$\mathbb{P}(\mu = 4.1 | \vec{X} = (3,4,7)) = \frac{\mathbb{P}(\vec{X} = (3,4,7) | \mu = 4.1)\mathbb{P}(\mu = 4.1)}{\mathbb{P}(\vec{X} = (3,4,7) | \mu = 4.1)\mathbb{P}(\mu = 4.1) + \mathbb{P}(\vec{X} = (3,4,7) | \mu = 5.2)\mathbb{P}(\mu = 5.2)}$$

Note

$$\begin{aligned}\mathbb{P}(\vec{X} = (3,4,7) | \mu = 4.1) &= \left(\exp(-4.1)4.1^3/3!\right)\left(\exp(-4.1)4.1^4/4!\right)\left(\exp(-4.1)4.1^7/7!\right) \\ &= \exp(-4.1 \cdot 3)(4.1)^{X_1+X_2+X_3}/[X_1!X_2!X_3!] \\ &\approx 0.002378805\ldots\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(\vec{X} = (3,4,7) | \mu = 5.2) &= \left(\exp(-5.2)5.2^3/3!\right)\left(\exp(-5.2)5.2^4/4!\right)\left(\exp(-5.2)5.2^7/7!\right) \\ &= \exp(-5.2 \cdot 3)(5.2)^{X_1+X_2+X_3}/[X_1!X_2!X_3!] \\ &\approx 0.002469372\ldots\end{aligned}$$

So

$$\mathbb{P}(\mu = 4.1 | \text{data}) = \frac{0.002378805(1/2)}{0.002378805(1/2) + 0.002469372(1/2)} \approx 0.4906.$$

- Note: $1/[3!4!7!]$ appeared in both expressions, so canceled, so didn't really need to be a part of things.

Example: frequentist approach

- The strong law of large numbers, for $X_1, X_2, \ldots \sim X$ iid, where $\mathbb{E}[|X|] < \infty$,

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n}\right) = \mathbb{E}[X].$$

- For $X \sim \mathsf{Pois}(\mu)$, $\mathbb{E}[X] = \mu$. Hence let

$$\hat{\mu}_n = \frac{X_1 + \cdots + X_n}{n}.$$

- Note: usually use "hat" of variable, $\hat{\mu}$, as estimate for μ.
- For simple data: $\hat{\mu} = (3 + 4 + 7)/3 \approx 4.666$.

[Time permitting, show how to do this in R.]

References The idea to have students take data from chocolate chip cookies came from a wonderful statistician, Herbie Lee. You can read more details at

Lee, H.K.H. (2007). "Chocolate Chip Cookies as a Teaching Aid." *The American Statistician*, 61, 4, 351–355.

Problems

6.1: True or false: if an experimenter is careful, they will always get the same result for their data.

6.2: Fill in the blank: For data $X_1, X_2, \ldots$, $(X_1 + \cdots + X_{15})/15$ and $\max_i X_i$ are examples of _____.

Chapter 7

Method of Moments Estimator

Question of the Day Suppose that the times needed for service are modeled using iid exponential random variables $X_1, X_2, \ldots$ with rate λ. If the first three services take time 34, 23, and 17 minutes, estimate the rate λ.

In this chapter

- Consistent estimators
- Strong Law of Large Numbers
- Method of moments

Last time

- Created a statistical model (probabilistic model) of how data was generated
- This model has parameters
- Want to know parameters!
  - **1:** Stocks, global temperatures, poverty: are they going up or down on average?
  - **2:** What about the volatility/spread/variance?

One thing to remember whenever you are creating a mathematical model: no model is perfect! There will always be deviations from the model. So the mark of a good model is not that it is 100% accurate, more that a good model doesn't break completely when there are small errors. Moreover, a good model allows one to make reasonable predictions about future behavior of the system. George Box is known for the following quote about statistical models:

> All models are wrong, but some are useful.

7.1. Consistent estimators

So what makes a good statistical estimator? Again there are many criteria that can be applied. One of the simplest is the notion of *consistency*. Intuitively, this means that as you take more and more data, the value of your statistic should get closer to the true value of the parameter.

Definition 17
A family of estimators $\{\hat{\theta}\}_n$ is **consistent** if $\lim_{n\to\infty} \hat{\theta} = \theta$ with probability 1.

- Variables with finite expectation can give consistent estimators.

Definition 18
Say that X has **finite first moment** or is **integrable** if $\mathbb{E}[|X|] < \infty$.

Definition 19
If X^i is integrable, say that $\mathbb{E}[X^i]$ is the ***i*-th moment of X**.

- All random variables $Y \geq 0$ have $\mathbb{E}[Y]$ defined, it might be in $[0, \infty)$, or it might equal ∞.
- $|X| \geq 0$, so $\mathbb{E}[|X|]$ always defined.
- Example: $X \sim \mathsf{Cauchy}$
  - Note if $U \sim \mathsf{Unif}([-\pi/2, \pi/2])$ then $X = \arctan(U)$ is Cauchy.
  - Density of X is $f_X(s) = [\pi(1 + s^2)]^{-1}$.
  - $\mathbb{E}[X]$ is undefined, but $\mathbb{E}[|X|] = \infty$, so X is not integrable.

Theorem 2 (Strong Law of Large Numbers)
Let $X_1, X_2, \ldots \sim X$ be iid, and $\mathbb{E}[|X|] < \infty$. Then

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mathbb{E}[X]\right) = 1.$$

In words: the sample average converges to the true average with probability 1.

7.2. How to build a Method of Moments estimate

- Suppose $X \sim D(\theta)$, where $D(\theta)$ is some family of distributions with vector parameter $\theta \in \mathbb{R}^n$.
  - In QotD, $X \sim \mathsf{Exp}(\lambda)$, $\theta = \lambda$, $n = 1$.
  - If $X \sim \mathsf{N}(\mu, \sigma^2)$, then $\theta = (\mu, \sigma)$, $n = 2$.
- Then $\mathbb{E}[X] = g_1(\theta)$, $\mathbb{E}[X^2] = g_2(\theta), \ldots$
- Use sample averages to get estimates $(\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_n)$ for $(\mathbb{E}[X], \mathbb{E}[X^2], \ldots, \mathbb{E}[X^n])$.
- Solve system of equations for $\theta = (\theta_1, \ldots, \theta_n)$:

$$\begin{aligned} \hat{g}_1 &= g_1(\theta) \\ \hat{g}_2 &= g_2(\theta) \\ \vdots &= \vdots \\ \hat{g}_n &= g_n(\theta) \end{aligned}$$

Qotd

- For $X \sim \mathsf{Exp}(\lambda)$, $\mathbb{E}[X] = 1/\lambda$.
- So let $\bar{X} = (X_1 + \cdots + X_n)/n$ (read X-bar), then solve

$$\mu_n = \frac{1}{\hat{\lambda}},$$

to get $\hat{\lambda} = 1/\bar{X}$.

- For Qotd: $\hat{\lambda} = [(34 + 23 + 17)/3]^{-1}$.

Doing this in R

- Can combine numbers into a vector with `c` command
  Find sample average with `mean` command

```
data <- c(34,23,17)
1/mean(data)
```

  gives 0.04054 as answer (to 4 sig figs)

SLLN gives consistency

- Remember, we know that
$$\mathbb{P}\left(\frac{X_1^i + \cdots + X_n^i}{n} = \mathbb{E}[X^i]\right) = 1,$$
  so as n goes to infinity, a MOM estimate will converge to true answer for θ as long as system of equations is "nice".

An important note

- Important: take sample average first, then solve.
- Estimate is $1/\bar{X}$, not $\overline{(1/X)}$
$$\frac{(1/34) + (1/23) + (1/17)}{3}.$$
  In general: $\mathbb{E}[f(X)] \neq f(\mathbb{E}[X])$.
- For $X \sim \mathsf{Exp}(\lambda)$, $\mathbb{E}[1/X] = \infty$, so SLLN does not apply!

An example that uses the second moment

- Suppose $X_1, X_2, \ldots \sim \mathsf{N}(\mu, \sigma^2)$.
- Then find MOM estimators for μ and σ^2.
$$\hat{g}_1 = \frac{X_1 + \cdots + X_n}{n}, \ \hat{g}_2 = \frac{X_1^2 + \cdots + X_n^2}{n}.$$
- μ easy: $\mathbb{E}[X] \approx \hat{g}_1$, so use $\hat{\mu} = \hat{g}_1$.
- σ^2: recall $\sigma^2 = \mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. Since $\mathbb{E}[X] \approx \hat{g}_1$ and $\mathbb{E}[X^2] \approx \hat{g}_2$, $\sigma^2 \approx \hat{g}_2 - \hat{g}_1^2$. Together:
$$\hat{\mu} = \hat{g}_1$$
$$\hat{\sigma}^2 = \hat{g}_2 - \hat{g}_1^2$$

Of course, the Method of Moments estimator is not perfect. We say that an estimate $\hat{\theta}$ for θ is *unbiased* if $\mathbb{E}[\hat{\theta}] = \theta$. The MOM estimate for σ^2 above turns out to *not* be unbiased. We say that it is *biased*. In the next chapter we will see how to build an unbiased estimator for the variance.

Problems

7.1: Suppose that $X_1, \ldots, X_n$ given θ are iid $\mathsf{Unif}([0, \theta])$. Find the Method of Moments estimate of θ.

7.2: Suppose I model X given θ as being $\mathsf{Unif}([\theta, 2\theta])$. Say $X_1, \ldots, X_n$ are iid draws from X.

(a) What is the likelihood function $L_{x_1,\ldots,x_n}(\theta)$ given $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$?
(b) Derive the MLE for θ given data $x_1, \ldots, x_n$.
(c) Evaluate your MLE at data $1.3, 2.1, 1.7$.
(d) Derive the MOM for θ given data $x_1, \ldots, x_n$.
(e) Evaluate your MOM at data $1.3, 2.1, 1.7$.

Chapter 8

Unbiased estimators

Question of the Day For $X_1, X_2, \ldots \sim \mathsf{N}(\mu, \sigma^2)$, give an unbiased estimator for μ and σ^2.

In this chapter

- Unbiased estimators
- Unbiased estimator of variance

Last time we saw that a good property for a statistical estimator to have is that it be *consistent*, so that as we take more data, the estimate gets closer to the true answer. Today we will will looks at another good property for estimators to have, namely, that they are *unbiased.*

Definition 20
Call an estimator $\hat{\theta}$ for θ **unbiased** if $\mathbb{E}[\hat{\theta}] = \theta$.

Why is unbiasedness good?

- Suppose that $\hat{\mu}_1$ and $\hat{\mu}_2$ are both unbiased estimates of μ. Then $(\hat{\mu}_1 + \hat{\mu}_2)/2$ is also an unbiased estimator of μ, and the spread in the estimator is generally smaller.

Example

- Give an unbiased estimator for $\mathbb{E}[X]$.
  - Answer: $\bar{X}$. Say $\mathbb{E}[X] = \mu$.

$$\begin{aligned}\mathbb{E}[\bar{X}] &= \mathbb{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] \\ &= \frac{\mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n]}{n} \\ &= \frac{\mu + \mu + \cdots + \mu}{n} \\ &= \frac{n\mu}{n} = \mu.\end{aligned}$$

8.1. The unbiased estimator of variance

Okay, so finding an unbiased estimator of $\mathbb{E}[X]$ is easy, just use the sample average. Finding an unbiased estimator for $\mathbb{V}(X)$ is a bit trickier, but can be done! Here are the steps.

- Recall:

$$\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

- So
$$a = \overline{(X - \mathbb{E}[X])^2},$$
is an unbiased estimator for variance, but we can't calculate a since we don't know $\mathbb{E}[X]$.
- Idea: why not use $\bar{X}$ instead of $\mathbb{E}[X]$.
$$b = \overline{(X - \mathbb{E}[X])^2}.$$
Problem: $\mathbb{E}[b] = \mathbb{V}(X)(n-1)/n$.
- So just scale to get the unbiased estimate.

Theorem 3 (Unbiased estimators)
Let $X_1, X_2, \ldots \sim X$ be iid with finite second moment. Then unbiased estimators for $\mathbb{E}[X]$ and $\mathbb{V}[X]$ are (for any n)
$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}, \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}.$$

[In R, use `var` to compute $\hat{\sigma}^2$.]
Before proving, it helps to have a lemma,

Lemma 1
Let $X_1, X_2, \ldots \sim X$ be iid with finite second moment. Then for any n,
$$\mathbb{E}[(\bar{X})^2] = (1/n)[\mathbb{E}[X_i^2] + (n-1)\mathbb{E}[X_i]^2]$$

Proof.
$$\begin{aligned}
\mathbb{E}[(\bar{X})^2] &= \mathbb{E}\left[((1/n)\sum_{i=1}^n X_i)^2\right] \\
&= (1/n^2)\mathbb{E}[\sum_{i=1}^n X_i^2 + \sum_{i \neq j} 2X_i X_j] \\
&= (1/n^2)[n\mathbb{E}[X_i^2] + (n)(n-1)\mathbb{E}[X_i]^2 \\
&= (1/n)[\mathbb{E}[X_i^2] + (n-1)\mathbb{E}[X_i]^2]
\end{aligned}$$
□

Now we are ready to prove Theorem 3.

Proof of Theorem 3. Since $\mathbb{E}$ is a linear operator,
$$\mathbb{E}\left[\frac{\sum_{i=1}^n X_i}{n}\right] = \frac{\sum_{i=1}^n \mathbb{E}[X_i]}{n} = \frac{n}{n}\mathbb{E}[X] = \mathbb{E}[X].$$

For the variance part, first note that $\sum_i X_i = n\bar{X}$, so
$$\begin{aligned}
\sum_{i=1}^n (X_i - \bar{X})^2 &= \sum_{i=1}^n (X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\
&= \sum_{i=1}^n X_i^2 - 2n\bar{X}\bar{X} + n\bar{X}\bar{X} \\
&= \sum_{i=1}^n X_i^2 - (n)(\bar{X})^2.
\end{aligned}$$

Using the previous lemma,

$$\begin{aligned}\mathbb{E}[\hat{\sigma}^2] &= \frac{1}{n-1}\left[n\mathbb{E}[X^2] - n(1/n)(\mathbb{E}[X^2] + (n-1)\mathbb{E}[X]^2)\right] \\ &= \frac{1}{n-1}[(n-1)\mathbb{E}[X^2] - (n-1)\mathbb{E}[X]^2] = \mathbb{V}(X).\end{aligned}$$

□

Intuition Now let us take a look at the intuition behind dividing by $n-1$ rather than by n.

We start with $X_1, \ldots, X_n$ able to freely vary. Statisticians like to say that there are n *degrees of freedom* for the random variable. On the other hand, $\hat{\sigma}$ uses $\sum(X_i - \bar{X})$. For a fixed value of $\bar{X}$, any $n-1$ values of X_i determine the last value. For instance, if for $n = 3$ we know $\bar{X} = 1.6$ and $X_1 = 2.1$, $X_3 = 1.0$, then we can solve for $X_2 = 1.7$.

So knowing $\bar{X}$ "uses up" one degree of freedom, leaving only $n-1$ degrees. That is definitely not a formal proof, but often these type of informal arguments turn out to be correct. Of course, they need to be justified by formal mathematics like in the proof above!

Problems

8.1: Given data $(1.7, 1.6, 2.4, 3.1)$,

(a) Give an unbiased estimate of the mean of the distribution.

(b) Give an unbiased estimate of the variance of the distribution.

Chapter 9

Maximum likelihood estimators

Question of the Day Suppose $X_1, X_2, X_3 \overset{\text{iid}}{\sim} \exp(\lambda)$. What is the maximum likelihood estimator for λ?

In this chapter

- Maximum likelihood estimator
- Consistency of the estimator

9.1. Likelihood function

Definition 21
Suppose data $X_1, \ldots, X_n$ is drawn from a statistical model whose density $f_\theta(x_1, \ldots, x_n)$ (also written $L(\theta|x_1, \ldots, x_n)$) is a function of the parameter θ. Then call f the **likelihood function**.

Example

- Suppose $X_1, X_2, X_3 \sim X$ are iid and $X \sim \mathsf{Exp}(\lambda)$. Then $f_X(s) = \lambda \exp(-\lambda s)\mathbb{1}(s \geq 0)$.
- Here $\mathbb{1}(\cdot)$ denotes the indicator function that equal 1 whenever it's argument is true, and 0 otherwise.
- Since X_1, X_2, X_3 are independent, the joint density is the product of the individual densities:

$$\begin{aligned} f_\lambda(x_1, x_2, x_3) &= \lambda^3 \exp(-\lambda x_1) \exp(-\lambda x_2) \exp(-\lambda x_3)\mathbb{1}(x_1, x_2, x_3 \geq 0) \\ &= \lambda^3 \exp(-\lambda(x_1 + x_2 + x_3))\mathbb{1}(x_1, x_2, x_3 \geq 0). \end{aligned}$$

Definition 22
The **maximum likelihood estimator** (or MLE) for data $X_1 = x_1, \ldots, X_n = x_n$ is any value of θ that maximizes the likelihood function. That is,

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta f_\theta(x_1, \ldots, x_n).$$

9.2. Maximizing functions

Recall that the maximum of a function is a value that is bigger than all the other values that the function can take on.

Definition 23
Say that $f(x)$ has **maximum** value M for $x \in A$ if $(\exists x^*)(f(x^*) = M)$ and $(\forall x \in A)(f(x) \leq f(x^*) = M)$. If this holds, write

$$\max_{x \in A} f(x) = M.$$

The word *argument* is another term for the input to a function. The argument maximum (often shortened to arg max) is the argument value which causes f to attain its maximum value.

Definition 24
The **argument maximum** of f over A is the set of points A^* such that for all $x^* \in A^*$, $f(x^*) = \max_{x \in A} f(x)$. Write

$$\arg\max_{x \in A} f(x) = A^*.$$

Note

- If f has no maximum value over A, $A^* = \emptyset$.
- If $A^* = \{x^*\}$ for some state $x^* \in A$, then x^* is the unique global maximizer of f over A.
- If I compose f with a strictly increasing function, that can change the maximum value, but it will not change the argument maximum!
- For example, the function $g(r) = r^2$ is strictly increasing for $r \in [0, \infty)$. So if $f(x) \in [0, \infty)$ for all $x \in A$:

$$\arg\max_{x \in A} f(x) = \arg\max_{x \in A} f(x)^2.$$

We know that for independent random variables, the joint density is the product of the individual densities. Now, logarithms turn products into sums (that are easier to deal with), plus the natural log function is strictly increasing. So we have the following.

Fact 20 (Argmax same for g and $\ln(g)$)
Suppose for all $\theta \in A$, $g(\theta) \geq 0$, and $g(\theta)$ attains its maximum value for at least one $\theta \in A$. Then

$$\arg\max_{\theta \in A} g_\theta(x_1, \ldots, x_n) = \arg\max_{\theta \in A} \ln(g_\theta(x_1, \ldots, x_n),$$

Proof. This works because natural log is a strictly increasing function and the likelihood is nonnegative and strictly positive somewhere. □

This is very useful when dealing with problems like the Question of the Day!

Example

- For the QotD:

$$\ln(f_\lambda(x_1, x_2, x_3)) = [3\ln(\lambda) - \lambda(x_1 + x_2 + x_3)]\mathbb{1}(x_1, x_2, x_3 \geq 0)$$
$$[\ln(f_\lambda(x_1, x_2, x_3))]' = \left[\frac{3}{\lambda} - (x_1 + x_2 + x_3)\right] \mathbb{1}(x_1, x_2, x_3 \geq 0).$$

- The derivative is ≥ 0 for $\lambda \leq [(x_1 + x_2 + x_3)/3]^{-1}$, and ≤ 0 for $\lambda \geq [(x_1 + x_2 + x_3)/3]^{-1}$. Hence

$$\arg\max_{\lambda \leq [(x_1+x_2+x_3)/3]^{-1}} \ln(f_\lambda) = [(x_1 + x_2 + x_3)/3]^{-1}$$
$$\arg\max_{\lambda \geq [(x_1+x_2+x_3)/3]^{-1}} \ln(f_\lambda) = [(x_1 + x_2 + x_3)/3]^{-1},$$

which means

$$\lambda_{\text{MLE}} = \left[\frac{x_1 + x_2 + x_3}{3}\right]^{-1},$$

the same as the MOM estimator!

9.3. Consistency of the MLE

Our goal in this section is to show that under mild regularity conditions, the MLE estimator for independently drawn data will be consistent. As with the Method of Moments estimator, our proof will be based upon the Strong Law of Large Numbers.

First note that for independent data drawn from density f_θ,

$$\hat{\theta}_n = \arg\max_\theta f_\theta(x_1, \ldots, x_n)) = \arg\max_\theta \frac{1}{n} \ln(f_\theta(x_1, \ldots, x_n)) = \arg\max_\theta \frac{1}{n} \sum_{i=1}^{n} \ln(f_\theta(x)),$$

and the thing being maximized looks kind of like a sample average.

Let θ_0 be the true value of the parameter. Define

$$L(\theta) = \mathbb{E}[\ln(f_{\theta_0}(X))],$$

where X is a draw from the statistical model with the true parameter θ_0. Assume that $L(\theta)$ exists and is finite.

That means that if we set

$$L_n(\theta) = \frac{1}{n} \ln(f_\theta(x)),$$

then by the SLLN, as $n \to \infty$, with probability 1, $L_n(\theta) \to L(\theta)$ for all θ.

Even more, it will converge from below by the following fact.

Fact 21
For any θ, $L(\theta) \leq L(\theta_0)$, and $L(\theta) = L(\theta_0) \Leftrightarrow \mathbb{P}(f(X|\theta)) = f(X|\theta_0) = 1$.

Another way to say this is $L(\theta) \leq L_n(\theta)$ for all values of θ and it is strict inequality unless θ and θ_0 index exactly the same statistical model.

Proof. We will write the proof for continuous random variables for convenience, but by integrating with respect to counting measure rather than Lebesgue measure the same proof can be adapted to discrete random variables.

We are after the difference

$$L(\theta) - L(\theta_0) = \mathbb{E}[\ln(f(X|\theta)) - \ln(f(X|\theta_0))] = \mathbb{E}[\ln(f(X|\theta)/f(X|\theta_0))].$$

We want this different

Recall that $\ln(t) \leq t - 1$ (the tangent line at $t = 1$ lies on or above the convex function $\ln()$) so that

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[\ln\left(\frac{f(X|\theta)}{f(X|\theta_0)}\right)\right] &\leq \mathbb{E}_{\theta_0}\left[\frac{f(X|\theta)}{f(X|\theta_0)} - 1\right] \\
&= \int_{x:f(x|\theta_0)>0} \left(\frac{f(x|\theta)}{f(x|\theta_0)} - 1\right) f(x|\theta_0)\, dx \\
&= \int_{x:f(x|\theta_0)>0} f(x|\theta)\, dx - \int_{x:f(x|\theta_0)>0} f(x|\theta_0)\, dx = 1 - 1 = 0.
\end{aligned}$$

Note that the two integrals in the last line are always 1 because densities always integrate to 1! So we have the inequality.

In order to obtain equality between $L(\theta)$ and $L(\theta_0)$, it must hold that $\ln(t) = t - 1$ which only occurs at $t = 1$, so $f(X|\theta) = f(X|\theta_0)$ with probability 1. That only happens if they have the same density!. □

Another way to state this fact is that $\arg\max L(\theta) = \{\theta_0\}$.

With this in hand and some regularity conditions, it is possible to prove the following.

Theorem 4 (The MLE is consistent)
Suppose that the space of parameter values θ can take on is compact, $\ln(f(x|\theta))$ is continuous in θ for all values of x. Also suppose there exists a function $F(x)$ such that $\mathbb{E}[F(X)] < \infty$ (where X is a draw from the statistical model at the true parameter value θ_0) and $|\ln(f(x|\theta))| \leq F(x)$ for all x and θ. Then

$$\mathbb{P}(\hat{\theta}_n \to \theta_0) = 1,$$

where $\hat{\theta}_n$ is the MLE estimator using the first n values from the data stream $X_1, X_2, \ldots \overset{\text{iid}}{\sim} [X|\theta_0]$.

Proof idea. We've seen that θ_0 is the maximizer for $L(\theta)$. And by the SLLN, $L_n(\theta) \to L(\theta)$. So use continuity to show that the maximizer of $L_n(\theta)$ (which is $\hat{\theta}_n$) must approach the maximizer of $L(\theta)$ (which is θ). □

Problems

9.1: Suppose $[X|\theta] \sim \mathsf{Unif}([0, \theta])$, and $[X_1, \ldots, X_n|\theta]$ are iid from $[X|\theta]$.

(a) What is the likelihood of θ given $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$?

(b) Find the MLE of θ given $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$.

9.2: True or false: The maximum likelihood estimator is always unbiased.

9.3: Suppose that an experimenter runs a sequence of trials that are each independently a success with parameter p.

(a) Let T be the number of trials needed for one success. So if the sequence was fail, fail, success, then $T = 3$. Find the MLE of p as a function of T.

(b) Find the Method of Moments estimate of p as a function of T.

Chapter 10

Bayesian point estimators

Question of the Day The chance a new drug lower cholesterol by 20 points or more is p, initially taken to be uniform over $[0, 1]$. Suppose the drug is tested on 10 patients, 3 of whom have success with the drug. What is the best Bayesian estimate of p?

In this chapter

- Point estimates with Bayesian statistics

10.1. How Bayesian statistics works

Bayesian statistics approaches the goal of estimating parameters of a statistical model in the following way. The parameters that we are trying to estimate are unknown, and so first build a probability model for the parameter. This probability model for the parameters is called a *prior*. Unlike most situations with random variables, we still tend to use the lowercase Greek letter θ rather than the uppercase one to denote the random value.

Definition 25
Given a statistical model $[X|\theta]$ where X is the observed data and θ is a parameter of the model, the distribution of θ is called the **prior**.

It is called the prior because this distribution is set prior to taking any data.

For instance, in the Question of the Day, the prior for the unknown probability p is uniform over $[0, 1]$ (write $p \sim \mathsf{Unif}([0, 1])$.)

As with frequentist statistics, the *likelihood* functon is the density of the model for the observed variable evaluated at the data value given the parameter value treated as a function of the parameter value. So if the observation is X, and the density of X given parameter $\Theta = \theta$ is $f_X(x|\theta = t)$, then the likelihood function is $L(t|X = x) = f_X(x|\theta = t)$.

Once you have observed the data, you can update the distribution on θ using Bayes' Rule, which is also known as Bayes' Law or Bayes' Theorem.

Definition 26
Given a statistical model $[X|\theta]$ where X is the observed data and θ is a parameter of the model, the distribution of $[\theta|X]$ is called the **posterior**.

This is called the posterior because this is the distribution of Θ after taking data. Bayes' Rule allows us to calculate the posterior density given the prior density and the likelihood.

$$\text{posterior density} \propto \text{prior density} \cdot \text{likelihood}.$$

Fact 22 (Bayes' Rule for densities)
For a parameter θ and data X,

$$f_{\theta|X=x}(t) = C(x) f_\theta(t) f_{X|\theta=t}(x),$$

where

$$C(x) = \left[\int_{-\infty}^{\infty} f_\theta(t) f_{X|\theta=t}(x)\ dt \right]^{-1}.$$

Here t is a dummy variable for the random variable θ which is the parameter, and x is a dummy variable for the random variable X which is the data. Note that often the data is really $(X_1, \ldots, X_n)$ and so the dummy variable is $(x_1, \ldots, x_n)$.

Because t and x are dummy variables, oftentimes the rule is written without them:

$$f_{[\theta|X]} \propto f_\theta f_{[X|\theta]}.$$

In order to find the constant of proportionality, you have to integrate $f_\theta(t) f_X(x|\theta = t)$ with respect to t and set it equal to 1.

Notation alert! Bayesian statisticians often use a lowercase θ both to denote the random variable Θ, and as a dummy variable for Θ. So you see statements like

$$\mathbb{E}[\theta] = \int_{\mathbb{R}} \theta f_\theta(\theta)\ d\theta,$$

when the same statement using correct notation is

$$\mathbb{E}[\theta] = \int_{\mathbb{R}} t f_\theta(t)\ dt.$$

Often the following mnemonic is used, mathematically however, this notation doesn't make sense because it mixes up random variables and dummy variables. Here p stands for the posterior, π for the prior, and L for the likelihood):

$$p(\theta|X) \propto L(\theta|X)\pi(\theta).$$

I encourage you to remember Bayes' Rule as: posterior density is proportional to prior density times likelihood.

To summarize, we have the following pieces of a statistical framework.

θ	The parameter for the statistical model.
X	The data collected (often an n-dimensional vector.)
$[\theta]$	The distribution of the parameter, known as the prior.
$[X\|\theta]$	The distribution of the data given the parameter, called the *statistical model*.
$L(\theta) = f_{X\|\theta}(x)$	The density of the statistical model at $X = x$ viewed as a function of θ is called the likelihood.
$[\theta\|X]$	The distribution of the parameter given the data, called the *posterior*.

10.2. Examples of calculating the posterior

Qotd

- Start with prior, $p \sim \mathsf{Unif}([0,1])$, so $f_p(s) = \mathbb{1}(s \in [0,1])$.
- Next likelihood: let $X =$ number of successful patients. Then

  $$[X|p] \sim \mathsf{Bin}(10, p).$$

  So density of $X|p$ (with respect to counting measure) is

  $$f_X(i|p = a) = \binom{10}{i} a^i (1-a)^{10-i}.$$

- Together, give proportional to posterior:

$$f_p(a|X=i) \propto \mathbb{1}(a \in [0,1])\binom{10}{i}a^i(1-a)^{10-i}$$
$$f_p(a|X=3) \propto \mathbb{1}(a \in [0,1])a^3(1-a)^7.$$

  [Note only stuff that depends on a matters!]

- To find the normalizing constant:

$$\int_{\mathbb{R}} \mathbb{1}(a \in [0,1])a^3(1-a)^7 = \int_{a=0}^{1} a^3(1-a)^7 = 1/1320.$$

- Some of you might recognize this distribution! This is the density of a Beta random variable with parameters $4 = 3+1$ and $8 = 7+1$.

Fact 23 (Betas and binomial are conjugate)
Suppose $\theta \sim \mathsf{Beta}(a,b)$ and $[X|\theta] \sim \mathsf{Bin}(n,\theta)$. Then $[\theta|X] \sim \mathsf{Beta}(X+a, n-X+b)$.

- This is lucky! For some likelihoods and priors, the posterior comes from a known distribution.

Definition 27
Suppose the likelihood $[X|\theta]$ and prior $[\theta]$ is such that the distribution of $[\theta|X]$ is known. Then the prior and posterior distributions are called **conjugate priors**.

- For $[X|\theta] \sim \mathsf{Bin}(n,p)$, a beta family prior is conjugate to a beta family posterior. Such a family is *self-conjugate*.

- Conjugate priors can speed up Bayesian calculations, but you can *always* just use Bayes' Rule to find the posterior.

10.3. Point estimates from the posterior

- So now you have a posterior distribution, but your boss isn't interested in that, they want a single value for the parameter, a best guess.

- There are of course several ways to go from the posterior to a point estimate. 3 common ways
  - Posterior mean
  - Posterior mode
  - Posterior median (only works in 1-dimension)

- In Question of the Day, $[p|X=3] \sim \mathsf{Beta}(4,8)$. Looking up on Wikipedia the mean, mode, and median of a beta random variable:

$$\mathbb{E}[p|X=3] = \frac{4}{4+8} = 0.3333,$$
$$\text{mode}[p|X=3] = \frac{4-1}{4+8-2} = 3/8 = 0.3000$$
$$\text{median}[p|X=3] = \frac{4-1/3}{4+8-2/3} \approx 0.3235.$$

- Important note: there is no one right way to estimate p. As $n \to \infty$ (so the amount of data goes to infinity) they will all converge to the true answer, however.

Example: tracking a car at an intersection using GPS

- Suppose that a car approaches an intersection at $(0,0)$ traveling north. If it goes straight, it will move to position $(0,1)$, if it turns right, it will move to position $(1,0)$, and if it moves left, it will move to position $(-1,0)$.
- The GPS says that the car's position is $D = (-0.2, 0.7)$. The distance of the GPS position from the true position follows a Rayleigh distribution with density

$$f_R(r) = r\exp(-r^2/2)\mathbb{1}(r \geq 0).$$

  Did the car go straight? Turn left? Turn right?
- To use a Bayesian analysis, need prior probabilities that the car turned left, went straight, or went right. A study of the intersection indicates that 40% of drivers turned right, 40% went straight, and 20% turned left.
- Let $\theta \in \{\ell, r, s\}$ for left, right, and straight. Then

$$\begin{aligned}\|(-0.2,0.7)-(0,1)\| &= 0.360555\\ \|(-0.2,0.7)-(-1,0)\| &= 1.06301\\ \|(-0.2,0.7)-(1,0)\| &= 1.38924\end{aligned}$$

  So that means

$$f_{[\theta|D]} \propto f_\theta f_{D|\theta}.$$

  Hence

$$\begin{aligned}f_{[\theta|D]}(s) &\propto (0.4)(0.360555\exp(-0.36055^2/2)) \approx 0.1351457\\ f_{[\theta|D]}(\ell) &\propto (0.2)(1.06301\exp(-1.06301^2/2)) \approx 0.1208351\\ f_{[\theta|D]}(r) &\propto (0.4)(1.38924\exp(-1.38924^2/2)) \approx 0.2117121.\end{aligned}$$

- That means the normalizing constant is $0.1351457 + 0.1208351 + 0.2117121$, which gives:

$$f_{[\theta|D]} = (0.2889625, 0.2583641, 0.4526733).$$

  So based on this information, the best guess is that the car turned right! [Here the mode is the desired way to summarize the information.]
- Where did the Rayleigh distribution come from? That's the density of the distance when the error in the x and y directions are both normal random variables.

Problems

10.1: Fill in the blank: A Beta prior and Binomial likelihood gives an example of _____ priors.

10.2: A rate of typos in a series of plays by an author is modeled as having a prior $\mu \sim \mathsf{Exp}(0.1)$, so $f_\mu(s) = 0.1\exp(-0.1s)\mathbb{1}(s \geq 0)$. Given μ, the number of typos found in a given play is modeled as Poisson distributed with mean μ, so if T denotes the number of typos, for $i \in \{0, 1, 2, \ldots\}$

$$\mathbb{P}(T = i|\mu = s) = \frac{\exp(-s)s^i}{i!}.$$

(a) What is the posterior distribution of μ given T?

(b) If $T = 5$, what is the posterior mean?

10.3: Suppose I have statistical model $[X|\theta] \sim \mathsf{Exp}(\lambda)$, and a prior on λ of $\lambda \sim \mathsf{Unif}([1,3])$.

(a) Find the density

$$f_{\lambda|X_1,\ldots,X_n=x_1,\ldots,x_n}(t)$$

of the posterior up to an unknown normalizing constant.

(b) For data $1.3, 2.1, 1.7$, what is the posterior mode?

(c) For general data $x_1, \ldots, x_n$, what is the posterior mode?

Chapter 11

Confidence intervals

Question of the Day Give a 95% confidence interval for the mean of heights in a population.

In this chapter

- Confidence intervals
- Pivoting

Why intervals?

- Point estimates are never right for continuous models
- Suppose $X_1, X_2, \ldots, X_{10} \sim \mathsf{N}(0,1)$:

$$1.52, -0.690, 0.875, \ldots, 0.502.$$

  In which case $\hat{\mu} = 0.0475977 \neq 0 = \mu$.
- The probability that $\hat{\mu} = \mu$ is 0.
- Repeat experiment multiple times:

$$\hat{\mu}_1 = 0.475,\ \hat{\mu}_2 = 0.231,\ \hat{\mu}_3 = 0.458, \ldots \hat{\mu}_8 = -0.0793.$$

- A *coverage interval* $[a, b]$ attempts to choose a and b so that $\mu \in [a, b]$ some specified percentage of the time.
  - Frequentist: confidence intervals
  - Bayesian: credible intervals
- Start with confidence intervals: build statistics $A(D)$ and $B(D)$, such that for data D,

$$\mathbb{P}(\mu \in [A(D), B(D)] | \mu) \geq 0.95.$$

- Call *95%* the level of the confidence interval.

Definition 28
Suppose we have a statistical model where $X_1, X_2, \ldots \sim X$. Then statistics A and B form an **α-level confidence interval** for n samples if for all θ

$$\mathbb{P}(\theta \in [A(X_1, \ldots, X_n), B(X_1, \ldots, X_n)] | \theta) \geq \alpha.$$

Notes

- If 100 experiments each independently create their own confidence intervals, on average 95 of them will actually contain the true answer.
- The rest will be wrong.
- Impossible to tell which are right and which are wrong.
- No special reason to use 95% as confidence interval!
  - Seems close to 1.
  - Early tables used it.
  - No mathematical reason to use
- CERN uses $1 - 10^{-6} = 0.999999$ CI.
  - Very conservative
  - Much greater than 10^{-6} chance that statistical model is wrong.
- Medical testing 99% confidence interval.

Using the CLT

- What can we say about $\hat{\mu} - \mu$?
- Recall that
$$\hat{\mu} = \frac{X_1 + \cdots + X_{10}}{10}.$$
- The CLT indicates that sums of independent random variables are roughly normal. If $\mathbb{E}[X] = \mu$, $\mathrm{SD}(X) = \sigma$, then $\mathbb{V}(X_1/10) = \sigma^2/100$, and
$$\mathbb{E}\left(\frac{X_1}{n} + \cdots \frac{X_n}{n}\right) = \frac{\mu}{n} + \cdots + \frac{\mu}{n} = \mu,$$
and
$$\mathbb{V}\left(\frac{X_1}{n} + \cdots + \frac{X_n}{n}\right) = \mathbb{V}\left(\frac{X_1}{n}\right) + \cdots + \mathbb{V}\left(\frac{X_n}{n}\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n},$$
so by CLT,
$$\hat{\mu} \approx \mathsf{N}(\mu, \sigma^2/n).$$
Subtracting μ from both sides,
$$\hat{\mu} - \mu \approx \mathsf{N}(0, \sigma^2/n).$$
Here's a problem: we don't know σ typically! So approximate with $\hat{\sigma}^2$:
$$\hat{\mu} - \mu \approx \mathsf{N}(0, \hat{\sigma}^2/n).$$
Divide by $\hat{\sigma}/\sqrt{n}$ to get rid of variance and use symmetry of normal distribution
$$\frac{\mu - \hat{\mu}}{\hat{\sigma}/\sqrt{n}} \approx \mathsf{N}(0, 1).$$
Let $Z \sim \mathsf{N}(0, 1)$. Solving for μ:
$$\frac{\mu - \hat{\mu}}{\hat{\sigma}/\sqrt{n}} \approx Z$$
$$\hat{\mu} - \mu \approx Z\frac{\hat{\sigma}}{\sqrt{n}} \Rightarrow \hat{\mu} \approx \mu + Z\frac{\hat{\sigma}}{\sqrt{n}}.$$

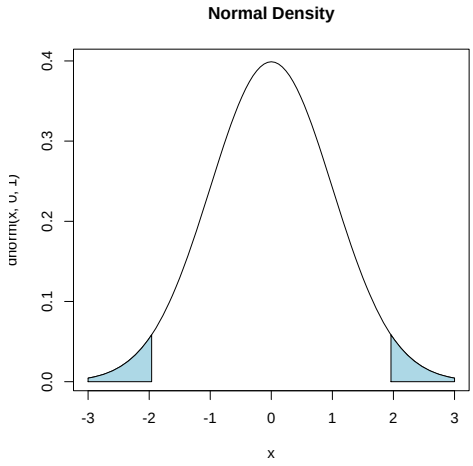


Figure 11.1: Shaded region is 5% of the probability under the normal curve

- Note that 2.5% of the time, $Z \leq -1.959964$ and 2.5% of time $Z \geq 1.959964$, so $\mathbb{P}(Z \in [-1.959964, 1.959964]) \approx 0.95$.
- So

$$\begin{aligned}\mathbb{P}\left(-1.95994 \leq Z \leq 1.95994\right) &\approx \mathbb{P}\left(-1.95994 \leq \frac{\mu - \hat{\mu}}{\hat{\sigma}/\sqrt{n}} \leq 1.95994\right) \\ &= \mathbb{P}\left(\hat{\mu} - 1.95994\frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \hat{\mu} + 1.95994\frac{\hat{\sigma}}{\sqrt{n}}\right)\end{aligned}$$

- Note: $z = 1.95994$ is sometimes called the *z-value* (or more specifically the two-sided z-value since the shaded area is broken into two sides) for the 95% confidence interval.
- How to find the 95% z-value in R: `qnorm(0.975)`. Using

  ```
  q[distribution name]
  ```

  gives inverse cdf functions, and this finds a such that $\mathbb{P}(Z \leq a) = 0.975$ (that way $1 - 0.975 = 2.5\%$ is above a. Note that since the normal distribution is symmetric, `qnorm(0.025)` gives $-a$.
- What command would you give R to find a 99% two sided confidence interval? What about a one sided (on the right) 95% confidence interval?

11.1. Pivoting

The *pivoting* method gives a way for turning a point estimate into a confidence interval.

Pivot This gives a general way to finding an interval for θ from an estimator $\hat{\theta}$.

Definition 29
Given a parameter θ and point estimate $\hat{\theta}$. Call $W = f(\theta, \hat{\theta})$ a **pivot** if the distribution of W is independent of θ.

1: Find or approximate the distribution of $W = f(\theta, \hat{\theta})$.

2: Find a and b such that $\mathbb{P}(a \leq W \leq b) = \alpha$.

3: Solve to write $\{a \leq W \leq b\} = \{a \leq f(\theta, \hat{\theta}) \leq b\} = \{a(\hat{\theta}) \leq \theta \leq b(\hat{\theta})\}$. Then

$$\mathbb{P}(a(\hat{\theta}) \leq \theta \leq b(\hat{\theta})) = \alpha.$$

Example: a uniform model Consider a statistical model where $X_1, X_2, \ldots \overset{\text{iid}}{\sim} X$ for

$$[X|\theta] \sim \mathsf{Unif}([0,\theta]).$$

That makes the density of X given θ

$$f_{X|\theta}(s) = \frac{1}{\theta}\mathbb{1}(s \in [0,\theta]).$$

Our $\hat{\theta}$ is the MLE (Maximum likelihood estimator)

$$\hat{\theta} = \max_{i\in\{1,2,\ldots,n\}} X_i.$$

Here's why: the likelihood of $X_1, \ldots, X_n$ given $\theta \geq \max X_i$ is the product of individual densities because the X_i are independent. That gives

$$f_{(X_1,\ldots,X_n)|\theta}(s_1,\ldots,s_n) = \prod_{i=1}^{n} \frac{1}{\theta}\mathbb{1}(s_i \in [0,\theta]) = \theta^{-n}\mathbb{1}((\forall i)(0 \leq s_i \leq \theta).$$

If $\theta < s_i$ for any i then this joint density is 0. But to make θ^{-n} as large as possible, θ needs to be as small as possible. Hence $\theta = \max s_i$ is the value of θ that maximizes f. In terms of the data,

$$\hat{\theta} = \max X_i$$

is the MLE.

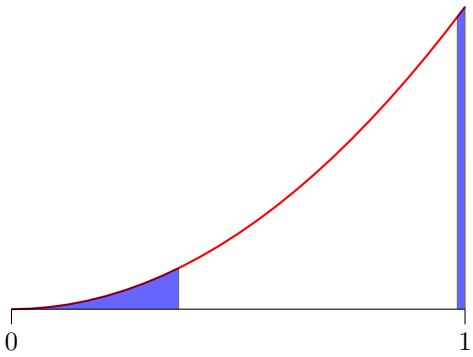
Followup question . Can you build a 90% confidence interval for θ using $\hat{\theta}$?

One thing to note is that the uniform distribution is scalable. That is, $X/\theta \sim \mathsf{Unif}([0,1])$. So

$$\frac{\hat{\theta}}{\theta} = \max\left\{\frac{X_1}{\theta}, \ldots \frac{X_n}{\theta}\right\},$$

which has distribution equal to the max of $U_1, \ldots, U_n$ where $U_i \overset{\text{iid}}{\sim} \mathsf{Unif}([0,1])$. This gives that $\hat{\theta}/\theta \sim \mathsf{Beta}(n,1)$.

That means that the distribution of $\hat{\theta}/\theta$ doesn't depend on θ at all!



5% in each shaded region

Also, $\mathbb{P}(\hat{\theta}/\theta \leq a) = a^n$ for $a \in [0,1]$. Hence

$$\mathbb{P}(\hat{\theta}/\theta \leq q^{1/n}) = q.$$

This gives

$$\mathbb{P}(0.05^{1/n} \leq \hat{\theta}/\theta \leq 0.95^{1/n}) = 0.9 \Rightarrow \mathbb{P}(0.05^{-1/n} \geq \theta/\hat{\theta} \geq 0.95^{-1/n}) = 0.9$$
$$\Rightarrow \mathbb{P}(\theta \in [(1/0.95)^{1/n}\hat{\theta}, (1/0.05)^{1/n}\hat{\theta}]).$$

Problems

11.1: Suppose that $X_1, X_2, \ldots, X_{10} \overset{\text{iid}}{\sim} X$, where $[X|\theta] \sim \mathsf{Unif}([0, \theta])$. What is

$$\mathbb{P}(2 \min_i X_i \leq \theta \leq 2 \max_i X_i)?$$

11.2: Dr. Pamela Isley measures the height of four plant samples, and finds them to be (in centimeters)

$$4.5, 3.7, 1.2, 6.2.$$

(a) Give an unbiased estimate of the mean height of the plants (including units).

(b) Give an unbiased estimate of the variance of the height of the plants (including units).

(c) Give a 90% z-value confidence interval for the mean plant height, using $\Phi(0.95) = 1.644854$.

11.3: Let $X_1, \ldots, X_n$ be modeled as iid draws from the uniform distribution on $[\theta, \theta + 1]$.

(a) What is the distribution of $X_i - \theta$? [You do not have to prove the result, simply give the distribution.]

(b) Show that $W = \bar{X} - \theta$ is a pivot.

Chapter 12

More Confidence intervals

12.1. Confidence intervals for population variance

Question of the Day [Example 6.4.2 Ramachandran & Tsokos] Suppose cholesterol levels (in mg/dL) of 10 patients are

$$360, 352, 294, 160, 146, 142, 318, 200, 142, 116.$$

Give a 95% confidence interval for σ^2.

In this chapter

- Confidence intervals for population variance

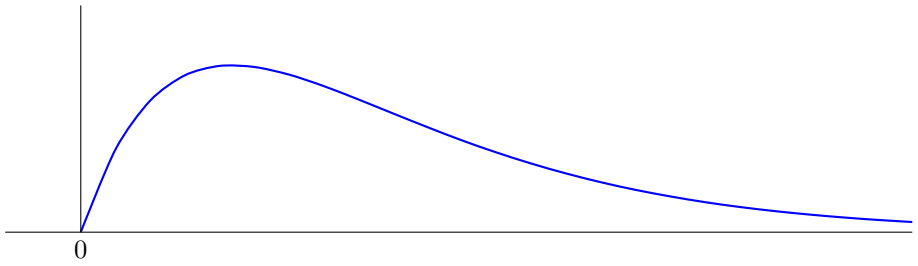
Need statistical model

- Recall that
$$\hat{\sigma}^2 = \frac{\sum(\bar{X} - X_i)^2}{n-1}$$
is an unbiased estimator of $\mathbb{V}(X)$.
- To make confidence interval, need model for how X, typically we use
$$X \sim \mathsf{N}(\mu, \sigma^2)$$
because it makes things easier to calculate.
- Need to know distribution of $[\hat{\sigma}^2 | \mu, \sigma]$.

Definition 30
If $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$, then $Z_1^2 + \cdots + Z_n^2 \sim \chi^2(n)$. [Read as chi-squared with n degrees of freedom.]

Fact 24 (Chi-squared distribution is also Gamma)
If $X \sim \chi^2(n)$, then $X \sim \mathsf{Gamma}(n/2, 1/2)$.

Proof outline. The joint distribution of two standard normals, (Z_1, Z_2), is rotationally symmetric. That means that if we take the point (Z_1, Z_2) and write it in polar coordinates, the angle from the x-axis is uniform from 0 to 360 degrees. Also, $Z_1^2 + Z_2^2$ has distribution of $\exp(1/2)$ (use polar transformation. Adding $n/2$ independent exponentials of rate $1/2$ together gives a gamma with parameters $n/2$ and $1/2$. □



Chi squared with 4 degrees of freedom

Remember that it our pivoting procedure, we want to find a random variable $W = f(\theta, \hat{\theta})$ such the distribution of W does not depend on θ. In the case of $\hat{\sigma}^2$, our pivot random variable is $W = (n-1)\hat{\sigma}^2/\sigma$.

Fact 25
Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$. Then $(n-1)\hat{\sigma}^2/\sigma^2$ has a distribution that does not depend on μ or σ.

Proof. Let $Z_1, \ldots, Z_n$ be standard normals, so $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$. Then $X_i = \mu + \sigma Z_i \sim \mathsf{N}(\mu, \sigma^2)$, and

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\mu + \sigma Z_i - \frac{(\mu + \sigma Z_1) + \cdots + (\mu + \sigma Z_n)}{n}\right)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(\sigma Z_i - \bar{\sigma}\bar{Z})^2 \\
&= \sigma\frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2
\end{aligned}$$

So $(n-1)\hat{\sigma}^2/\sigma^2$ has the same distribution as $\sum_{i=1}^{n}(Z_i - \bar{Z})^2$, where the Z_i are iid standard normals, and so do not depend in any way on μ or σ. □

Fact 26
For $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$, $(n-1)(\hat{\sigma}^2/\sigma^2) \sim \chi^2(n-1)$.

Given the previous fact, this proof is equivalent to showing that $\sum_{i=1}^{n}(Z_i - \bar{Z})^2 \sim \chi^2(n-1)$, which turns out to be true but very difficult to show. So here we will omit the proof.

- Can use this fact to pivot!

$$\mathbb{P}\left(\text{cdf}^{-1}_{\chi^2(n-1)}((1-\alpha)/2) \leq \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \leq \text{cdf}^{-1}_{\chi^2(n-1)}(\alpha + (1-\alpha)/2)\right) = \alpha$$

- Note that χ^2 not symmetric around 0, so left and right cdf value are not same in absolute value.
- Now pivot: solve for σ^2 inside of probability:

$$\mathbb{P}\left(\frac{\hat{\sigma}^2(n-1)}{\text{cdf}^{-1}_{\chi^2(n-1)}(1/2 + \alpha/2)} \leq \sigma^2 \leq \frac{\hat{\sigma}^2(n-1)}{\text{cdf}^{-1}_{\chi^2(n-1)}(1/2 - \alpha/2)}\right) = \alpha.$$

Fact 27
An α level confidence interval for σ^2 when $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$ is

$$\left[\frac{\hat{\sigma}^2(n-1)}{\text{cdf}^{-1}_{\chi^2(n-1)}(1/2+\alpha/2)}, \frac{\hat{\sigma}^2(n-1)}{\text{cdf}^{-1}_{\chi^2(n-1)}(1/2-\alpha/2)}\right]$$

Qotd

- Here $\alpha = 0.95$, so to get Chi-squared quantiles in R

```
a <- qchisq(0.025,df=9)
b <- qchisq(1-0.025,df=9)
x <- c(360,352,294,160,146,142,318,200,142,116)
m <- mean(x)
s <- sd(x)
(length(x) - 1)*s^2/a
(length(x) - 1)*s^2/b
```

  Gives $[4440, 31290]$ as interval (to 4 sig figs).

12.2. Confidence intervals for difference of two population parameters

- Consider data $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$.
- Then for $d = \mu_X - \mu_Y$, $\hat{d} = \bar{X} - \bar{Y}$ is the unbiased estimate of the difference between the two means. Find an α significance level confidence interval for d.
- Can't do without assuming a statistical model. Suppose

$$X_1, \ldots, X_n \sim \mathsf{N}(\mu_X, \sigma_X^2) \text{ and } Y_1, \ldots, Y_m \sim \mathsf{N}(\mu_Y, \sigma_Y^2).$$

- That makes $\bar{X} \sim \mathsf{N}(\mu_X, \sigma_X^2/n)$ and $\bar{Y} \sim \mathsf{N}(\mu_Y, \sigma_Y^2/m)$. So

$$\hat{d} \sim \mathsf{N}(\mu_X - \mu_Y, \sigma_X^2/n + \sigma_Y^2/m).$$

- Unfortunately, we don't know σ_X^2 or σ_Y^2, use our standard technique of approximating with $\hat{\sigma}_X^2$ and $\hat{\sigma}_Y^2$. Then scale and pivot to obtain the confidence interval.

Fact 28
For $X_1, \ldots, X_n \sim \mathsf{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_m \sim \mathsf{N}(\mu_Y, \sigma_Y^2)$ independent, let $\hat{d} = \bar{Y} - \bar{X}$ and

$$w = \text{cdf}^{-1}_{\mathsf{N}(0,1)}(1/2+\alpha/2)\sqrt{(\hat{\sigma}_X^2/n + \hat{\sigma}_Y^2/m)}.$$

Then

$$\mathbb{P}\left(\mu_Y - \mu_X \in \left[\hat{d} - w, \hat{d} + w\right]\right) = \alpha$$

Example

- A group of 13 patients on a control has a white blood cell count that averages 7824 per mm^3 with sample standard deviation of 2345. The 20 patients taking a drug has a sample mean of 5672 per mm^3 with sample standard deviation of 1834. Find a 99% confidence interval for the drug average minus the control average.

- Here $\hat{d} = 5672 - 7824 = -2152$. `qnorm(0.995)` gives 2.575829. Using the formula from above gives 1980.509.

$$w = 2.575829\sqrt{\frac{2345^2}{13} + \frac{1834^2}{20}} = 1980.509.$$

Therefore, to 4 sig figs the 99% CI is $[-4133, -171.4]$.

Problems

12.1: Suppose $X_1, \ldots, X_{10}$ are modeled as normal random variables with unknown mean μ and variance σ^2. What is the chance that the relative error in $\hat{\sigma}^2$ is greater than 10%? In other words, what is $\mathbb{P}(\hat{\sigma}^2 \geq 1.1\sigma)$?

$$\mathbb{P}(\hat{\sigma}^2 \geq 1.1\sigma) = \mathbb{P}(C \geq 9.9),$$

where $C \sim \chi^2(9)$. Using `1-pchisq(9.9,df=9)` then gives $\boxed{0.3586}$.

Chapter 13

Credible intervals

Question of the Day Bayesian statistical model: $X_1, X_2, \ldots \sim \mathsf{Unif}([0,\theta])$, $\theta \sim \mathsf{Exp}(0.01)$. If $(X_1, X_2, X_3) = (16.3, 48.8, 17.5)$, find a 90% credible interval for θ.

Let D be the data and a and b be functions of the data. Then recall that $[a(D), b(D)]$ is an α-level confidence interval for θ if

$$\mathbb{P}(a(D) \leq \theta \leq b(D)|\theta) = \alpha.$$

Note that here we are saying that no matter what the true value of θ is, conditioned on the value, data generated from the model when plugged into functions a and b will give an interval that contans the target parameter θ with probability α. Of course, the functions a and b are dependent on the statistical model $[D|\theta]$.

In Bayesian statistics, we model the parameter θ using a prior distribution $[\theta]$. Given the data that is drawn $[D|\theta]$, we can then use Bayes' Rule to build the posterior distribution $[\theta|D]$.

Then we can ask the question, are there functions a and b such that $\mathbb{P}(\theta \in [a(D), b(D)]|\theta) = \alpha$?

Definition 31
Given data D, statistical model $[D|\theta]$, and a prior on θ, an **α-level credible interval** $[a, b]$ is any choice of a and b such that $\mathbb{P}(\theta \in [a, b]|D) \geq \alpha$.

The big difference between confidence intervals and credible intervals:

- For confidence intervals, the probability statement holds when conditioning on the value of the parameter.
  - Only requires statistical model $[X|\theta]$.
- For credible intervals, the probability statement holds when conditioning on the value of the data.
  - Requires both statistical model $[X|\theta]$ and prior $[\theta]$.

There are some similarities between confidence and credible intervals. For both, there is typically more than one choice of functions a and b that work.

13.1. Equal tailed interval

The simplest choice of functions is to choose $a(D)$ and $b(D)$ in a balanced fashion.

Definition 32
Say that a credible interval is **equal tailed** or **balanced** if it is of the form $[a, b]$ where

$$\mathbb{P}(\theta \geq b) = \mathbb{P}(\theta \leq a) = (1 - \alpha)/2.$$

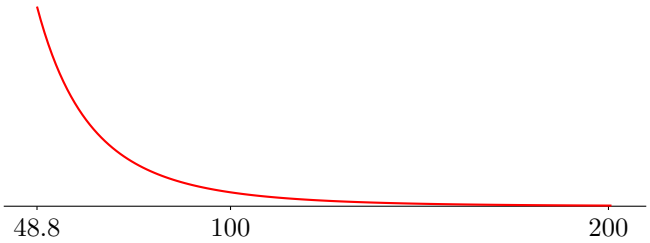
Solving the Question of the Day To find a credible interval, start by finding the posterior distribution $[\theta|X]$ (here $X = (X_1, \ldots, X_n)$.)

$$f_\theta(t|X = (x_1, x_2, x_3)) \propto f_\theta(t) f_X(x_1, x_2, x_3|\theta = t)$$

Since for a single data point, $[X_i|\theta] \sim \mathsf{Unif}([0, \theta])$, $f_{X_i|\theta=t}(x_i) = \frac{1}{t-0}\mathbb{1}(x_i \in [0, t])$. The prior density of $\theta \sim \mathsf{Exp}(0.01)$ is $f_\theta(t) = 0.01e^{-0.01t}\mathbb{1}(t \geq 0)$.

$$\begin{aligned} f_\theta(t|X = (x_1, x_2, x_3)) &\propto 0.01 \exp(-0.01t)\mathbb{1}(t \geq 0) \cdot \\ &\quad (1/t)(1/t)(1/t)\mathbb{1}(x_1 \in [0, t], \ldots, x_3 \in [0, t]) \\ &\propto [\exp(-0.01t)/t^3]\mathbb{1}(\max x_i \leq t). \end{aligned}$$

In other words, since each $X_i \in [0, \theta]$, we know that $\theta \geq X_i$ for all i, so $\theta \geq \max_i X_i$. But for $\theta \geq \max_i X_i$, the density of θ looks like $\exp(-0.01t)/t^3$. For the data, $\max\{16.3, 48.8, 17.5\} = 48.8$. Hence it looks as follows.



Now let us use WolframAlpha to integrate the posterior and find the normalizing constant.

```
integrate exp(-0.01t)/t^3 from 48.8 to infinity   0.0000947178  ll
```

Find lower end of interval with WolframAlpha by narrowing in:

```
integrate exp(-0.01t)/t^3/0.0000947178 from 48.8 to 60     0.438242
integrate exp(-0.01t)/t^3/0.0000947178 from 48.8 to 52     0.15983
integrate exp(-0.01t)/t^3/0.0000947178 from 48.8 to 49.8   0 .0538311
```

Recall

$$\frac{dy}{dx} = \frac{d}{dx}\int_{48.8}^{x} \exp(-0.01t)/t^3/0.0000947178\ dt = \exp(-0.01x)/x^3/0.0000947178.$$

So at $x = 49.8$, $dy/dx = 0.05195$. We want $dy = (0.053811 - 0.05) = 0.003811$ so $dx = dy/(dy/dx) = 0.003811/0.05195 = 0.0733$. So try $49.8 - 0.0733 = 49.726$.

```
integrate exp(-0.01t)/t^3/0.0000947178 from 48.8 to 49.726   0.0499767
```
(should we round up?)
```
integrate exp(-0.01t)/t^3/0.0000947178 from 48.8 to 49.73    0.0501856
```

Now for the upper limit:

```
integrate exp(-0.01t)/t^3/0.0000947178 from 100 to infinity   0.115809
integrate exp(-0.01t)/t^3/0.0000947178 from 110 to infinity   0.0836595
integrate exp(-0.01t)/t^3/0.0000947178 from 120 to infinity   0.0615385
```

Same derivative approach:

$$\frac{dy}{dx} = 0.00184023,\ dy = (0.0615385 - 0.05) \Rightarrow dx = 8.36$$

```
integrate exp(-0.01t)/t^3/0.0000947178 from 128.6 to infinity   0.0478409
integrate exp(-0.01t)/t^3/0.0000947178 from 128 to infinity     0.0486724
integrate exp(-0.01t)/t^3/0.0000947178 from 127 to infinity     0.05000958
integrate exp(-0.01t)/t^3/0.0000947178 from 127.1 to infinity   0.0499513
```

Therefore, to 4 sig figs, the 95% equal tailed credible interval is

$$\boxed{[49.72,127.1]}$$

Check: `integrate exp(-0.01t)/t^3/0.0000947178 from 127.1 to infinity` returns 0.900385.

- Notice that we round here down at the lower end of the interval and up at the higher end to ensure at least 95% coverage.

These are the easiest to calculate, but there are other common ways to get α-level credible intervals.

13.2. Narrowest interval

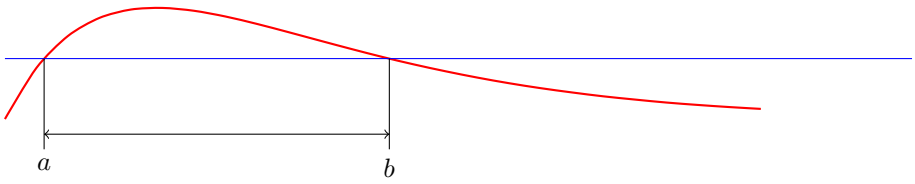
- Try to put as much probability as possible near high points of density.
- When unimodal density (one maximum), credible interval should contain maximum.
- For **qotd**, this means intervals of form $[48.8, b]$.
- The narrowest interval with 90% probability is:

$$\boxed{[48.80, 104.5]}.$$

- Width of narrowest interval: $104.5 - 48.80 = 55.70$, width of equal tailed interval: 77.38.

Ups and downs

- Suppose mode is in middle of range, not at end.
- Then narrowest interval of the form $[a, b]$ where mode is in $[a, b]$, and $f(a) = f(b)$.



Here $[a, b]$ is a candidate for the narrowest interval.

13.3. Centering at the posterior mean

When forming confidence interval, they often have the form (where $\hat{\mu}$ is a point estimate for μ):

$$[\hat{\mu} - \text{error}, \hat{\mu} + \text{error}].$$

Often, it is possible to make a credible interval of that form as well. On the other hand, it might not be possible to do this and stay in the range where the posterior density is positive.

The first two types of credible intervals that we talked about, equal tailed and narrowest, always exist. The posterior mean centered confidence interval might not.

Problems

13.1: Suppose a drug works with a probability p that is modeled as $\mathsf{Beta}(1, 9)$.

(a) What is the prior mean that the drug works?

(b) Suppose that 40 independent trials are run, in 13 of which the drug is a success. What is the posterior distribution of p given this data?

(c) Give a balanced two-tailed 95% credible interval for this data.

(d) Prove that your balanced interval is *not* the narrowest interval.

Chapter 14

Nonparametric point and interval estimates

Question of the Day Counts of a water lily at four locations are 34, 24, 71, 48. Estimate the median of the water lily count distribution.

In this chapter

- Nonparametric point estimates
- Nonparametric interval estimates

So far we have been taking advantage of the strong law of large numbers to build consistent estimators. This has a couple drawbacks. First, the statistical model for our data might not have a mean, in which case the data does not converge. Second, the standard deviation might be very high, in which case the confidence intervals formed from the estimate might converge very slowly or not at all.

The canonical example of a random variable that does not have a mean is $X \sim \mathsf{Cauchy}(0)$ with density $f_X(s) = \frac{2}{\tau(1+s^2)}$. Because this random variable has no mean, if $X_1, X_2, \ldots \sim X$, and $\hat{\mu}_n$ is the sample average of the first n data points, then $\hat{\mu}_n$ will not converge to anything, no matter how large n gets!

The other problem is that perhaps the mean exists, but there is an extremely small chance of a very large data point. Then the overall mean will converge, but very slowly.

14.1. Nonparametric methods

Nonparametric methods solve these problems by not assuming that the data comes from a particular distribution ahead of time. That is, there is no statistical model giving rise to a likelihood. These methods have several good properties.

- Does not require finite mean
- Very robust to outliers
- Because they assume so little, no problems if the data comes from a different distribution than expected.

On the other hand, there are downsides to these methods as well.

- Problem. because assume so little, the intervals tend to be wider than with other approaches. Typically

$$\text{Bayesian interval} \subset \text{Frequentist interval} \subset \text{Nonparametric interval}$$

14.2. Nonparametric Point estimates

Definition 33
The **median** of a random variable X is any value a such that $\mathbb{P}(X \leq a) \geq 1/2$ and $\mathbb{P}(X \geq a) \geq 1/2$. The median of a distribution is the median of a random variable with that distribution.

- Ex: If $X \sim \mathsf{Unif}([0, 10])$, median$(X) = 5$.
- Ex: If $X \in \{1, 2, 3, 4, 5, 6\}$, then median$(X) = [3, 4]$.
- Note, every random variable has at least one median. [Not all random variables have a mean.]

Definition 34
Let $(X_1, \ldots, X_n)$ be a vector of n values. Let f be a permutation of $\{1, 2, \ldots, n\}$ so that

$$X_{f(1)} \leq X_{f(2)} \leq \cdots \leq X_{f(n)}.$$

Then let $X_{(i)} = X_{f(i)}$, and call this value the ***i*th order statistic** of $(X_1, \ldots, X_n)$.

- Ex: for numbers 34, 24, 71, 68, the order statistics are $X_{(1)} = 24$, $X_{(2)} = 34$, $X_{(3)} = 68$, and $X_{(4)} = 71$.

Definition 35
The **sample median** of a set of data $(x_1, \ldots, x_n)$ is

$$\text{median}(x_1, \ldots, x_n) = \frac{x_{(\lfloor (n+1)/2/\rfloor)}, x_{(\lfloor (n+2)/2\rfloor)}}{2}.$$

To find the sample median

1: Sort the data: $(X_1, \ldots, X_n) \to (X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ [Recall the $X_{(i)}$ are called *order statistics*.]

2: Sample median is $\hat{m} = (X_{(\lfloor (n+1)/2\rfloor)} + X_{\lfloor (n+2)/2\rfloor})/2$.

- Estimate the median of $34, 24, 71, 48$.
  - First sort: $24 < 34 < 48 < 71$.
  - Average middle two values: $(34 + 48)/2 = 41$.
- Estimate the median of $34, 24, 71, 48, 61$.
  - First sort: $24 < 34 < 48 < 61 < 71$.
  - Average middle two values: $(48 + 48)/2 = 48$. (Same as picking middle value.)
- Note, for $X_1 = (34, 24, 71, 48)$, $\hat{\mu}_1 = 44.25$, but for $X_1 = (34, 24, 7100, 48)$, $\hat{\mu}_2 = 1801.5$.
- But median$(34, 24, 71, 48) =$ median$(34, 24, 7100, 48) = 41$.
- Because the sample median is unaffected by outliers, say the the estimate is *robust*.

14.3. Nonparametric Confidence Intervals

- Now let's try to build a confidence interval for the median.
- Suppose that we draw 9 values $X_1, X_2, \ldots, X_9$, sort them to get $X_{(1)}, X_{(2)}, \ldots, X_{(9)}$. What is a good confidence interval for $\hat{\mu} = X_{(5)}$?
- Idea: Use intervals of the form $[X_{(a)}, X_{(b)}]$. For instance, what is the probability that $m \in [X_{(3)}, X_{(7)}]$? or $m \in [X_{(2)}, X_{(8)}]$?

- What has to happen for $m \notin [X_{(2)}, X_{(8)}]$. Either $m > X_{(8)}$, or $m < X_{(2)}$. The chance that $X_{(8)} < m$ is the chance that of 9 draws, at least 8 were smaller than m.
- Let $N = \#\{i : X_i < m\}$. Then for continuous X_i,

$$N \sim \mathsf{Bin}(n, 1/2).$$

- So

$$\begin{aligned}\mathbb{P}(N \leq 1 \text{ or } N \geq 8) &= 2\mathbb{P}(N \leq 1) = 2[\mathbb{P}(N = 0) + \mathbb{P}(N = 1)] \\ &= 2\left((1/2)^9 + \binom{9}{1}(1/2)^9\right) \\ &= 0.0390625.\end{aligned}$$

- This makes $[X_{(2)}, X_{(8)}]$ a $1 - 0.0390625 = 0.9609375\%$ confidence interval for the median.
- A similar calculation gives $[X_{(3)}, X_{(8)}]$ is a 82.03% CI, and $[X_{(1)}, X_{(9)}]$ is an 99.60% CI.

14.4. Exact confidence intervals

- Suppose that I want a 95% confidence interval for the median.
- For 9, $[X_{(2)}, X_{(8)}]$ is slightly too big 96.09375, $[X_{(3)}, X_{(7)}]$ is too small at 82.03125.
- Mix the two options to get an exact CI
- Note that 95% is a *convex linear combination* of 96.09% and 82.03%. That is, $(\exists p \in [0,1])(p(96.09375) + (1-p)(82.03125) = 0.95)$. Solving gives $p \approx 0.922222$.
- So with probability 92.22222% percent, use $[X_{(2)}, X_{(8)}]$ as your confidence interval, otherwise use $[X_{(3)}, X_{(4)}]$.
- Recall that Conf. Int. for mean require knowledge of distribution of data.
- Recall that Cred. Int. require knowledge of dist. of data + prior for parameter.
- Conf. Int. for median requires no prior, no knowledge of dist. of data.
- Typically wider than Conf. Int. or Cred. Int.

Exact confidence intervals for $\mathsf{Exp}(\lambda)$ data

- Recall that if you scale an exponential random variable, the result is just an exponential random variable with different rate parameter:

Fact 29 (Exponential facts)
The following are some useful facts about exponential random variables.

1: Let $X \sim \mathsf{Exp}(\lambda)$ and $c \in \mathbb{R}$. Then $cX \sim \mathsf{Exp}(\lambda/c)$.

2: Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Exp}(\lambda)$. Then $X_1 + \cdots + X_n \sim \mathsf{Gamma}(n, \lambda)$.

3: Let $X \sim \mathsf{Gamma}(n, \lambda)$ and $c \in \mathbb{R}$. Then $cX \sim \mathsf{Gamma}(n, \lambda/c)$.

- Let $X_1, X_2, \ldots, X_n \sim \mathsf{Exp}(\lambda)$.
- Probability fact: $X_1 + X_2 + \cdots + X_n \sim \mathsf{Gamma}(n, \lambda)$.

- So
$$\frac{X_1 + \cdots + X_n}{n} \sim \mathsf{Gamma}n, n\lambda.$$
- So let
$$\hat{\lambda} = \frac{n-1}{X_1 + X_2 + \cdots + X_n} \sim \mathsf{InvGamma}(n, (n-1)\lambda).$$
Then
$$\frac{\hat{\lambda}}{\lambda} = \frac{n-1}{\lambda X_1 + \cdots + \lambda X_n} \sim \mathsf{InvGamma}(n, n-1),$$
which doesn't depend on λ!
- Why $n-1$ rather than n? So $\mathbb{E}[\hat{\lambda}] = \lambda$

Example

- Suppose the model for data $2.3, 0.7, 1.7$ is $\mathsf{Exp}(\lambda)$. Find a 99% confidence interval for λ with equal tails using $\hat{\lambda}/\lambda$ to pivot.
- First find $\hat{\lambda}$:
$$\hat{\lambda} = \frac{3-1}{2.3 + 0.7 + 1.7} = 0.4255\ldots.$$
Now for CI:
$$\mathbb{P}\left(a \le \frac{\hat{\lambda}}{\lambda} \le b\right) = 0.99 \Leftrightarrow \mathbb{P}\left(a^{-1} \ge \frac{\lambda}{\hat{\lambda}} \ge b^{-1}\right) = 0.99.$$
Since $\lambda/\hat{\lambda} \sim \mathsf{Gamma}(n, n-1)$, use in R
```
qgamma(0.005,shape=3,rate=2)
qgamma(0.995,shape=3,rate=2)
```
which gives $b^{-1} = 0.1689317$ and $a^{-1} = 4.636896$, so
$$\boxed{[0.07188, 1.974]}$$
as the 99% confidence interval.

Problems

14.1: For the distribution $\mathsf{Unif}([0, \theta])$, find the median as a function of θ.

14.2: (a) Find the sample median of $\{1.2, 7.6, 5.2\}$.
(b) Find the sample median of $\{3.4, 2.3, 7.3, 5.0\}$.

Chapter 15

Statistical Modeling

Question of the Day What makes a good model?

In this chapter

- The language of models

What is a model?

- One way is to think of a model as a function $f : \mathbb{R}^m \to \mathbb{R}^n$. The m inputs to the function are called *explanatory variables* and the n outputs are called *response variables*.
- For instance, the amount of time I run my air conditioner affects my electric bill. The interest rate set by the Fed affects the inflation rate.
- No model is perfect however. The difference between the model value $f(e_1, \ldots, e_m)$ and the true values of the response variables are called *residuals*.

15.1. What makes a good model?

- Keep the number of explanatory variables small (KISS).
- Keep the residuals small.
- Good ability to predict how response variables change as explanatory variable change.

15.2. Notation for models

Notation 5
Statisticians use the following notation for describing models. Put the response variables on the left of a tilde symbol, $\sim$, the explanatory variables on the right, use plus signs if more than experimental vararіable affects the model, and use a colon to indicate terms that interact within the model.

- Example:
$$\text{wage} \sim 1 + \text{gender}$$
indicates that the average wage paid depends on a constant term, and upon the gender of the person.
- Example:
$$\text{wage} \sim 1 + \text{gender} + \text{education} + \text{gender} : \text{education}$$
Now we've added another explanatory variable, education, which also interacts with gender.

- The $\sim$ notation is often used in R as well! For instance, the command in R

```
boxplot(wage ~ gender,data=example)
```

  will plot the `wage` values against the `gender` values in a dataset named `example`.

Explanatory and Response Variables

- Note that explanatory variables do *not* necessarily "cause" the response variable.

$$\text{age} \sim 1 + \text{wrinkles}.$$

  Wrinkles do not cause aging. In fact, one could say that aging causes wrinkles. Just because you can fit a model doesn't mean that you have proved something causes something else!

- Model stock prices:

$$\text{price} \sim 1 + \text{time}$$

  Time doesn't *cause* the stock price, but over time the stock price might be rising or falling in a predictable way.

- Two main types of variables. Quantitative variables assign a number to the variable
  - Height, GPA, # of insect bites

  Categorical variables put the data point into a category:
  - Gender, Spray

15.3. Linear models

- One response, one explanatory variable:

$$y = mx + b.$$

- Example: from Galton's height data:

$$\begin{aligned}\text{height} &\sim 1 + \text{mother}\\ \text{height} &= 46.7 + 0.313\ \text{mother}.\end{aligned}$$

- A more sophisticated model adds father's height and interaction:

$$\begin{aligned}\text{height} &\sim 1 + \text{mother} + \text{father} + \text{father} : \text{mother}\\ \text{height} &= 132.3 - 1.43\ \text{mother} - 1.21\ \text{father} + 0.0247\ \text{father} \cdot \text{mother}.\end{aligned}$$

  Note: 3 explanatory variables here: father, mother, and father$\times$mother.

Definition 36
A **linear model** with n explanatory variables, the prediction is a fixed linear combination of the values of the explanatory variables plus a constant term.

- Suppose there are k explanatory variables and m response variables. Each of these variables is measured n times. Then the model is

$$Y = X\beta + \epsilon.$$

  Here Y is a column matrix of length n, β is an $k \times 1$ vector of coefficients for the different explanatory variables, X is an n by k matrix and ϵ is a $k \times 1$ vector is independent draws from the residual distribution.

Example: for the Galton height model with constant, mother's height, father's height, mother-father interaction, the first column of X would be a constant (for the constant term), the second is the mother's height, the third column is the father's height, and the fourth column the product of the heights. The least square β values are

$$\beta = (132.3, -1.21, -1.43, 0.0247)^T.$$

The first five families in Galton's height data (where Height is the mean of the children) are:

Family	Father	Mother	Height
1	78.5	67.0	70.1
2	75.5	66.5	69.25
3	75.0	64.0	69.5
4	75.0	64.0	67.625
5	75.0	58.5	65.7

- Which makes Y and X the following:

$$Y = \begin{pmatrix} 70.1 \\ 69.25 \\ 69.5 \\ 67.625 \\ 65.7 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 78.5 & 67.0 & 5259.50 \\ 1 & 75.5 & 66.5 & 5020.75 \\ 1 & 75.0 & 64.0 & 4800.00 \\ 1 & 75.0 & 64.0 & 4800.00 \\ 1 & 75.9 & 58.5 & 4387.50 \end{pmatrix}$$

- That makes the residuals:

$$\epsilon = Y - X\beta = \begin{pmatrix} -1.612 \\ -0.9025 \\ 0.6191 \\ -1.255 \\ 0.2237 \end{pmatrix}$$

Linearity

- Note that the response is a linear function of X, but that X can be quadratic or any other function. For instance, suppose that the model is

$$y_i = c_0 + c_1 x_i + c_2 x_i^2 + \epsilon_i.$$

Then X could be:

$$\begin{pmatrix} 1 & 3 & 9 \\ 1 & -2 & 4 \\ 1 & 1 & 1 \end{pmatrix}$$

where the second column contains the x_i terms and the third column contains the x_i^2 terms. In other words, the explanatory variables do *not* have to be independent of each other.

15.4. Modeling residuals

- There is no one right way to model the residuals.
- Since $\epsilon_i = y_i - \mu - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik}$, CLT considerations make a normal distribution a popular model. Since the X typically includes a first column of 1's for the constant term, the mean of the residuals can be taken to be 0. This gives:

$$\epsilon_i \sim \mathsf{N}(0, \sigma_\epsilon^2)$$

as the model for the residuals.

- Hence the density of residuals is:

$$f_{(\epsilon'_1,\ldots,\epsilon'_n)}(a_1,\ldots,a_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left(\frac{-a_i^2}{2\sigma_\epsilon^2}\right).$$

Problems

15.1: Fill in the blank: $Y = X\beta + \epsilon$ where X is an m by k matrix, β is a k by 1 column vector, is a _____ model.

Chapter 16

MLE for linear models with normal residuals

Question of the Day Suppose that the residuals ϵ are normally distributed. What is the MLE for β given $Y = X\beta + \epsilon$?

In this chapter

- Least squares
- The pseudoinverse

Linear model

- Recall: A linear model has the form
$$Y = X\beta + \epsilon,$$
where Y is $n \times 1$, X is $n \times k$, β is $k \times 1$ and ϵ is $n \times 1$.
- So for a given μ and β, and data Y and explanatory variables X,
$$\epsilon = Y - X\beta.$$
- Each $\epsilon_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma_\epsilon^2)$, so the likelihood is
$$L(\sigma_\epsilon^2 | a) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left(\frac{-a_i^2}{2\sigma_\epsilon^2}\right).$$
so the log-likelihood is:
$$\begin{aligned}\ln(L(\sigma_\epsilon^2|\epsilon)) &= \sum_{i=1}^{n} -(1/2)\ln(2\pi\sigma_\epsilon^2) - a_i^2/(2\sigma_\epsilon^2) \\ &= -n\ln(\sigma_\epsilon) - \frac{1}{2\sigma_\epsilon^2}\sum_i a_i^2 - (n/2)\ln(2\pi).\end{aligned}$$
Call the right hand side $f(\sigma_\epsilon)$. Then
$$\begin{aligned}f'(\sigma_\epsilon) &= -\frac{n}{\sigma_\epsilon} + \frac{1}{\sigma_\epsilon^3}\sum_i a_i^2 \\ &= -\frac{n}{\sigma_\epsilon}\left[1 - \frac{1}{n\sigma_\epsilon^2}\sum_i a_i^2\right].\end{aligned}$$
This means the derivative is

- Positive when $\sigma_\epsilon^2 < n^{-1}\sum_i a_i^2$.
- Zero when $\sigma_\epsilon^2 = n^{-1}\sum_i a_i^2$.
- Negative when $\sigma_\epsilon^2 > n^{-1}\sum_i a_i^2$.

So the maximum of f occurs when $\sigma_\epsilon^2 = n^{-1}\sum_i a_i^2$.

This says that no matter what the a_i are, the likelihood is maximized when $\sigma_\epsilon^2 = n^{-1}\sum_i a_i^2$. But the a_i are not chosen by us, they are a function of the data and the values of the β variables. That is, $a_i = y_i - x_i\beta$, where y_i is the ith observation and x_i is the ith vector of explanatory variable values. Using the best possible choice of σ_ϵ^2 makes

$$\ln(L(\beta)) = -(n/2)\ln\left(n^{-1}\sum_{i=1}^{n} a_i^2\right) - (n/2) - (n/2)\ln(2\pi).$$

This is maximized when the β values are chosen so that

$$\sum_{i=1}^{n} a_i^2$$

is as small as possible. This MLE then gives the *least squares method*.

Definition 37
The **least squares** choice of β minimizes $\sum_i \epsilon_i^2$, where $\epsilon = Y - X\beta$.

The discussion above proved the following fact.

Fact 30 (Least-squares is MLE for normal residuals)
The MLE for the linear model $Y = X\beta + \epsilon$, where the residuals are normal, occurs at the least squares choice of β.

To find the choice of β that gives the least squares of the residuals, we'll the need the following linear algebra fact, where X^T denote the transpose of the matrix X.

Fact 31
For matrices A and B with compatible dimensions,

$$[AB]^T = B^T A^T.$$

Another way to look at it, is that if there was no random variation in the data we could solve $Y = X\beta$ exactly. Since there is variation in the data, we can only get $X\beta$ close to Y.

Here close is measured by minimizing the sum of the squares of the residual, or in mathematical terms, the L_2-norm of the residuals:

$$\|Y - X\beta\|_2 = \sqrt{\sum_i \epsilon_i^2}.$$

Recall that the length $\|x\|$ of a matrix can be found as

$$\|x\|_2^2 = x^T x, \quad \begin{pmatrix}3 & 1 & 2\end{pmatrix}\begin{pmatrix}3\\1\\2\end{pmatrix} = 3^2 + 1^2 + 2^2.$$

When applied to the residuals, we get:

$$\begin{aligned}\|\epsilon\|_2^2 &= (Y - X\beta)^T(Y - X\beta)\\ &= Y^TY - Y^TX\beta - (X\beta)^TY + (X\beta)^T(X\beta)\\ &= Y^TY - Y^TX\beta - \beta^TX^TY + \beta^TX^TX\beta\end{aligned}$$

Each of these four terms in the sum is a 1×1 matrix, that is, it is a real number. Since the third term is the transpose of the second term, it must be that $-Y^T X\beta = -\beta^T X^T Y$. Hence

$$\|\epsilon\|_2^2 = f(\beta) = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta$$

So $f(\beta)$ is the sum of a constant ($Y^T Y$), a linear form ($-2Y^T X\beta$), and a quadratic form ($\beta^T X^T X\beta$). To minimize this function, we need some help from Multivariable Calculus.

16.1. Derivatives in Multivariable Calculus

For a one dimensional vector $f \in C^1$ (so the first derivative is continuous), if $f'(x) = 0$, we call x a *critical point*. If $f''(x) \leq 0$ and $f'(x) = 0$ then $(x, f(x))$ is a *local minimum*. If there is only one local minimum, then it is a *global minimum*.

If the function is linear, so $f(x) = c_1 x$ for a constant c_1, then $f'(x) = c_1$. If the function is quadratic, $f(x) = c_1(x - c_2)^2$, then $f'(x) = 2c_1(x - c_2)$, and $f''(x) = 2c_1$.

Now let's move up to a function of n variables, $f(\beta_1, \ldots, \beta_n)$. Here the first derivative is called the *gradient*, and is

$$\nabla(f) = \left(\frac{\partial f}{\partial \beta_1}, \ldots, \frac{\partial f}{\partial \beta_n}\right)^T.$$

For example, $f(a_1, a_2, a_3) = 3a_1 - 2a_2 + 6a_3$, then $\nabla f = (3, -2, 6)^T$.

A *critical point* is any place where the gradient of f evaluates to the 0 vector. The equivalent of the second derivative is called the *Hessian*, and is:

$$H(f(x)) = \begin{pmatrix} \frac{\partial f}{\partial \beta_1 \partial \beta_1} & \frac{\partial f}{\partial \beta_2 \partial \beta_1} & \cdots & \frac{\partial f}{\partial \beta_n \partial \beta_1} \\ \frac{\partial f}{\partial \beta_1 \partial \beta_2} & \frac{\partial f}{\partial \beta_2 \partial \beta_2} & \cdots & \frac{\partial f}{\partial \beta_n \partial \beta_2} \\ & & \vdots & \\ \frac{\partial f}{\partial \beta_1 \partial \beta_n} & \frac{\partial f}{\partial \beta_2 \partial \beta_n} & \cdots & \frac{\partial f}{\partial \beta_n \partial \beta_n} \end{pmatrix}$$

A matrix A is *positive definite* if for all nonzero vectors $(\beta_1, \ldots, \beta_n)$, $\beta^T A\beta > 0$.

A value β that is both a critical point and a point where the Hessian is negative definite is a *local minimum*, and if there is a unique local maximum over all values of β, it must be a *global minimum*.

Call $f(\beta_1, \ldots, \beta_n)$ a *linear form* if

$$f(\beta) = w^T\beta = (w_1, \ldots, w_n)\beta = \sum_{i=1}^{n} w_i\beta_i.$$

For linear forms:

$$\nabla f(\beta) = (w_1, w_2, \ldots, w_n)^T = w.$$

Recalling our earlier example, $f(a_1, a_2, a_3) = 3a_1 - 2a_2 + 6a_3$ is a linear form, and $\nabla f = (3, -2, 6)^T$.

Because all the second partial derivatives of a linear form will be 0, the Hessian is just the matrix of all zeros.

Next, call $f(\beta)$ a *quadratic form* if it has the form:

$$f(\beta) = \sum_{i,j} \beta_i A_{i,j} \beta_j = \beta^T A\beta.$$

The gradient of a quadratic form also is not too bad, if we consider $\partial\beta^T A\beta/\partial\beta_i$ we can calculate it to be $2\sum_j A_{i,j}\beta_j$, and so altogether,

$$\nabla(\beta^T A\beta) = 2A\beta.$$

For example, if

$$g(a_1, a_2) = \begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} 3 & -1 \\ -1 & 7 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = 3a_1^2 - 2a_1a_2 + 7a_2^2,$$

$$\nabla g = (6a_1 - 2a_2, -2a_1 + 14a_2)^T = 2\begin{pmatrix} 3 & -1 \\ -1 & 7 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}.$$

The Hessian of a quadratic form $\beta^T A\beta$ is just $2A$. Continuing our earlier example:

$$H(3a_1^2 - 2a_1a_2 + 7a_2^2) = \begin{pmatrix} 2(3) & 2(-1) \\ 2(-1) & 2(7) \end{pmatrix} = 2A.$$

It is also important to note that both the gradient and Hessian are linear operators, so

$$\nabla(c_1 f + c_2 g) = \nabla(c_1 f) + \nabla(c_2 g),\ H(c_1 f + c_2 g) = H(c_1 f) + H(c_2 g).$$

Now we can use these facts to minimize

$$f(\beta) = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta.$$

First the gradient:

$$\nabla(f(\beta)) = (0, \ldots, 0) - 2X^T Y + 2X^T X\beta.$$

Set equal to zero to find the critical points:

$$-2X^T Y + 2X^T X\beta = 0.$$

so $X^T X\beta = Y^T X$, and for $X^T X$ invertible,

$$\beta = (X^T X)^{-1} X^T Y$$

is the unique critical point.

What is the Hessian at this point? Well,

$$H(f(\beta)) = 2X^T X,$$

for all β, so the Hessian is either positive definite everywhere, or nowhere.

Fact 32
If X is an $n \times k$ matrix where $n > k$, and X has rank k, then $X^T X$ is positive definite (and so is invertible.)

Proof. Let v be a nonzero matrix. Then Xv is a nonzero linear combination of the columns of the matrix X, and X has full column rank, so they must be linearly independent. That means $Xv \neq 0$. Hence $\|Xv\| = (Xv)^T(Xv) = v^T X^T Xv \neq 0$. □

Definition 38
For an $n \times k$ matrix X of rank k, the matrix $(X^T X)^{-1} X^T$ is called the **pseudoinverse** of X.

Theorem 5 (Linear Regression)
Consider the linear model $Y = X\beta + \epsilon$ where the components of ϵ are normal with equal variance. If X has full column rank, then the unique maximum likelihood estimator is

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Example Let's try this on some data in R. To build a matrix from columns, use the `cbind` command. To take the transpose of a matrix `X`, use `t(X)`. To multiply matrices `A` and `B`, use `A %*% B`. To find A^{-1}, use `solve(A)`. Putting this together, we can analyze some data of eruption height versus time waiting for the eruption for Old Faithful geyser in Yellowstone National Park with:

```
data(faithful)
head(faithful)
X <- cbind(c(rep(1,nrow(faithful))),faithful$eruptions)
head(X)
Y <- cbind(faithful$waiting)
solve(t(X) \%*\% X) \%*\% t(X) \%*\% Y
```

The result is $\beta = (33.47, 10.72)$, so our fit is that

$$y_i = 33.47 + 10.72x_i$$

Of course, R has commands that do all this automatically. Use

```
summary(lm(waiting ~ eruptions,data=faithful))
```

to get the coefficients, along with much more information that we will delve into in later chapters.

16.2. Why is it called linear regression?

- Sir Francis Galton predicted that if a short mother and tall father (or tall mother and short father) had a child, the height would be closer to the mean height.
- This behavior Galton called *regression to the mean*.
- He initially believed that after many generations, a population would completely regress and all be the same height!
- What he didn't know was that our genes our discrete, not continuous, so the average height can never fully converge. Moreover, random effects of nutrition and diet continue to add random effects to adult height.
- Even though regression to the mean never fully occurs, we still call fitting a linear model *linear regression*.

Problems

16.1: The form $Y = X\beta + \epsilon$ is what kind of model?

16.2: Consider some data from Old Faithful geyser showing the length of the eruption together with the waiting time until the next eruption (both measured in minutes.)

3.600	79
1.800	54
3.333	74
2.283	62
4.533	85
2.883	55

We wish to fit a model where the waiting times y_i are predicted by the eruption length x_i using constant, linear, and quadratic terms. So

$$y_i = c_0 + c_1 x_i + c_2 x_i^2 + \epsilon_i$$

(a) What is the vector Y in $Y = X\beta + \epsilon$?

(b) What is the matrix X in $Y = X\beta + \epsilon$?

(c) Using numerical software, find the pseudoinverse of X.

(d) What is the vector β in $Y = X\beta + \epsilon$?

(e) What is the maximum likelihood estimate $\hat{\beta}$ for β?

(f) What is the estimate of the residuals $Y - X\hat{\beta}$?

Chapter 17

Hypothesis testing

Question of the Day Suppose $(0.23, 0.17, 0.42, 0.34)$ oz. are the weight gains of 4 rats on an experimental diet. Should we reject the hypothesis that the weight gain is at most 0.1 ounces?

In this chapter

- Popper and philosophy of science
- The null hypothesis
- Type I error
- Rejecting the null hypothesis at a level

17.1. Popper and falsifiability

Karl Popper was perhaps the most influential philosopher of science of the 20th century. He attempted to tackle the Demarcation problem: What is science? What makes one set of experiments, theories, and facts science, whereas another set is not?

For instance, why is astronomy considered a science, where astrology is not? Both have complex theories, and both make predictions. However, we say that astronomy is supported by evidence while astrology is not. How do we make such decisions?

This has been answered in various ways throughout history. Francis Bacon proposed a framework where we begin by observing nature, then we propose a law, we confirm the law in action through more observations (or discard if we fail), generalize the law, take more observations, and so on. So for Bacon, science was a continually evolving thing that took more and more information into account.

The difficulty with a purely observational approach is that it is difficult to distinguish causation from correlation. The cannonical example in statistics is the observation that children with big feet tend to be better at spelling. Of course, the foot size does not cause the increase in spelling ability, it is simply the fact that children that are older tend both to have bigger feet and better spelling ability.

Karl Popper followed this framework to its logical endpoint, and proposed that it is impossible for science to ever prove that something is ever true, instead, it is only capable of falsifying ideas. What makes something a scientific statement is that it is falsifiable.

For example, "All swans are white" can be proved false by an observation of a black swan, and so is a scientific statement. Popper's ideas gained widespread traction, and so have heavily influenced how statistics has developed.

Implications for hypothesis testing

- Want to test if a hypothesis is false. (Can never prove it is true.)
- Example: The average weight gain of rats on a certain drug is at least 0.2 ounces.

Definition 39
The **null hypothesis** is the hypothesis that we are trying to disprove.

17.2. Frequentist hypothesis testing

- Let w be the average weight gain of rats in the experiment.
- The null hypothesis is $H_0 : \{w < 0.1\}$.
- Should we reject? Note null hypothesis is one-sided
  - Find a one-sided confidence interval for w of form $[a, \infty)$
  - If this doesn't overlap $(-\infty, 0.1)$, reject H_0.

For data that is normally distributed, a pivot for μ is the *t-statistic*.

Definition 40
The ***t*-statistic** of data $(d_1, \dots, d_n)$ for null $H_0 : \{\text{mean is } \mu\}$ with unbiased mean and variance estimates $\hat{\mu}$ and $\hat{\sigma}^2$ is

$$t = \frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}}.$$

The distribution of t is called the **Student *t*-distribution** with $n - 1$ degrees of freedom, and written $t(n-1)$.

Example

- Suppose $w = (0.23, 0.17, 0.42, 0.34)$. Then $\hat{w} = 0.29$, $\hat{\sigma} = 0.1116542$.
- One sided CI: Want $\mathbb{P}(a(w) \le w < \infty) = 0.95$.
- Recall $(w - \hat{w})/(\hat{\sigma}/\sqrt{n}) = T_{n-1} \sim t(n-1)$. Want

$$\mathbb{P}(a \le T_{n-1} < \infty) = 0.95 \Rightarrow a = -2.353363,$$

so

$$\mathbb{P}\left(-2.353363 \le \frac{w - \hat{w}}{\hat{\sigma}/\sqrt{n}} < \infty\right) = 0.95$$

And solving for w gives:

$$\mathbb{P}\left(\hat{w} - 2.353363\hat{\sigma}/\sqrt{n} \le w < \infty\right) = 0.95$$

```
hatw - hatsigma*qt(0.05,df=3)/sqrt(length(w))
```

gives $[0.1586, \infty)$. Does not overlap with H_0!

- So we reject the hypothesis at the 5% significance level.

Example: Suppose that I test each of 20 people on a standardized test. I then give them a study regime, then test them again. Did the regime help?

Suppose improvement for a person is $d_i \sim d$. I want to know, is $\mathbb{E}[d] > 0$? Standardize by using $\delta = \mathbb{E}[d]/\operatorname{SD}[d]$. Note δ is same regardless of units for d_i. The question I hope to know the answer to: is $\delta > 0$?

Back to the question: if $\hat{\mu} = 4.9$ and $\hat{\sigma} = 3.2$, then $t = (4.9/3.2)\sqrt{20} = 6.847958\ldots$. Is that big enough to say that $\mathbb{E}[d] > 0$?

Fact 33
If $d \sim \mathsf{N}(0, \sigma^2)$, then the distribution of the t-statistic only depends on n.

Definition 41
For $d \sim \mathsf{N}(0, \sigma^2)$, call the distribution of t the **Students t distribution** (or just *t* **distribution** for short) with $n-1$ degrees of freedom.

- Note: even if $d \not\sim \mathsf{N}(\mu, \sigma^2)$, since $\hat{\mu}$ is the sum of random variables, the t statistic might be close to the t distribution.
- `pt(6.847958,df=19,lower.tail=FALSE)` in R tells me that there is only a $7.775 \cdot 10^{-7}$ chance that t would be at least 6.87958. So yes, it's big enough to say there is strong evidence that there is an effect from the regime!

Definition 42
If the hypothesis contains only a single parameter value, it is **simple**. If it contains more than one parameter value, it is **compound**.

- Ex: $H_0 = \{w \in [0,1]\}$ is compound, while $H_0 = \{w = 0\}$ is simple.

Matching confidence interval to the hypothesis

- Reject H_0 when θ is too large. CI $= [a, \infty)$.
- Reject H_0 when θ is too small. CI $= (-\infty, a]$.
- Reject H_0 when θ too large or too small. CI $= [a, b]$.

Error

- Note that we can always get unlucky.
- Even though H_0 is true, we might reject it anyway.

Definition 43
If you reject the null hypothesis, even though H_0 is true, this is called by statisticians a *Type I error*.

Definition 44
If the chance of making a Type I error is at most α, then call α the **significance level** of the hypothesis test.

In the example above, $S = \hat{w} - 2.353363\hat{\sigma}/\sqrt{n}$. If $S < 0.1$ then we did not reject H_0, and if $S \geq 0.1$ then we did reject H_0.

Definition 45
Let S be a statistic and R a region such that if $S \in R$ we reject H_0, otherwise we do not reject H_0. Then call S a **test statistic** and R a **rejection region**.

In the example above, S is the test statisting, and $R = [0.1, \infty)$ is the rejection region.

17.3. p-values

Rats Let's go back to the rats gaining weight example. For that data set, we rejected the hypothesis at the 5% level.

- At what level would we have not rejected? Recall confidence interval

$$\left[\hat{w} + q_\alpha \hat{\sigma}/\sqrt{n}, \infty\right),$$

  where $\mathbb{P}(q \leq T_3 < \infty) = 1 - \alpha$.

- When $\alpha = 0.05$, $q = -2.353363$.
  When $\alpha = 0.01$, $q = -4.50703$, CI $= [0.03650566, \infty)$.
  Would not reject at 1% level!

- The smaller the significance level, the less likely you are to reject!

- At exactly what level do you change from not reject to reject? That is the p-value.

$$\hat{w} + q\hat{\sigma}/\sqrt{n} = 0.1 \Rightarrow q = -3.403364.$$

  and $\mathbb{P}(T_3 \leq -3.403364) = 0.02117825$. So the p-value for not rejecting the null is 2.1%.

Intuition 3
For a constant p, suppose that a test rejects a hypothesis at significance level $\alpha \geq p$ and does not reject when $\alpha < p$. Then call p the p-value for the hypothesis.

Common usage:

- When $p \leq 0.05$, the rejection of the null hypothesis is said to be *statistically significant*.

Avoid these common confusions!

- The p-value is **not** the probability that the null hypothesis is false.

- The words *statistically significant* and *significant* should not be confused.

  – Example: Suppose a sample of 10,000 children studies a new learning technique. The average score of the children was raised by 0.02 percentage points on a standardized test. Because of the large sample size, this could be enough to make the rejection of the null hypothesis statistically significant, but a 0.02 percentage point rise is not significant. A more accurate way of reporting the results would be a confidence interval for the difference between mean test scores with or without the new technique (effect sizes).

Problems

17.1: A hypothesis containing only a single parameter value is called what?

17.2: Suppose that $T(X) \in R$ where X is our data, T is our test statistic and R is our rejection region. What does that mean for the null hypothesis?

17.3: True or false: t statistics under the null hypothesis have a t distribution.

17.4: Say if the following hypothesis are simple or compound.

(a) $H_0 : \mu = 0$.
(b) $H_0 : \mu < 0$.
(c) $H_a : \mu \geq 0$.

(d) $H_0 : \mu \in \{0, 1\}$.

17.5: Suppose that a group of students is trying to assess whether or not the mean price of textbooks has risen more than \$20 in the past five years. Let μ_{-5} be the mean price of textbooks 5 years ago, and μ_0 be the current price.

(a) State the null hypothesis in terms of the μ_i.

(b) State the alternate hypothesis in terms of the μ_i.

17.6: A researcher is considering the effects of childhood income on graduation from college. Let μ_0 be the mean graduation rate for children born in poverty, and μ_1 be the mean graduation rate for children not born in poverty.

(a) State the null hypothesis.

(b) If the researchers only cared that being not born into poverty increased the college graduation rate, state the alternative.

(c) If the researchers only care that being not born into poverty increased the college graduation rate by at least 50%, state the alternative.

(d)

Chapter 18

Hypothesis selection

Question of the Day A new drug lowers the blood pressure of patients with chance p. We want to test the hypothesis $H_0 : p = 0.6$ against $H_a : p = 0.3$ using an experiment involving ten patients. If $\alpha = 0.03$, find a test with the lowest chance of rejecting H_a given that H_a is correct.

In this chapter

- The alternative hypothesis.
- Type II error.

Working with more than one hypothesis Often, there is more than one hypothesis involved. That is, either H_0 (the null hypothesis) is true, or H_a (the alternative hypothesis) is true, but not both. Since knowing that either H_0 or H_a is give us better information, so we should be able to make better decisions. On the other hand, we also can make a new type of error.

Definition 46
If we reject H_a even though H_a is true, say that we have made a *Type II error*.

Notation 6
We usually use α as an upper bound on the Type I error, and β as an upper bound on the Type II error.

18.1. Maximizing power

So how can we maximize the power? Try to make the rejection region for H_0 as large as possible in order not to reject H_a unnecessarily.

Question of the Day Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Bern}(p)$. If H_0 is true, since p is higher than if H_a is true, a good test statistic is $X = X_1 + \cdots + X_{10}$. High values of X indicate that H_a should be rejected, and low values indicate that H_0 should be rejected. Therefore, if $X < c$, reject H_0, if $X \geq c$, reject H_a. Type I error is $\alpha = 0.3$. If H_0 is true, then the chance of Type I error is

$$\mathbb{P}(X < c | H_0 \text{ true}) = \mathbb{P}(X < c | p = 0.6).$$

Using R

$$\texttt{pbinom(3,10,0.6)} \to \mathbb{P}(X \leq 3 | p = 0.6) = \mathbb{P}(X < 4 | p = 0.6) = 0.05476188$$
$$\texttt{pbinom(2,10,0.6)} \to \mathbb{P}(X \leq 2 | p = 0.6) = \mathbb{P}(X < 3 | p = 0.6) = 0.01229455.$$

So set $c = 3$. That makes the chance of rejecting H_a when it is true:

$$\mathbb{P}(X \geq 3) = 0.6172.$$

So the smallest probability of Type II error (given Type I error of at most 3%) is 61.72%.

Some comments Here α does have to exactly equal the chance of Type I error, it is just an upper bound on the probability. Similarly, β is an upper bound on Type II error, and not equal to it. That means we get the same answer if our null and alternate are $H_0 : p \in [0.6, 1]$ and $H_a : p \in [0.3, 1]$ rather than just $H_0 : p = 0.6$ and $H_a : p = 0.3$.

Definition 47
The **power** of a test is 1 minus the chance of a Type II error.

Remember we like tests with small Type II error (so small β) which means that we like tests that have high power.

18.2. Sample sizes

- In last examples, once α, n fixed, β fixed as well.
- Only way to decrease β for fixed α is increase n.
- Can ask question, how big does n have to be for fixed α, β?

In last example, once α and n fixed, then our bound β on the Type II error probability is fixed as well. The only way to decrease β for a fixed value of α is increase n. A natural question to ask is: how big does n have to be to meet fixed values of α, β?

Example: Going back to the question of the day, how large does n have to be in order to make $\alpha = \beta = 0.05$? And what c should we use?

Recall that for the question of the day, our test statistic $X \sim \binom{(}{n}, p)$, where $p = 0.6$ if H_0 holds, and $p = 0.3$ if H_a holds. The Central Limit Theorem gives us that $(X - \mathbb{E}[X])/\operatorname{SD}(X)$ is approximately normally distributed. Hence

$$\begin{aligned}\mathbb{P}(X < c | p = 0.6) &= \mathbb{P}\left(\frac{X - 0.6n}{\sqrt{n(0.6)(0.4)}} < \frac{c - 0.6n}{\sqrt{n(0.6)(0.4)}}\right) \\ &\approx \mathbb{P}\left(Z < \frac{c - 0.6n}{\sqrt{n(0.6)(0.4)}}\right) = \operatorname{cdf}_Z\left(\frac{c - 0.6n}{\sqrt{n(0.6)(0.4)}}\right) = 0.05,\end{aligned}$$

where $Z \sim \mathsf{N}(0, 1)$. So

$$\frac{c - 0.6n}{\sqrt{n(0.6)(0.4)}} = \operatorname{cdf}_Z^{-1}(0.05) = -1.644854.$$

Similarly, for the Type II error:

$$\begin{aligned}\mathbb{P}(X \geq c | p = 0.3) &= \mathbb{P}\left(\frac{X - 0.3n}{\sqrt{n(0.3)(0.7)}} \geq \frac{c - 0.3n}{\sqrt{n(0.7)(0.3)}}\right) \\ &\approx \mathbb{P}\left(Z \geq \frac{c - 0.3n}{\sqrt{n(0.3)(0.7)}}\right) = 1 - \operatorname{cdf}_Z\left(\frac{c - 0.3n}{\sqrt{n(0.3)(0.7)}}\right) = 0.05,\end{aligned}$$

where $Z \sim \mathsf{N}(0, 1)$. So

$$\frac{c - 0.3n}{\sqrt{n(0.3)(0.7)}} = \operatorname{cdf}_Z^{-1}(0.95) = 1.644854.$$

Therefore we have two equations in two unknowns. Solving both for c, and setting them equal to each other yields an quadratic equation in $\sqrt{n}$ which can be solved to yield

$$n = 27.0254, c = 12.0261.$$

Of course, the CLT was just an approximation here. But the exact values can be found with R using the `pbinom` command. A quick table of integer values near to the values above gives:

n	c	$\mathbb{P}$(Type I error)	$\mathbb{P}$(Type II error)
27	12	0.03369	0.07980
27	13	0.07432	0.01425
28	12	0.02150	0.1028
28	13	0.0499495	0.0491038.

Therefore, the sample size and test are

$$\boxed{n = 28, \text{ reject } H_0 \text{ if } X < 13.}$$

Problems

18.1: Rejecting the null when the null is true is what type of error?

18.2: Fill in the blank: _____ is usually used to represent an upper bound on Type II error.

18.3: True or false: The power of a test plus the chance of Type II error must add to 1.

18.4: True or false: We want Type II error to be as low as possible.

18.5: When deciding which is the null and which is the alternate, the hypothesis that an intervention does not change the mean is typically which hypothesis?

Chapter 19

p-values

Question of the Day How can we quantify evidence that the null is false?

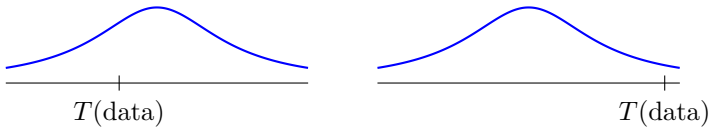
In this chapter

- p-values

19.1. What is a p-value?

Suppose that null hypothesis is actually true, so that we know that exactly distribution of the data (parameters and all.) Then the test statistic T that we are using has some known distribution

The idea behind p-values is when T value for the data is far out in one of the tails of its distribution, that provides evidence that the null is false.



Little evidence against null Evidence against null

Definition 48
Let $[X|\theta]$ be the statistical model, $\theta = \theta_0$ the null hypothesis, and T be a test statistic. For $d = (d_1, \ldots, d_n)$ a data set, and $X = (X_1, \ldots, X_n)$ where $X_i \overset{\text{iid}}{\sim} [X|\theta = \theta_0]$, the **$p$-value** of the data is

$$\mathbb{P}(T(X) \text{ is at least as extreme as } T(d)).$$

If $p = \mathbb{P}(T(X) \geq T(d))$ or $p = \mathbb{P}(T(X) \leq T(d))$, then p is a one-sided (or one-tailed) p-value. If $p = \mathbb{P}(|T(X)| \geq |T(d)|)$, then p is a two-sided (or two-tailed) p-value.

In words, the p-value is the chance that if the null hypothesis is true, that we would have attained a test statistic as weird or weirder as we did simply by chance.

Our earlier hypothesis tests can be written in terms of p-value.

1: Let α be the maximum chance of Type I error.

2: Draw data, and calculate p.

3: If $p \leq \alpha$, reject the null. Otherwise, do not reject the null.

Example Suppose that we measure arithmetic ability in students after receiving a caffeine pill using 20 questions. The results for four students are:

$$17, 14, 16, 19.$$

If the average score by noncaffinated students is 15, what is the p-value to reject the null hypothesis that caffeine does not improve scores?

To answer this question, first we need a statistical model: for each question, each student has a $15/20 = 0.75$ chance of answering correctly. That gives the results that we say for the noncaffinated students, an average of 15 out of 20 correct questions.

Let $S = X_1 + X_2 + X_3 + X_4$ be the test statistic. The null hypothesis is $S \sim \mathsf{Bin}(80, 0.75)$. Since we are testing if caffeine raises scores, our notion of weird will be that S is much larger than it would be if the null hypothesis was true.

So the p-value is $\mathbb{P}(S(X) \geq S(\text{data})) = \mathbb{P}(S(X) \geq 17+14+16+19)$. We can find this probability with R

```
1 - pbinom(65,80,0.75)
```

which gives a p-value of 0.07398627.

So I shouldn't report it right? Yes and no. It's low, so there might be something there. You need to rerun the experiment with a larger number of subjects if possible.

19.2. If the null hypothesis is true, then p-values are uniform over $[0, 1]$

Under the best conditions p-values are difficult to interpret. They are not the chance that the null hypothesis is false. So what are they? One insight is that they are random variables. Moreover, if the null hypothesis is true, then we can say exactly what the distribution of the p-value will be!

Fact 34
If the data set $d = (D_1, \ldots, D_n)$ where $D_i \overset{\text{iid}}{\sim} [X|\theta = \theta_0]$ is a continuous distribution, then

$$p \sim \mathsf{Unif}([0, 1]).$$

This is difficult to prove in general, but a special case can be proved with the help of a simple fact from probability.

Fact 35
Let X be a continuous random variable with cdf F_X. Then $F_X(X) \sim \mathsf{Unif}([0, 1])$.

Proof. Let $a \in (0, 1)$. Then $\mathbb{P}(F_X(X) \leq a)$. Let $F_X^{-1}(b) = \inf\{c : F_X(c) \leq b\}$. Since X is continuous and nondecreasing, F_X^{-1} is nondecreasing as well, $F_X^{-1}(F_X(x)) = x$, and $F_X(F_X^{-1}(x)) = x$. Hence

$$\mathbb{P}(F_X(X) \leq a) = \mathbb{P}(X \leq F_X^{-1}(a)) = F_X(F_X^{-1}(a)) = a.$$

Therefore $F_X(X)$ has the cdf of a uniform over $[0, 1]$. □

- Now consider the p-value special case where $p = \mathbb{P}(T(X) \leq T(D))$.
- Let $\alpha \in [0, 1]$. Then

$$\begin{aligned}\mathbb{P}(p \leq \alpha) &= \mathbb{P}(T(X) \leq T(D)) \\ &= F_{T(X)}(T(D)) \\ &= F_{T(D)}(T(D)) \sim \mathsf{Unif}([0, 1]).\end{aligned}$$

- In particular, if $\alpha = 5\%$, then *just by chance even if the null is true* 5% of the time p will be at most 0.05.

Another way to state this fact.

Fact 36
If the null hypothesis is true, then the p-value for a test statistic will be uniformly distributed over $[0, 1]$.

19.3. Relating p-values to confidence intervals

- Suppose we have a family of confidence intervals $\{[a_\alpha, b_\alpha]\}$ such that interval $[a_\alpha, b_\alpha]$ is an α-level conf. int., and
$$(\alpha_1 < \alpha_2) \to ([a_{\alpha_1}, b_{\alpha_1}] \subseteq [a_{\alpha_2}, b_{\alpha_2}]).$$
  - Most intervals found through pivoting have these properties.
- Note that as $\alpha \to 0$, the $1-\alpha$ level confidence interval grows in size. So if α is large, then the confidence interval is small, less likely to contain the parameter value in the null, and rejection is likely. As α shrinks, the confidence interval grows until it just includes the null hypothesis as a possiblility. That value of α is the p-value.

  Put another way, given an alternative hypothesis, the p-value is the maximum value of α such that the $1-\alpha$-level confidence interval includes the null hypothesis.

19.4. p hacking

Remember that if you run an experiment, you are 5% likely to have a p-value that is at most 5% just by chance. That means that if you look at 20 different test statistics on different data sets, just by chance you expect to have one that is "statistically significant".

How science should work:

- Determine statistical procedure. Publish your planned experiment so that others know what effects you are looking for, and how you plan to test for them.
- Run the experiment.
- Find p-value from the data.
- If p-value greater than 0.05 (or whatever the cutoff is in your field), report the result to the appropriate journal.
- If p-value at most 0.05, report result to appropriate journal.

Unfortunately, many journals do not accept results with $p > 0.05$. So process becomes:

- Determine statistical procedure.
- Run experiment.
- Find p-value.
- Only report in journal if p-value greater than 0.05.

That's a big problem, because then you can't tell if the experiment obtained the result because there truly is an effect there, or if it merely happened by chance. It gets worse: because it is so important to have a small p-value, researchers go hunting for a test statistic that gives a small p-value.

- Run experiment
- Find p-value for multiple subsets of the data until find $p > 0.05$.
- Only report there subsets to the journal if p-value greater than 0.05.

An example of this was a study that looked at prayer versus medical outcomes:

- Does prayer cause shorter hospital stays? $p = 0.7$
- Does prayer cause fewer complications? $p = 0.35$

$\vdots$

- Does prayer cause fewer return visits to hospital? $p = 0.03$

This quickly gets put in a headline:

> Prayer results in fewer return visits to the hospital!

Followed in short order by the following appearing on social media:

> Prayer heals!

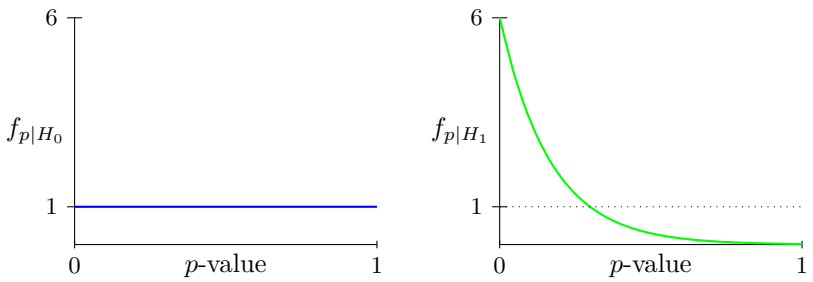
The result is a large number of false positives. One study indicated that as many as half of the articles in the psychology literature report incorrect results.

Should we never use p-values?

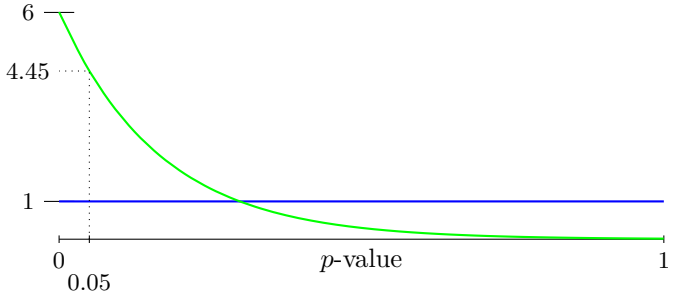
- p-values can be a useful tool for determining which effects are worth following up on.
- Does not provide a good framework for evaluation of experiments.
- Confidence intervals for predetermined variables preferred.
- Easy to misinterpret–be very careful when looking at studies reporting p-values near 0.05.

19.5. How much can we learn about a null and an alternate from a p-value?

The short answer: not as much as we often think! Remember, if the null hypothesis H_0 is true, $p \sim \mathsf{Unif}([0,1])$. If the alternate hypothesis H_1 is true, then we hope that the p-value is more likely to occur than for large values. So the density of p looks like this:



So now the question is we see a p-value of 0.05. How much evidence does that give for H_1 over H_0 if initially each of the two were considered equally likely?



Not as much as you might think, since the density is only about 4.45 times as high under the alternate as under the null. So given a p-value of 0.05, we should only give 4.45 to 1 odds that the true hypothesis is the alternate rather than the null. Call the 4.45 the *MPR* or Maximum p-Ratio.

Of course, we don't know the green line! It could be even worse! The line in the picture above is if the p-value is an exponential with rate 6 conditioned to lie in $[0,1]$.

What if under the alternate, the true p-value has an exponential distribution with rate 1000 conditioned to lie in $[0,1]$? Then the height of the density at 0.05 is $100\exp(-100 \cdot 0.05)/(1-\exp(-100)) \approx 0.6737$.

In other words, under the p-value of 0.05 found here, the alternative is *less likely* to be true than the null hypothesis!

Now that's an extreme case, but in a more common case where under the alternate, $p \sim \mathsf{Beta}(1, 2)$, the odds of the hypothesis being true at $p = 0.05$ is a mere 1.9 to 1. And in fact, no matter how small the p-value is in this case (for instance $p = 10^{-6}$), the odds of H_0 versus H_1 never go higher than 2 to 1!

Problems

19.1: Under the null hypothesis, the chance that a p-statistic is in $[0.3, 0.34]$ is what?

Chapter 20

The Neyman-Pearson Lemma

Question of the Day Let $X_1, X_2, \ldots \sim \mathsf{Exp}(\lambda)$. If $H_0 : \lambda = 1.2$ and $H_1 = \lambda = 3.2$, what is the most powerful test for distinguishing between H_0 and H_1 with Type I error of 5%.?

In this chapter

- Test statistics
- The Neyman-Person Lemma

In the first half the 20th century, statisticians struggled with the question of what it meant to have a good test of a hypothesis. One approach was to try to bound the probabilities of making a Type I or Type II error. This is illustrated from a quote from Neyman and Pearson in 1933.

> But we may look at the purpose of tests from another view-point. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong.
>
> Neyman & Pearson, 1933

The goal of Neyman and Pearson was to build a means of rejecting a hypothesis in such a way that in the long run we were not too often wrong. That is, we do not reject the null too often when it is true (Type I error) and we do not reject the alternate too often when it is true (Type II error). It turns out they were able to show a simple but powerful theorem that gives the most powerful test for a given Type I error.

In the case of two hypothesis H_0 and H_1, Neyman and Pearson suggested that the *likelihood ratio* be used to determine if we reject or not.

Definition 49
Let $L(\theta|x)$ denote the likelihood function of parameter θ given data x. Then for two hypothesis $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, the **likelihood ratio** between H_0 and H_1 is

$$\frac{L(\theta_0|x)}{L(\theta_1|x)}.$$

A natural approach to testing is to reject H_0 when $L(\theta_0|x)$ is small compared to $L(\theta_1|x)$, and otherwise reject H_1. That is, reject H_0 when

$$\frac{L(\theta_0|x)}{L(\theta_1|x)} \leq K,$$

for some constant K. What they were able to show is that under mild conditions, this also gives the most powerful test with that level of Type I error. That is, for all tests with a given Type I error, this type of test has the smallest Type II error. This result has become known as the Neyman-Pearson Lemma.

Theorem 6 (Neyman-Pearson Lemma)
Given null $H_0 : \theta = \theta_0$ and alternative $H_a : \theta = \theta_1$, and data $X_1, \ldots, X_n \overset{\text{iid}}{\sim} X$, let $L(\theta|x_1, \ldots, x_n)$ be the likelihood function of θ. Then let R_K be the set of $(x_1, \ldots, x_n)$ such that

$$\frac{L(\theta_0|x_1, \ldots, x_n)}{L(\theta_1|x_1, \ldots, x_n)} \leq K.$$

Then R_K is the most powerful test with respect to the alternative. That is, for any other rejection region R such that

$$\mathbb{P}((X_1, \ldots, X_n) \in R|H_0) \leq \mathbb{P}((X_1, \ldots, X_n) \in R_K|H_0)$$

we have that

$$\mathbb{P}((X_1, \ldots, X_n) \notin R_K|H_1) \leq \mathbb{P}((X_1, \ldots, X_n) \notin R|H_1).$$

Note: since natural log is an increasing function, this is equivalent to the rejection region being where the log-likelihood ratio is at most a constant:

$$\left\{(x_1, \ldots, x_n) : \ln\left(\frac{L(\theta_0|x_1, \ldots, x_n)}{L(\theta_1|x_1, \ldots, x_n)}\right) \leq K\right\}$$

As with the maximimum likelihood estimator, the log-likelihood ratio is often simpler to deal with.

Question of the Day Let's start by applying this lemma to the Question of the Day. In this case, our likelihood function given data $x_1, \ldots, x_n$ all nonnegative is

$$\ln\left(L(\lambda|x_1, \ldots, x_n)\right) = \sum_{i=1}^{n} \ln(\lambda \exp(-\lambda x_i) = n\ln(\lambda) - \lambda\left[\sum_{i=1}^{n} x_i\right].$$

Since the log-likelihood (and hence likelihood ratio) only depends on the data through the value of $s = \sum_{i=1}^{n} x_i$, let $S = \sum_{i=1}^{n} X_i$ be our test statistic.

Then the log-likelihood ratio is

$$\ln\left(\frac{L(1.2|s)}{L(3.2|s)}\right) = n\ln(\lambda) - 1.2s - [n\ln(\lambda) - 3.2s] = 2s$$

Therefore, the Neyman-Pearson rejection region has the form: reject H_0 if for some constant K,

$$2s \leq K,$$

which of course can be changed by dividing by 2 to get, reject when

$$s \leq K'$$

where K' is just a different constant.

So we want the probability of Type I error (that we reject H_0 even though it is true) to be 5%. That means that we choose K' so that $\mathbb{P}(S \leq K') = 0.05$.

Now $[S|\lambda] = [X_1 + \cdots + X_n|\lambda] \sim \mathsf{Gamma}(n, \lambda)$ by the properties of exponential random variables. So that means we can find the correct value of K' by using the inverse cdf of a gamma distribution under the assumption that H_0 is true:

$$K' = \text{cdf}^{-1}_{\mathsf{Gamma}(n,1.2)}(0.05).$$

Then the probability of Type II error is the probabiltiy that $S > K'$ conditioned on H_1 being true. That is

$$1 - \text{cdf}_{\mathsf{Gamma}(n,3.2)}(\text{cdf}^{-1}_{\mathsf{Gamma}(n,1.2)}(0.05)).$$

For instance, suppose I have 4 data points. Using the R command `qgamma(0.05,4,rate=1.2)` gives $K' = 1.138599$. So the most powerful test for distinguishing H_0 from H_1 is to reject H_0 when $S \leq 1.38599$ and reject H_1 when $S > 1.38599$.

The chance of Type II error is found by looking at the probability that $S > 1.38599$ when H_1 is actually true. That happens with probability 0.4939955, so that is the chance of making a Type II error.

Suppose the data taken were $0.075, 0.248, 0.294, 0.568$. Then $s = 1.86$, and we would reject H_1 and not reject H_0.

A normal example Suppose we have a data set of 5 points $X_1, \ldots, X_5$ that are iid with the same distribution as X. The null hypothesis H_0 is that $X \sim \mathsf{N}(10, 4^2)$, while the alternate hypothesis H_a is that $X \sim \mathsf{N}(8, 4^2)$. If we want a test that distinguishes between H_0 and H_a with Type I error probability of 0.01, then what is the most powerful test?

First let's calculate the likelihood ratio for the two hypotheses. Because both have the same variance, the normalizing constant for the normals cancels out, leaving

$$\ln\left(\frac{L(10|x)}{L(8|x)}\right) = \ln\left(\frac{\prod_i(\exp(-(x_i-10)^2/32))}{\prod_i(\exp(-(x_i-8)^2/32))}\right) = \sum_i \left[(x_i-8)^2 - (x_i-10)^2\right]/32$$
$$= ((20-16)/32)\sum_{i=1}^{n} x_i + C.$$

Making this less than or equal to a constant is the same as making

$$\sum_{i=1}^{n} x_i \leq K,$$

where K is a constant. Again, this comports well with our intuition: reject the hypothesis that the mean is larger whenever the sum of the random variables is too small.

Again we can use R to find $\text{cdf}^{-1}_{N(10,4^2)}(0.05)$:

```
qnorm(0.05,mean=10,sd=4)
```

returns $K = 3.420585$. Then `1-pnorm(3.420585,mean=8,sd=4)` gives us our Type II error chance of 0.8738651.

A discrete example Now consider a data set from a discrete distribution. Suppose that a drug trial is testing $H_0 : p = 0.5$ versus $H_1 : p = 0.7$ after recording the success or failure of the drug on 41 patients. Assuming the patients are independent, then gives $[X|p] \sim \mathsf{Bin}(n, p)$. Suppose we want the uniformly most powerful test that distinguishes between these two at the 1% level.

The density of discrete random variables is just the probability mass function, so

$$L(p|x) = f_{X|p}(x) = \binom{n}{x} p^x (1-p)^{n-x},$$

and

$$\ln\left(\frac{L(0.5|x)}{L(0.7|x)}\right) = x[\ln(0.5) - \ln(0.7) - \ln(1-0.5) + \ln(1-0.7)] \leq K.$$

Note that that number inside the brackets is negative, so this is equivalent to saying reject if $x \geq K'$ for some constant K'. (The direction of the inequality flips when we divide by a negative number.)

Again this fits with our intuition: reject the null hypothesis that p is the smaller number when the value of the data is too large.

The next step is to find the value of K' such that

$$\mathbb{P}(X \geq K') \leq 0.01.$$

This can be done by using the `pbinom` command in R for various values of K'. Trial and error quickly narrow the correct value down to $K' = 28$. That is, $\mathbb{P}(X \geq 27|p = 0.5) = 0.1376$ (so 27 is too low) while $\mathbb{P}(X \geq 28|p = 0.5) = 0.00575$.

The chance of rejecting $p = 0.7$ when that is true is then $\mathbb{P}(X < 28|p = 0.7) = 0.3345$, so that is the chance of Type II error.

20.1. Proof of the Neyman Pearson Lemma

Proof. Suppose that X has density f with respect to measure μ, where $X \in \Omega$. Let R be any rejection region. Then

$$\mathbb{P}(X \in R) = \mathbb{E}[\mathbb{1}(X \in R)].$$

For R_k the rejection region where the likelihood ratio is at most $k > 0$, let R be a second rejection region with the same level or lower of Type I error. So

$$\mathbb{E}_{\theta_0}[\mathbb{1}(X \in R)] \leq \mathbb{E}_{\theta_0}[\mathbb{1}(X \in R_k)].$$

We're interested in showing that the test using R is less powerful than the test using R_k. That is, we want to show

$$\mathbb{E}_{\theta_1}[\mathbb{1}(X \notin R)] \geq \mathbb{E}_{\theta_1}[\mathbb{1}(X \notin R_k)].$$

Since $\mathbb{1}(X \notin A) = 1 - \mathbb{1}(X \in A)$, this is equivalent to showing that

$$\mathbb{E}_{\theta_1}[\mathbb{1}(X \in R)] \leq \mathbb{E}_{\theta_1}[\mathbb{1}(X \in R_k)].$$

In other words, we want to show that

$$\mathbb{E}_{\theta_0}[\mathbb{1}(X \in R_k) - \mathbb{1}(X \in R)] \geq 0 \Rightarrow \mathbb{E}_{\theta_1}[\mathbb{1}(X \in R_k) - \mathbb{1}(X \in R)] \geq 0.$$

To do this, define the function

$$g(x) = (\mathbb{1}(x \in R_k) - \mathbb{1}(x \in R))(kf_{\theta_1}(x) - f_{\theta_0}(x)).$$

If $\mathbb{1}(x \in R_k) - \mathbb{1}(x \in R) \geq 0$, then $x \in R_k$ and so $f_{\theta_0}/f_{\theta_1}(x) \leq k$. So if the first term is positive, then so is the second term. Similarly, if the first term is negative, then $x \notin R_k$ and the second term is negative. Hence $g(x) \geq 0$ whether or not x is in R_k. Therefore,

$$\int_\Omega g(x)\ d\mu \geq 0.$$

This integral can be written as the difference of two expected values, since we are multiplying functions times densities. That is,

$$\int_\Omega g(x)\ d\mu = k\mathbb{E}_{\theta_1}[\mathbb{1}(x \in R_k) - \mathbb{1}(x \in R)] - \mathbb{E}_{\theta_0}[\mathbb{1}(x \in R_k) - \mathbb{1}(x \in R)] \geq 0,$$

which gives

$$\mathbb{E}_{\theta_1}[\mathbb{1}(x \in R_k) - \mathbb{1}(x \in R)] \geq (1/k)\mathbb{E}_{\theta_0}[\mathbb{1}(x \in R_k) - \mathbb{1}(x \in R)].$$

since $k > 0$. Therefore, if the right hand side is nonnegative, then so is the left hand side. □

Problems

20.1: True or false: Likelihood ratio tests require two possible hypotheses.

20.2: Suppose a research groups gathers a data that is summarized by a statistic X. The group forms a hypothesis that X comes from either density f_0 (the null), or it will come from density f_1 (the alternate).

Describe how you would construct a test for the collected dataset s of the null versus the alternate at the 5% significance level.

20.3: Suppose that a researcher models their summary statistic X as coming (null) from a beta with parameters 2 and 1 (so density $2s\mathbb{1}(s \in [0,1])$) or, alternatively, coming from a beta with parameters 3 and 1 (so density $3s^2\mathbb{1}(s \in [0,1])$.)

(a) Construct the uniformly most powerful test at the 5% for testing the null versus the alternate. Be sure to state any theorems that you are using.

(b) Evaluate your test at data $X = 0.8$. Would you reject the null at the 5% level?

Chapter 21

Bayes factors

Question of the Day A public relations firm polls 100 randomly selected people to see who is favorable toward a new product. The company is trying to determine if at least 40% of people would support the new product (H_a), versus at most 20% (H_b). The poll reveals 24 people in favor of the new product. What evidence does this provide for H_b?

In this chapter

- Hypothesis testing with priors

Example

- Both H_a and H_b have the same statistical model:

$$[X|p] \sim \mathsf{Bin}(100, p).$$

- They differ in where they put p:

$$H_a = \{p \geq 0.4\}, \quad H_b = \{p \leq 0.2\}.$$

- Now suppose we ask 100 people about the new product, and 24 say yes. How does that affect our belief in these hypotheses?

- To find the posterior probability, use Bayes' Rule:

$$\mathbb{P}(H_a|X = 24) \propto \mathbb{P}(H_a) \cdot \mathbb{P}(X = 24|H_a)$$
$$\mathbb{P}(H_b|X = 24) \propto \mathbb{P}(H_b) \cdot \mathbb{P}(X = 24|H_b)$$

  Note they have the *same* constant of proportionality! So

$$\frac{\mathbb{P}(H_a|X = 24)}{\mathbb{P}(H_b|X = 24)} = \frac{\mathbb{P}(H_a) \cdot \mathbb{P}(X = 24|H_a)}{\mathbb{P}(H_b) \cdot \mathbb{P}(X = 24|H_b)}$$

  Another way to write this:

$$\text{posterior ratio} = \text{prior ratio} \cdot \text{Bayes factor}$$

  Can generalize to densities.

Definition 50
Let $f(x|\theta)$ be a statistical model data x, the **Bayes factor** between hypothesis $H_a : \theta = \theta_a$ and $H_b : \theta = \theta_b$ is

$$\frac{f(x|\theta_a)}{f(x|\theta_b)}.$$

- Note that in qotd, the hypothesis are not simple. In this case, the Bayes factor is usually taken to be

$$\frac{\max_{\theta \in H_a} f(x|\theta)}{\max_{\theta \in H_b} f(x|\theta)}.$$

- For the qotd, this gives:

$$F = \frac{\binom{100}{24}(0.4)^{24}(0.6)^{100-24}}{\binom{100}{24}(0.2)^{24}(0.8)^{100-24}}.$$

  Today, just put it on computer, but for large data, often ln is useful:

$$\ln(F) = 24\ln(2) + 76\ln(0.75) = -5.228, \quad F = 0.005362.$$

21.1. How to interpret Bayes Factors:

- First: Remember that Bayes factors multiply prior ratio to get posterior ratio. So if numerator hypothesis unlikely to start, could still be unlikely after evidence.
- With that caveat in mind, two scales.

  H. Jeffreys, *The Theory of Probability (3rd ed.)*, 1961, p. 432

F	Strength of evidence
$[0, 1)$	negative, supports denominator hypothesis
$[1, \sqrt{10})$	barely worth mentioning
$[\sqrt{10}, 10)$	substantial
$[10, 10^{1.5})$	strong
$[10^{1.5}, 10^2)$	very strong
$[100, \infty)$	decisive

  R. E. Kass, A. E. Raftery, *Bayes Factors*, JASA, **90**:430, p. 791

$2\ln(F)$	F	Strength of evidence
$[0, 2)$	$[1, 3)$	Not worth more than a bare mention
$[2, 6)$	$[1, \sqrt{10})$	Positive
$[6, 10)$	$[20, 150)$	Strong
$[10, \infty)$	$[150, \infty)$	Very strong.

- Note this second table is only given in terms of positive evidence. If the evidence for the numerator hypothesis is F, then the evidence for the denominator hypothesis is $1/F$.
- Why $2\ln(F)$? Makes the value similar to frequentist log-likelihood ratio statistics.

21.2. Diffuse hypothesis testing

- Now suppose the hypotheses are not $p \geq 0.4$ or $p \leq 0.2$, but are diffuse. $H_a : p \sim \mathsf{Beta}(4, 6)$, $H_b : p \sim \mathsf{Beta}(2, 8)$. [Note: $\mathbb{E}[p|H_a] = 0.4$, $\mathbb{E}[p|H_b] = 0.2$.

- Makes finding Bayes factor a bit harder. Have to integrate out p:

$$\begin{aligned}\mathbb{P}(X=24|H_a) &= \int_{r=0}^{1} \mathbb{P}(X=24|p=r) f_p(r)\, dr \\ &= \binom{100}{24} B(4,6)^{-1} \int_{r=0}^{1} r^{24}(1-r)^{100-24} r^{4-1}(1-r)^{6-1}\, dr \\ &= CB(4,6)^{-1} B(24+4, 100-24+6)\end{aligned}$$

  After similar calculation for H_b,

$$F = \frac{B(28,82)B(2,8)}{B(4,6)B(26,84)} = 2457/3403 \approx 0.7220.$$

- Notice, spread out priors on p make experiment give less evidence for hypothesis. Tighter priors:

$$[p|H_a] \sim \mathsf{Beta}(40,60), \quad [p|H_b] \sim \mathsf{Beta}(20,80).$$

  ($\mathsf{Beta}(40,60)$ like taking 41st order statistic of 101 iid uniforms on $[0,1]$.

- Tighter prior allows data to give more evidence:

$$F = \frac{B(64,136)B(20,80)}{B(44,156)B(40,60)} \approx 0.07092$$

  Tightest prior $H_a : p = 0.4$ and $H_b : p = 0.2$ ($F = 0.005362$).

- Why not use point priors? Difficult to update to posteriors. No matter what evidence you find still would have $p \in \{0.2, 0.4\}$, cannot change value. Even with tight prior $p \sim \mathsf{Beta}(40,60)$, if you saw $X = 24$, that would alter the posterior mean towards 0.24.

21.3. Bayes Factors for one sample testing

- Suppose we have a group of 20 people take a standardized test. They are then given a study regime and they take another standardized test. If $\hat{\mu} = 4.9$ and $\hat{\sigma} = 3.9$, did the study regime improve their scores?

- If the model is that person i increased their score $d_i \sim d$, then let $\delta = \mathbb{E}[d]/\operatorname{SD}[d]$ be the standardized increase using the regime.

- Then $H_a : \delta = 0$. $H_b : \delta \sim |Y|$, $Y \sim \mathsf{Cauchy}(0, \sqrt{2}/2)$.

$$f_{\delta|H_b}(s) = \frac{\sqrt{2}}{\pi(1+x^2/2)}$$

Definition 51
Let $X \sim \mathsf{Cauchy}$, so $f_X(x) = [\pi(1+x^2)]^{-1}$. Then for $Y = \mu + \sigma X$, write $Y \sim \mathsf{Cauchy}(\mu, \sigma)$, and say that Y is a **Cauchy with location parameter μ and scale σ**.

- Recall the t-statistic: $t = (\hat{\mu}/\hat{\sigma})\sqrt{n}$. Before we said that for normal data with mean 0, the distribution was Student t. Now let's add a "noncentrality parameter".

Definition 52
Let $d_1, \ldots, d_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$ where $\mu/\sigma = \delta$. Then $(\hat{\mu}/\hat{\sigma})\sqrt{n}$ has a **Student t distribution with $n-1$ degrees of freedom and noncentrality parameter δ**.

Let $f_{t(k,\mu_0)}(a)$ denote the density of a Student t with k degrees of freedom and a noncentrality parameter of μ_0. Then the Bayes Factor in favor of nonzero δ is:

$$F = \left[\int_{s=0}^{\infty} \frac{\sqrt{2}}{\pi(1+x^2/2)} f_{t(n-1,s)}(t)\ ds\right] / f_{t(n-1,0)}(t).$$

To integrate in R we can use the `integrate` command.

```
f <- function(x) return(dt(6.847958,df=19,ncp=x))
g <- function(x) return(sqrt(2)*f(x)/(1+x^2/2))
integrate(g,lower=0,upper=Inf)
0.06430118/dt(6.847958,df=19,ncp=0)
```

For $n = 20$, $t = 6.847958$, $F \approx 41112.91$

Problems

21.1: Suppose that $X_1, \ldots, X_n$ are iid Unif$([0, \theta])$. Say $H_0 : \theta = 1$ and $H_a : \theta = 1.1$.

(a) Suppose the data drawn is $\{0.47, 0.76, 0.48\}$. Find the Bayes Factor for H_0 versus H_a.

(b) Suppose the data drawn is $\{0.47, 1.01, 0.76, 0.48\}$. Find the Bayes Factor for H_0 versus H_a.

(c) How much data would we need to take to guarantee a Bayes Factor that is either at least 10 or 0?

Chapter 22

Two sample tests

Question of the Day 10 incoming patients are randomly split into a group that receives a drug, and a group that receives a placebo. The drug group cholesterol is:

$$111, 131, 145, 125, 152$$

whereas the placebo group is

$$195, 142, 156, 110, 134$$

Does the drug lower cholesterol?

In this chapter

- Two sample data

22.1. Paired data

Suppose that we have data from two groups and we are trying to discover if they come from the same distribution or not. The first group data will be $X_1, \ldots, X_n$, and the second group will be $Y_1, \ldots, Y_m$.

When $n = m$, we could just pair up the data

$$Z_1 = X_1 - Y_1, \ldots, Z_n = X_n - Y_n,$$

and then test if $\mathbb{E}[Z_i] \neq 0$.

If $n \neq m$, then we need to be a bit more clever.

22.2. Welch's t-test

Definition 53
The two-sample T statistic for data $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ is

$$T = \frac{\hat{\mu}_X - \hat{\mu}_Y}{\sqrt{\hat{\sigma}_X^2/n + \hat{\sigma}_Y^2/m}}.$$

- Like the one dimensional t statistic, T is unitless.
- If the two samples have the same mean, expect $|T|$ to be small.
- If $\mu_X < \mu_Y$, expect T to be negative.
- If $\mu_X > \mu_Y$, expect T to be positive.

Fact 37
For X and Y normal, Welch's T has an approximately Student t-distribution with

$$\nu = \frac{(\hat{\sigma}_X^2/n + \hat{\sigma}_Y^2/m)^2}{(\hat{\sigma}_X^2/n)^2/(n-1) + (\hat{\sigma}_Y^2/m)^2/(m-1)}$$

degrees of freedom.

Notes

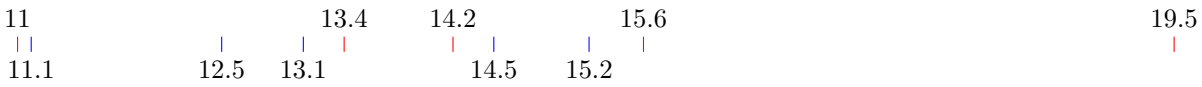
- Fortunately, this can be computed easily in R using `t.test(x,y)`.
- Also known as Welch's t test:

  B. L. Welch, *The generalization of "Student's" problem when several different population variances are involved*, Biometrika, **34**(1–2), pp. 28–35, 1947.

- For qotd, $T = -0.9315$, $\nu = 6.071$, gives p-value of 0.3871.
- Too much variance in cholesterol levels to say that the drug was effective, even though μ_X was less than μ_y.

22.3. A nonparametric test: The Wilcoxon Rank-Sum Test

- Like with the one-sample t-test, can build a Bayes Factor equivalent to two-sample test.
- Both t-test and Bayes Factor assume that data is normal.
- What if it is not?
- Can still say something about data: plot the points on a line:



- As order statistics, the drug group is

  $$2, 3, 4, 7, 8$$

  Are those numbers big or small?
- If they were random, then each order statistic would be a uniform draw from $\{1, \ldots, 10\}$. So on average each number would be $(1 + 10)/2 = 5.5$. So on average they would add up to $5 \cdot 5.5 = 27.5$. They actually add up to $2 + 3 + 4 + 7 + 8 = 24$. Is 24 small compared to 27.5?
- Note that 111 has order statistic 2 because there are exactly 2 numbers in both data sets less than or equal to 111 (itself and one other). This gives rise to the following definition of the statistic.

Definition 54
Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two samples. Then the rank of X_i is

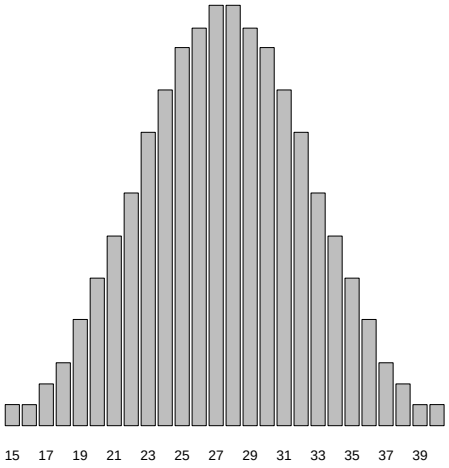
$$\text{rank}(X_i) = \sum_{j=1}^{n} \mathbb{1}(X_j \leq X_i) + \sum_{j=1}^{m} \mathbb{1}(Y_j \leq X_i).$$

Then Wilcoxon's rank-sum statistic is

$$W = \sum_{i=1}^{n} \text{rank}(X_i).$$

Fact 38
For W the Wilcoxon rank-sum statistic, $\mathbb{E}[W] = n(n+m+1)/2$.

- A two-sided test rejects when $|W - n(n+m+1)/2|$ is too big.
- If W is small then that provides evidence that $X < Y$, and if W is big, that provides evidence that $X > Y$.
- Is 24 small compared to 27.5? Note $|24 - 27.5| = 2.5$.
- Suppose I uniformly draw 5 out of 10 spots without replacement. What does the distribution of this random variable look like? Here's a plot of the density:



- Note that a good chunk of the probability is either at most 24 or at least 30. About 54% to be precise! This value can be estimated using Monte Carlo or approximating the distribution with a normal.

22.4. One-tailed versus two-tailed tests

- Two sided T-tests reject null when $|T|$ large, or $|W - n(n+m+1)|$ is large.
- If only care if $\mu_A > \mu_B$, reject null when T large, or W large.
- This is a one sided test.
- If only care if $\mu_A < \mu_B$, reject null when T small, or W small.

Problems

22.1: Suppose that a drug used for decreasing anxiety is tested on ten patients that are randomly divided into two groups. One group $(X_1, \ldots, X_n \sim X)$ receives the drug, while the other group $(Y_1, \ldots, Y_m \sim Y)$ does not.

Each group initially started with 5 participants, but one of the drug receiving patients left the study part way through. Over the next month, the number of anxiety attacks are recorded, and found to be

patients	1	2	3	4	5
X_i	13	14	17	22	
Y_i	24	30	15	23	24

(a) What should the null and alternate hypothesis be if the company is interested in testing if the drug decreases anxiety?

(b) What is the Wilcoxon rank sum for the data?

(c) What is the average of the Wilcoxon statistic given that your null hypothesis is true?

(d) Write the calculation of the p-value for the Wilcoxon test as p is equal to the probability of an event.

(e) If $p \approx 0.032$, would you reject your null hypothesis at the 5% level?

Chapter 23

Fisher Information

Question of the Day How much information does a single draw of a random variable give us about the distribution?

In this chapter

- Fisher information

Recall

- Suppose $X_1, X_2, \ldots \overset{\text{iid}}{\sim} X$.
- $\hat{\mu} = \bar{X}$ is unbiased estimator for $\mathbb{E}[X]$.

$$\begin{aligned}
\mathbb{V}(\hat{\mu}) &= \mathbb{V}\left(\frac{X_1 + \cdots + X_n}{n}\right) \\
&= \frac{\mathbb{V}(X_1 + \cdots + X_n)}{n^2} \\
&= \frac{\mathbb{V}(X_1) + \cdots + \mathbb{V}(X_n))}{n^2} \\
&= \frac{n\mathbb{V}(X)}{n^2} = \frac{\mathbb{V}(X)}{n}, \\
\mathrm{SD}(\hat{\mu}) &= \sigma/\sqrt{n}.
\end{aligned}$$

Example, coin flips

- Ex: if $X \sim \mathsf{Bern}(p)$, then $\sigma = \sqrt{p(1-p)}$, and

$$\mathrm{SD}(\hat{\mu}) = \sqrt{\frac{p(1-p)}{n}}.$$

- Can we make the standard deviation smaller?

$$\hat{\mu}_1 = \frac{n-1}{n}\hat{\mu}, \quad \mathrm{SD}(\mu_1) = \frac{n-1}{n}\,\mathrm{SD}(\hat{\mu}) < \mathrm{SD}(\hat{\mu})$$

  But $\hat{\mu}_1$ is biased!

- So can we do better with unbiased estimator?
- The answer is no!

23.1. Fisher information

- To understand the smallest standard deviation possible, need to know the amount of information in a single coin flip.

Definition 55
Let random variable X have a density $f_\theta(x)$ that depends upon a parameter θ, where $f_\theta(x)$ is differentiable with respect to θ except possibly at a countable number of places. Then the **score** of the random variable at x with $f_\theta(x) > 0$ is

$$S(x) = \frac{\partial \ln(f_\theta(x))}{\partial \theta} = \frac{\partial f_\theta(x)/\partial \theta}{f_\theta(x)}.$$

Example

- For $X \sim \mathsf{Bern}(p)$, $f_{X|p}(s) = p\mathbb{1}(s=1) + (1-p)\mathbb{1}(s=0)$. So

$$\ln(f_{X|p}(s)) = \ln(p)\mathbb{1}(s=1) + \ln(1-p)\mathbb{1}(s=0)$$
$$S(x) = \frac{1}{p}\mathbb{1}(s=1) - \frac{1}{1-p}\mathbb{1}(s=0).$$

- Here's something weird: suppose I draw $X \sim \mathsf{Bern}(p)$, and plug into the score:

$$Y = S(X)$$

Then $\mathbb{P}(Y = 1/p) = p$ and $\mathbb{P}(Y = -1/(1-p)) = 1-p$. So

$$\mathbb{E}[Y] = \mathbb{E}[S(X)] = (1/p)p + (1-p)(-1/(1-p)) = 1 - 1 = 0.$$

- This isn't a coincidence!

Definition 56
Say that a score is **regular** if for any function $g(s)$,

$$\int_s g(s)\frac{\partial f_\theta(s)}{\partial \theta}\, d\nu = \frac{\partial}{\partial \theta}\int_s g(s) f_\theta(s)\, d\nu.$$

Example: $[X|p] \sim \mathsf{Bern}(p)$ has a regular score. (The proof is beyond the scope of this course.)

Fact 39
For all θ, a regular score has $\mathbb{E}[S(X)|\theta] = 0$.

Proof. Then

$$\begin{aligned}\mathbb{E}\left[\frac{\partial f_\theta(X)/\partial\theta}{f_\theta(X)}|\theta\right] &= \int_s \left[\frac{\partial f_\theta(s)/\partial\theta}{f_\theta(s)}\right] f_\theta(s)\, d\nu \\ &= \int_s \partial f_\theta(s)/\partial\theta\, d\nu \\ &= \frac{\partial}{\partial\theta}\int_s f_\theta(s)\, d\nu \\ &= \frac{\partial}{\partial\theta}1 = 0.\end{aligned}$$

□

So that means for regular scores $\mathbb{V}(S(X)) = \mathbb{E}[S(X)^2]$.

Definition 57
The **Fisher information** of X with score $S(x)$ is $I(\theta) = \mathbb{E}[S(X)^2]$.

- The Fisher information of $[X|p] \sim \mathsf{Bern}(p)$ is

$$\mathbb{E}[Y^2] = p(1/p)^2 + (1-p)(-1/(1-p))^2 = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}.$$

- How much information is in $X_1, X_2, \ldots \overset{\text{iid}}{\sim} \mathsf{Bern}(p)$?

Fact 40
Suppose $X_1, X_2, \ldots$ are independent and use the same parameter θ. Then for all n,

$$S((x_1, \ldots, x_n)) = S(x_1) + S(x_2) + \cdots + S(x_n),$$

and if the scores are regular:

$$I_{(X_1,\ldots,X_n)}(\theta) = I_{X_1}(\theta) + I_{X_2}(\theta) + \cdots + I_{X_n}(\theta).$$

In particular, for $X_1, X_2, \ldots \overset{\text{iid}}{\sim} X$, $I_{(X_1,\ldots,X_n)}(\theta) = nI_{X_1}(\theta)$.

Proof. By independence:

$$\begin{aligned}\ln(f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)) &= \ln\left(\prod_{i=1}^{n} f_{X_i}(x_i)\right) \\ &= \sum_{i=1}^{n} \ln(f_{X_i}(x_i)),\end{aligned}$$

and since $\partial/\partial\theta$ is a linear operator,

$$S(x_1, \ldots, x_n) = \sum_{i=1}^{n} S(x_i).$$

Because the $X_1, \ldots, X_n$ are independent,

$$\mathbb{V}[S(X)] = \mathbb{V}\left[\sum_{i=1}^{n} S(X_i)\right] = \sum_{i=1}^{n} \mathbb{V}(S(X_i)).$$

□

- So for n flips of a coin, the Fisher information is n times the flips of a single coin.
- So $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Bern}(p)$, has information:

$$\frac{n}{p(1-p)}.$$

- Recall that $\mathbb{V}(\hat{\mu}) = p(1-p)/n$. The inverse of the Fisher information! That is not a coincidence!

Theorem 7 (Cramér-Rao Inequality)
Let $X = (X_1, \ldots, X_n)$ have Fisher information $I_X(\theta)$ for parameter θ, and a regular score function. Let $\hat{\theta}$ be an unbiased estimate for θ. Then

$$\mathbb{V}(\hat{\theta}) \geq 1/I_X(\theta).$$

- High information means low variance in unbiased estimator.
- Low information mean high variance in unbiased estimator.
- Note, for $X_1, \ldots, X_n \sim \mathsf{Bern}(p)$, lower and upper bounds match!

Definition 58
Given a random variable X with Fisher information $I_X(\theta)$ for parameter θ, an unbiased estimate $\hat{\theta}(X)$ is **efficient** if $\mathbb{V}(\hat{\theta}) = 1/I_X(\theta)$.

This is also known as a **uniformly minimum variance unbiased estimator (UMVUE)** of θ.

- For $X_1, X_2, \ldots \overset{\text{iid}}{\sim} \mathsf{Bern}(p)$, sample mean is efficient!
- Best you can do (in terms of standard deviation) for unbiased estimate.

Problems

23.1: True or false: Fisher information is always nonnegative when it exists.

23.2: Let $X \sim \mathsf{Gamma}(4, \lambda)$. Then X has density

$$f_{X|\lambda}(s) = \frac{\lambda^4}{6} s^3 \exp(-\lambda s) \mathbb{1}(s \geq 0).$$

This density is regular.

(a) What is the Fisher information in a single draw X about λ, $I_X(\lambda)$?

(b) What is the minimum variance of an unbiased estimator for λ? (Be sure to explain your answer.)

Chapter 24

The Crámer-Rao Inequality

Question of the Day Show that the sample mean $\bar{X}$ is an efficient estimate for μ, where $X_1, X_2, \ldots \overset{\text{iid}}{\sim} \mathsf{Pois}(\mu)$.

In this chapter

- Example of efficient estimators.
- Proof of Crḿer-Rao Inequality.

The Crámer-Rao inequality gives a lower bound on the variance of any unbiased estimator under mild regularity conditions. Essentially it says that the variance is lower bounded by the inverse of the Fisher information of the data. When this lower bound is reached by an estimator, we call the estimator *efficient*.

Poisson

- Density (with respect to counting measure) is $f_\mu(i) = \exp(-\mu)\mu^i/i!$. So
$$\ln(f_\mu(i)) = -\mu + i\ln(\mu) - \ln(i!),$$
partial differentiation with respect to μ gives score:
$$S(x) = i/\mu - 1.$$
Check: $\mathbb{E}[S(X)] = \mu/\mu - 1 = 0$.
- Fisher information is
$$\mathbb{V}(X/\mu - 1) = \mathbb{V}(X)/\mu^2 = \mu/\mu^2 = 1/\mu.$$
So best variance from n samples is μ/n.
- Here $\mathbb{V}(\bar{X}) = \mathbb{V}(X)/n = \mu/n$, the same! So the unbiased estimate $\bar{X}$ is efficient.

24.1. Proof of the Crámer-Rao inequality

Before getting into the proof, let's review some linear algebra.

Definition 59
An **inner product** $\langle x, y\rangle$ satisfies four properties (here x, y and z are vectors, and α is a scalar:

1: $\langle u + v, w\rangle = \langle u, w\rangle + \langle v, w\rangle$.

2: $\langle \alpha v, w\rangle = \alpha\langle v, w\rangle$

3: $\langle v, w\rangle = \langle w, v\rangle$

4: $\langle v, v\rangle \geq 0$ where equality holds if and only if $v = 0$.

You can also use inner products to form a norm.

Definition 60
An **inner product norm** has the form $\|v\| = \sqrt{\langle v, v \rangle}$.

Example: the L_2 norm comes from the dot product of vectors in $\mathbb{R}^n$.

Lemma 2 (The Cauchy-Schwarz inequality)
For any inner product and vectors x and y:

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \cdot \langle y, y \rangle.$$

or expressed using the inner product norm:

$$|\langle x, y \rangle| \leq \|x\| \cdot \|y\|.$$

Moreover, you only get inequality when there exists a scalar α such that $x = \alpha y$ or $y = \alpha x$.

Fact 41
In probability, random variables with finite second moment form a vector space where 0 is any random variable with $\mathbb{V}(X) = 0$, and the inner product is the covariance.

$$\langle X, Y \rangle = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))].$$

The inner product norm of a random variable is its standard deviation.

Proof of Crámer-Rao Inequality. Remember that $\hat{\theta} = \hat{\theta}(x)$ is an unbiased estimate of θ. So

$$\int_s \hat{\theta}(x) f_\theta(x) \, d\nu = \theta.$$

Differentiate both sides with respect to θ, and by regularity bring the partial derivative inside the integral.

$$\int_s \hat{\theta}(x) \frac{\partial f_\theta(x)}{\partial \theta} \, d\nu = 1.$$

Since $\mathbb{E}[S(X)] = 0$, $\mathbb{E}[\theta S(X)] = 0$, or in integral terms:

$$\int_s \theta \frac{\partial f_\theta(x)/\partial \theta}{f_\theta(x)} f_{\theta(x)} \, d\nu = 0.$$

Subtracting this last equation from the second to last gives:

$$\int_s (\hat{\theta} - \theta) \frac{\partial f_\theta(x)/\partial \theta}{f_\theta(x)} f_\theta(x) \, d\nu = 1.$$

Writing this in terms of expected value:

$$\mathbb{E}\left[(\hat{\theta} - \theta) \frac{\partial \ln(f_\theta(X))}{\partial \theta} \middle| \theta \right] = 1.$$

Since $\mathbb{E}[\hat{\theta} - \theta] = \mathbb{E}[S(X)] = 0$, Cauchy-Schwarz applies:

$$1 = \mathbb{E}\left[(\hat{\theta} - \theta) \frac{\partial \ln(f_\theta(X))}{\partial \theta} \middle| \theta \right]^2 \leq \mathbb{V}[\hat{\theta} - \theta | \theta] \mathbb{V}(S(X)).$$

Noting that $\mathbb{V}(\hat{\theta} - \theta | \theta) = \mathbb{V}(\hat{\theta})$ then gives the result.

□

- Since we used the Cauchy-Schwarz inequality in the proof, the only time you get equality is when you get equality in Cauchy-Schwarz, which means the vectors v and w are equal.
- In the probability setting any constant is the 0 vector, so inequality holds for Cauchy-Schwarz if $c_1X - c_2Y =$ constant.
- Our vectors are $\hat{\theta}(X) - \theta$ and $S(X)$. "Constants" here means any function of θ (since given θ, any function of θ is deterministic, not random.

Lemma 3
The unbiased estimator $\hat{\theta}$ is efficient if and only if it is of the form

$$\hat{\theta}(X) = a(\theta) + b(\theta)S(X).$$

24.2. A nonregular random variable

Where regularity fails

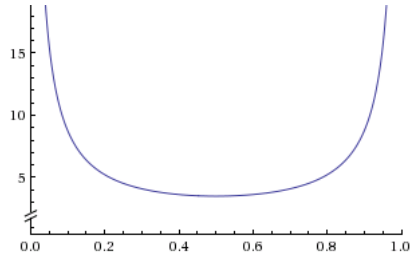
- Note not all random variables are regular.
- Recall: The *support* of a random variable is the values where the density is positive.
- Suppose the support depends on the parameter.
- Example: $[X|\theta] \sim \mathsf{Unif}([0,\theta])$, $f_X(s) = \theta^{-1}\mathbb{1}(s \in [0,\theta])$.
  - The score:

$$\begin{aligned} S(s) &= (\partial/\partial\theta)\ln(\theta^{-1})\mathbb{1}(s \in [0,\theta]) \\ &= (\partial/\partial\theta)(-\ln(\theta))\mathbb{1}(s \in [0,\theta]) \\ &= -\theta^{-1}\mathbb{1}(s \in [0,\theta)), \end{aligned}$$

    and is undefined when $s = \theta$.
  - So $\mathbb{E}(S(X)|\theta) = -\theta^{-1} \neq 0$. Hence $[X|\theta]$ cannot be regular!
- Get information from $X_1, \ldots, X_n$ about θ not just from mean.
- Recall MLE is $\max\{x_1, \ldots, x_n\}$, not $\bar{x}$ like earlier examples.

Extreme example

- Say $Y \sim \mathsf{Beta}(a,a)$ where a is close to 0, density is $f_Y(s) \propto \frac{1}{s^{1-a}(1-s^{1-a})}\mathbb{1}(s \in [0,1])$, so concentrates near 0 and 1.



- Now suppose $[X|\theta] \sim \theta Y$. Then the values of X are either close to 0 or close to θ. So $\max X_i$ is a very good estimate for θ.

- As $a \to 0$, X converges (in probability) to being uniform over $\{0, \theta\}$, so only have to wait for first large answer to determine θ with high accuracy.
- But even as $a \to 0$, $\mathbb{E}[X] = \theta/2$ and $\mathrm{SD}(X) \approx \theta/2$. So even though the standard deviation stays large relative to θ, the approximation is much better than Crámer-Rao would imply. Which you expect, because this is a nonregular distribution.

Problems

24.1: Suppose $\langle x, y \rangle = 4$. What is $\langle 3x, -2y \rangle$?

24.2: Suppose $\mathrm{Cov}(X, Y) = 4$. What is $\mathrm{Cov}(3X, -2Y)$?

24.3: Suppose that an unbiased estimator for parameter θ that uses data $x = (x_1, \ldots, x_n)$, has the form

$$\hat{\theta} = \theta^2 + \bar{x}/\theta.$$

Is the estimator efficient?

Chapter 25

Analysis of Variance

Question of the Day Suppose that a chain is considering three ad campaigns for a new product in their 14 stores. They randomly assign 5 stores to campaign A, 5 to B, and 4 to C. The amount of product sold is then:

A	B	C
11	14	26
23	17	13
09	16	24
10	16	19
12	8	

Is there enough evidence to suggest that one campaign is superior? In other words, can we reject the null hypothesis $H_0 = \mu_A = \mu_B = \mu_C$?

In this chapter

- Analysis of Variance (ANOVA)

The idea of *Analysis of Variance*, or ANOVA for short, was created by Fisher to analyze the difference (or lack thereof) between the group means of subjects that receive different treatments.

In a way, the idea of ANOVA can be though of as generalizing the t-test from earlier to more than two treatments. Since there are often cases (such as in the question of the day, when there are more than two possble ways to treat subjects, ANOVA is useful in determining if any of the treatments differs from the rest.

There are three main assumptions in the ANOVA framework:

1: The observed data for each subject is independent of the other observations.

2: These observations are normally distributed.

3: The variance of observations are all the same, which is called *homoscedasticity*.

Recall

- A model says that some variables are related to others. Example:

$$\text{income} \sim 1 + \text{age} + \text{state} + \text{age} : \text{state}.$$

- In a linear model $Y = X\beta + \epsilon$, that functional relationship was made explicit.
- ANOVA is an alternate approach to testing relationships in models.

Example

- Three stores try different price for a product.
- Each records sales volumes, model is volume $\sim 1+$ price
- Can we say one price is the best?
- Linear model: specific relationship between price and sales
- Both assume population normal, variance identical, independent.
- Look for differences in means given different prices.

Definition 61
In ANOVA, explanatory variables are called **factors**. The different values that the factors can take on are called **levels**. A specific factor and level value is called a **treatment**.

Qotd The sole factor is the ad campaign. For this factor the levels are A, B, or C. An example of a treatment is: the ad campaign used B. So there is one factor and three treatments in this table of data.

Notation 7
In a table of data, let n_j denote the number of experiments in treatment j. Let x_{ij} be the entry in the ith row and j column (so $i \in \{1, \dots, n_j\}$). A dot $\cdot$ is a wildcard character, so $x_{\cdot j}$ represents all the entries in column j of the data.

So for the qotd, $(n_A, n_B, n_C) = (5, 5, 4)$. Given this notation, we can now describe the overall mean of the entries.

$$\bar{x} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{\sum_{j=1}^{k} n_j} = 15.57.$$

Note that we can also find the means for each individual treatment:

$$(\bar{x}_{\cdot 1}, \bar{x}_{\cdot 2}, \bar{x}_{\cdot 3}) = (13, 14.2, 20.5).$$

Just from that, it could be that the last column has higher mean than the first two and the overall mean. But is there enough evidence to make that statistically significant?

To answer this question, Consider the variance in the x_{ij} entries. The overall population variance is:

$$\frac{1}{5+5+4-1} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2.$$

The $1/(5+5+4-1) = 1/13$ in front is just a constant, so we're not going to worry about it. The rest of it is called the total sums of squares:

$$SS_{\text{Total}} = SS_T = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2.$$

- The key idea of ANOVA is that we can break this sum of squares into a sum of squares from random variation (error) plus a sum of squares that comes from treatment effects.
- For instance, for treatment 1:

$$SS_1 = \sum_{i=1}^{n_1} (x_{ij} - \bar{x}_1)^2.$$

If we add those up over treatments, we get the sum of squares from within groups:

$$SS_W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{\cdot j})^2.$$

- Now we can also think of the group means as being a vector centered around $\bar{x}$. So they have a between groups sum of squares

$$SS_B = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{x}_{\cdot j} - \bar{x})^2 = \sum_{j=1}^{k} n_j(\bar{x}_{\cdot j} - \bar{x})^2.$$

25.1. Partitioning the sum of squares

Fact 42
For a one factor ANOVA,

$$SS_T = SS_W + SS_B.$$

Proof. It will help to have the following equation, which says that the average distance the entries of a column are away from their average value is 0:

$$\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{\cdot j}) = n_j\bar{x}_{\cdot j} - n_j\bar{x}_{\cdot j} = 0.$$

So

$$\begin{aligned}
SS_T &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2 \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{\cdot j} + \bar{x}_{\cdot j} - \bar{x})^2 \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{\cdot j})^2 + (\bar{x}_{\cdot j} - \bar{x})^2 + 2(x_{ij} - \bar{x}_{\cdot j})(\bar{x}_{\cdot j} - \bar{x}) \\
&= SS_W + SS_B + 2\sum_{j=1}^{k}(\bar{x}_{\cdot j} - \bar{x})\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{\cdot j}) \\
&= SS_W + SS_B.
\end{aligned}$$

□

- Let $N = \sum_{j=1}^{k} n_j$ be the total number of entries in the table.
- Assuming the data is normal, then $(SS_T/\sigma^2)\sqrt{N-1} \sim \chi^2(N-1)$.
- Say that SS_T has $N-1$ degrees of freedom.
- Similarly, SS_B has $k-1$ degrees of freedom.
- And SS_W has $N-k$ degrees of freedom.
- Put all this together to get an ANOVA table!

Definition 62
For a sum of squares SS with r degrees of freedom, let $MS = SS/r$ be the **mean square** of the statistic.

Recall that for the sum of squares of normals with variance σ^2 and df degrees of freedom: $(MS/\sigma^2)df \sim \chi^2(df)$. Putting this information into an ANOVA table gives something that looks like this.

Source of variation	df	MS
Among treatments/groups	$k-1$	$\hat{\sigma}^2 = [\sum_{j=1}^{k} n_j(\bar{x}_{\cdot j} - \bar{x})^2]/[k-1]$
Within treatments/groups	$N-k$	$s_p^2 = [\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{x}_{ij} - \bar{x}_{\cdot j})^2]/[N-k]$

Next time we will look at how to use the entries of this table to assess the null hypothesis!

Problems

25.1: Fill in the blank: A specific choice of level for every factor is called a _____.

25.2: The first factor has two levels, the second factor has 3. How many total possible treatments are there?

25.3: An experiment for student performance places students into a group given a soda with no caffeine but with sugar, coffee with caffeine but no sugar, or tea with neither sugar nor cafeine. Their scores on the exam are

Soda:	88	93	93	88	93
Coffee:	89	88	79	94	100
Tea:	90	90	88	91	

(a) Find the overall averages of the scores on the exam.

(b) Find the averages for each of Soda, Coffee, and Tea.

(c) Find SS_B, SS_W, and SS_T.

(d) Verify that $SS_T = SS_w + SS_B$.

Chapter 26

ANOVA: The F-statistic

Question of the Day Suppose that a product is considering three ad campaigns for a product in their 14 stores. They randomly assign 5 stores to campaign A, 5 to B, and 4 to C. The amount of product sold is then:

A	B	C
11	14	26
23	17	13
09	16	24
10	16	19
12	8	

Is there enough evidence to suggest that one campaign is superior? In other words, can we reject the null hypothesis $H_0 : \mu_A = \mu_B = \mu_C$?

In this chapter

- Using the ANOVA table to test hypothesis about the means of treatments.

26.1. Testing with ANOVA

- Need a model to proceed.

Definition 63
The **factor effects model** for data x_{ij} is

$$x_{ij} = \mu + \alpha_j + \epsilon_{ij}.$$

- So each entry in the table has a mean value μ, plus a mean value caused by being part of treatment j, plus a random error ϵ_{ij}.

- The α_j represent the effect of using treatment j.

- With the factor effects model, the null hypothesis that the means are the same can be rewritten as:

$$H_0 : \alpha_1 = \cdots = \alpha_k = 0.$$

- Recall our sums of squares:

$$SS_W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{\cdot j})^2$$

$$SS_B = \sum_{j=1}^{k} n_j (\bar{x}_{\cdot j} - \bar{x})^2.$$

  If the null is true, then we would expect SS_B to be small, and for SS_W to be big.

- More precisely, if $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma_\epsilon^2)$, then adding up variances gives:

$$\mathbb{E}[MS_W] = \sigma_\epsilon^2, \ \mathbb{E}[MS_B] = \sigma_\epsilon^2 + \sum_j n_j \alpha_j^2.$$

  So under the null hypothesis MS_B and MS_W are both expected to be σ_ϵ^2, so their ratio will be around 1. So let our test statistic $F = MS_B/MS_W$.

Definition 64
Let $Y_1 \sim \chi^2(d_1)$ and $Y_2 \sim \chi^2(d_2)$ be independent. Then call the distribution of $[Y_1/d_1]/[Y_2/d_2]$ an **F distribution with parameters d_1 and d_2**, and write $F \sim F(d_1, d_2)$.

Fact 43
Under the hypothesis that $\alpha_j = 0$ for all j in the factor effects model, $MS_B/MS_W \sim F(k-1, N-k)$.

To summarize this data, build an ANOVA table.

	df	Sum Sq.	Mean Square	F value	p-value
between blocks	$k-1$	SS_B	$MS_B = SS_B/(k-1)$	$F = MS_B/MS_W$	$\mathbb{P}(X > F)$
residuals	$N-k$	SS_W	$MS_W = SS_W/(N-k)$		

where $X \sim F(k-1, N-k)$ in the last column.

- For the qotd:

	df	Sum Sq.	Mean Square	F value	p-value
blocks	2	139.6	69.81	2.706	0.1107
residuals	11	283.80	25.800		

- So we say that the null hypothesis has a p-value of 11.07% based off of this data.

26.2. Completely randomized design

- There are always multiple ways to design an experiment.
- Some lead to easier analyses than others.

Definition 65
Design of experiments (DOE) is the study of how to run an experiment in order to test hypothesis concerning response and explanatory variables.

- In the QotD, the stores were randomly assigned to groups to use either treatment A, B, or C. This is the simplest form of what is called randomized block design.

Definition 66
In a **completely randomized design**, subjects are chosen uniformly at random to be part of the experiment, and subjects are assigned uniformly at random to different treatments.

- The point of random assignation is to get rid of the effect of other factors that might influence the results.
- Only completely random assignment can let us say that the effect comes from the treatment.
- Unfortunately, it is not always possible to have a randomized block design.
- If subjects select their own groups, not rbd
  - For example, if stores allowed to choose their own campaign, higher volume stores might choose C, would throw off results
  - Would believe that C causes higher sales, whereas it is the higher sales stores that are choosing C.

Problems

26.1: What statistics are produced by a one factor ANOVA table?

26.2: When using the F statistic, when do we reject the null hypothesis that the treatment leaves the mean effect unchanged?

26.3: True or false: In an ANOVA table the F statistics must have an F distribution even if the null hypothesis is not true.

26.4: An experiment for student performance places students into a group given a soda with no caffeine but with sugar, coffee with caffeine but no sugar, or tea with neither sugar nor caffeine. Their scores on the exam are

Soda:	88	93	93	88	93
Coffee:	89	88	79	94	100
Tea:	90	90	88	91	

The team decides to do an ANOVA analysis.

(a) For this data set, fill out the following:

Number of subjects =
Number of factors =
Number of treatments =

(b) Your research partner starts filling out an ANOVA table. Fill out the rest.

	df	Sum Squares	Mean Squares	F-statistic
drink		4.107		
Residuals		276.750		

(c) Let $\text{cdf}_{F(a,b)}$ denote the cdf of an F distributed random variable. Write the p-statistic for this table using this function.

(d) Calculate the p-statistic.

(e) The ANOVA analysis requires a major assumption about the distribution of residuals. Name the assumption and define what the assumption means.

26.5: A researcher wants to understand how much student belief affects exam scores. Before taking the exam, the students are made to watch a video that attempts to affect their confidence level. Some students watch an affirming video, others a discouraging video, and a third group a video which is neutral.

Their scores on the exam are

Boost:	8.8	9.2	8.1	9.5	
Discouraged:	9.6	4.5	6.0	7.1	
Neutral:	8.1	7.9	8.0	5.2	7.3

The team decides to do an ANOVA analysis.

(a) For this data set, fill out the following:

Number of subjects =
Number of factors =
Number of treatments =

(b) Fill out the following ANOVA table.

	df	Sum Squares	Mean Squares	F-statistic	p-statistic
video					
Residuals					

Chapter 27

Correlations

Question of the Day Consider the adult mass (in kg) and gestation period (in weeks) for several animals.

	Af. Elep.	Horse	Grizzly	Lion	Wolf	Badger	Rabbit	Squirrel
Adult mass	6000	400	400	200	34	12	2	0.5
Gestation	88	48	30	17	9	8	4.5	3.5

Are they independent?

In this chapter

- Estimating correlation.

Recall from probability

- The *covariance* between two random variables X and Y each with finite second moment is
$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$
- Covariance is an inner product, $\|X\| = \langle X, X\rangle^{1/2} = \text{SD}(X)$.
- Cauchy-Schwarz: $|\text{Cov}(X, Y)| \leq \text{SD}(X)\,\text{SD}(Y)$
- In geometry, the ratio
$$\frac{\langle X, Y\rangle}{\|X\|^{1/2}\,\|Y\|^{1/2}}$$
turns out to be the cosine of the angle between two vectors. In probability, we use this ratio to define the *correlation* between two random variables.
$$\text{Cor}(X, Y) = \frac{\langle X, Y\rangle}{\|X\|^{1/2}\,\|Y\|^{1/2}} = \frac{\text{Cov}(X, Y)}{\text{SD}(X)\,\text{SD}(Y)}.$$
- Vectors with $\theta = \tau/4$ are called *orthogonal*. For orthogonal random variables, $\text{Cor}(X, Y) = \cos(\tau/4) = 0$ are called *uncorrelated*.

Fact 44
If random variables X and Y with correlation $\text{Cor}(X, Y)$ are independent, then they are uncorrelated.

- Note that the inverse is not true: if two variables are uncorrelated, then they might still be dependent. The one exception is if they are bivariate normally distributed.

Fact 45
If (X, Y) are bivariate normal, then they are independent if and only if they are uncorrelated.

27.1. Estimating the correlation

- Suppose we have data $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$. If we knew the values $\mathbb{E}(X)$ and $\mathbb{E}(Y)$, then we could just estimate $\text{Cov}(X, Y)$ by

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbb{E}(X))(y_i - \mathbb{E}(Y)).$$

  But we don't know, so use $\bar{X}$ for $\mathbb{E}(X)$ and $\bar{Y}$ for $\mathbb{E}(Y)$. So have

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

  Now suppose we use our biased estimate for standard deviation:

$$\hat{\sigma}_X = \left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}, \ \hat{\sigma}_Y = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2}.$$

  Note, regular population variance unbiased for σ^2, still biased for σ anyway. So might as well not worry about n versus $n-1$. And this way the n's cancel out!

Definition 67
For data $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$, **Pearson's correlation coefficient** r is

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

The following fact lends credence to this choice of estimate.

Fact 46
If (X, Y) are bivariate normally distributed, then r is the MLE for $\text{Cor}(X, Y)$.

Qotd

- Let's try this out in R:

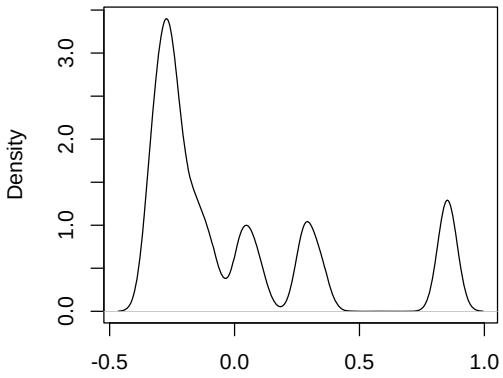
```
mass <- c(6000,400,400,200,34,12,2,0.5)
gest <- c(88,48,30,17,9,8,4.5,3.5)
cor(mass,gest)
```

- That returns an r of 0.8923541.

27.2. Confidence intervals for r

- Fisher noted that if the X vector is truly independent of the Y vector, then it is also independent of any permutation of Y.
- For $n = 8$, $n! = 40320$, not too bad, but for $n = 100$, $n! = 9.332 \cdot 10^{157}$, so won't be able to test all permutations when n is large
- Use Monte Carlo!

```
rsamp <- replicate(100000,cor(mass,sample(gest,length(gest))))
plot(density(rsamp))
```




```
sum(rsamp<cor(mass,gest))/length(rsamp)
```

returns 0.99996, so we have very strong evidence from this data set that the correlation does not equal 0.

27.3. The coefficient of determiniation R^2

- Suppose that we have a simple model:

$$y_i = c_0 + c_1 x_i + \epsilon_i.$$

- Then the *residual sum of squares* is:

$$SS_{\text{res}} = \sum_i \epsilon_i^2 = \sum_i (y_i - (c_0 + c_1 x_i))^2$$

- If I just knew about the y_i values, I could get the *total sum of squares*, which measures the sum of the squares of the distances of the y_i values from their mean:

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2.$$

[Note $\hat{\sigma}^2 = SS_{\text{tot}}/(n-1)$.]

- If $c_1 = 0$, then the least square estimator for $c_0 = \bar{y}$, and

$$SS_{\text{res}} = SS_{\text{tot}}.$$

If c_1 is allowed to be nonzero, we can fit the y_i better, and

$$SS_{\text{res}} \leq SS_{\text{tot}}.$$

Definition 68
The **coefficient of determination** has value for pairs of points $(x_i, y_i)_{i=1}^n$ is

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}.$$

Fact 47
The value of R^2 is just the square of the Pearson's sample correlation coefficient. [Hence the name R^2.]

Problems

27.1: True or false: Two random variables with positive correlation cannot be independent.

27.2: For X with finite second moment, what is $\text{Cor}(X, X)$?

27.3: If $R^2 = 0.36$, what is Pearson's r?

27.4: True or false: For data $\{X_i\}$ and $\{Y_i\}$ drawn iid from distributions with finite mean, variance, and covariance, Pearson's r converges to the true correlation as the number of sample points goes to infinity.

27.5: If $Y = 3X + 3$, what is $\text{Cor}(X, Y)$?

27.6: True or false.

(a) Covariance is an inner product.

(b) Correlation is an inner product.

27.7: True or false: If $U_1, \ldots, U_n \sim \mathsf{Unif}([0, 1])$ where the U_i are independent, then $U_1^2 + \cdots + U_n^2 \sim \chi^2(n)$.

27.8: Suppose that Z_1 and Z_2 are independent, standard normal random variables. Let $X_1 = (1/\sqrt{2})Z_1 + (1/\sqrt{2})Z_2$ and $X_2 = Z_1$.

(a) What is the distribution of X_1?

(b) What is the distribution of X_2?

(c) True or false: The distribution of $X_1^2 + X_2^2$ is $\chi^2(2)$.

27.9: Find the Pearson's correlation coefficient for

$$(1.1, 0.4), (-3.2, 4.6), (0.1, 5.1).$$

Chapter 28

Contingency Tables

Question of the Day Suppose that a survey of 100 adults that use social media between the ages of 30 and 40 reveals that their prefered social media platform is

Platform	Percentage
Facebook	42
Instagram	32
Twitter	26

Is this enough evidence to reject the null hypothesis that a user is equally likely to pick each of the three possibilities?

In this chapter

- Contingency tables

Definition 69
Let $\vec{x}$ be a vector of data subject to linear constraints:

$$Ax \geq b.$$

Then the data is called a **contingency table**.

- Example: linear constraint $x_1 - 2x_2 \geq 4$.
- Example nonlinear constrain $x_1^2 \geq 4$.
- Includes $\leq$ constraints as well ($x_1 + x_2 \geq 2 \Leftrightarrow -x_1 - x_2 \leq -2$).
- Includes $=$ constraints as well:

$$x_1 + x_2 = 2 \Leftrightarrow (x_1 + x_2 \leq 2) \text{ and } (x_1 + x_2 \geq 2).$$

- Note that the QotD data has linear constraints:

$$\begin{aligned} x_1 + x_2 + x_3 &= 200 \\ x_1 &\geq 0 \\ x_2 &\geq 0 \\ x_3 &\geq 0. \end{aligned}$$

Recall

- Suppose outcomes of each experiment are 0 or 1, number of experiments is n, and probability of a 1 is p.
- Then number of times 1 comes up is $\mathsf{Bin}(n, p)$
- When each experiment can return $1, 2, \ldots, k$, call vector of results *multinomial*.

Definition 70
Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} X$, where $X \in \{1, 2, \ldots, k\}$, and $p_i = \mathbb{P}(X = i)$. Let $Y_i = \sum_{j=1}^{n} \mathbb{1}(X_j = i)$. Then $(Y_1, \ldots, Y_k)$ has a **multinomial distribution** with parameters $p_1, \ldots, p_k$ and n. Write

$$(Y_1, \ldots, Y_k) \sim \mathsf{Multinom}(n, p_1, \ldots, p_k)$$

The following fact follows immediately from the definition.

Fact 48
Say $(Y_1, \ldots, Y_k) \sim \mathsf{Multinom}(n, p_1, \ldots, p_k)$. Then for any i, $Y_i \sim \mathsf{Bin}(n, p_i)$.

- Note the Y_i are not independent since $Y_1 + \cdots + Y_k = n$.
- In Qotd can state the null hypothesis using multinomial notation:

$$H_0 : (X_1, X_2, X_3) \sim \mathsf{Multinom}(200, 1/3, 1/3, 1/3)$$

From our fact:

$$\begin{aligned}\mathbb{E}[(X_1, X_2, X_3)] &= (200/3, 200/3, 200/3) \\ &= (66.66\ldots, 66.66\ldots, 66.66\ldots)\end{aligned}$$

- One way to check if data unusual is to measure how far away it is from expected value.

Definition 71
The **chi-squared** statistic for a vector $(x_1, \ldots, x_n)$, which has expected value $(\mu_1, \ldots, \mu_n)$ under the null hypothesis is

$$\chi^2(x_1, \ldots, x_n) = \sum_i \frac{(x_i - \mu_i)^2}{\mu_i}.$$

Since x_i and μ both have the same units, so does the χ^2 statistic.

Question of the Day In the question of the day, under the null hypothesis each of the three choices is equally likely to be picked. That means that the expected values for each entry in the contingency table are

$$\mu_i = 100\frac{1}{3} = 33.33\ldots$$

for all i. Hence the χ^2-statistic for this data is

$$\chi^2 = \left[(42 - 100/3)^2/(100/3) + (32 - 100/3)^2/(100/3) + (26 - 100/3)^2/(100/3)\right] = 3.92.$$

The next question that we always ask with a test statistic: is that value unusually big? One way to answer is to repeat the experiment of drawing from the test distribution many times and look at the mean. Consider drawing 10000 times from the $\mathsf{Multinom}(100, 1/3, 1/3, 1/3)$ distribution and calculating the chi-squared statistic as follows.

```
qotd <- function(n=100) {
  choice <- sample(c("f","i","t"),n,prob=c(1/3,1/3,1/3),replace=TRUE)
  x <- c(sum(choice=="f"),sum(choice=="i"),sum(choice=="t"))
  mu <- 100/3
  return(sum((x-mu)^2/mu))

}
results <- replicate(100000,qotd())
cat(mean(results >= 3.92),"ś",sd(results >= 3.92)/sqrt(length(results)),"\n")
```

The result for my run was something like 0.146 ± 0.001 as an estimate for the p-value, which indicates that with 100 people sampled there is not enough data at the 5% level to reject.

Now suppose that 200 people have been surveyed, and the table of data is $(84, 64, 52)$. In this case, $\chi^2 = 2 \cdot 3.92 \ldots = \ldots$, and 1000 replications gives an estimate of the exact p-value estimate of

$$0.0197 \pm 0.0005.$$

If this had been the original table, then we would have had enough information to reject the null at the 5% level.

28.1. Using χ^2 to test goodness of fit

- What if we hadn't had access to a computer for Monte Carlo simulation? Then the following would have had to do:

Fact 49
Suppose $X_1, X_2, \ldots \overset{\text{iid}}{\sim} X$, where $\mathbb{P}(X = i) = p_i$ for $i \in \{1, 2 \ldots, k\}$. For all i and n, let

$$Y_i^n = \sum_{j=1}^{n} \mathbb{1}(X_j = i).$$

Then for

$$\chi_n^2 = \sum_{i=1}^{n} \frac{(Y_i - np_i)^2}{np_i},$$

as $n \to \infty$, $\mathbb{P}(\chi_n^2 \leq a) \to \mathbb{P}(A \leq a)$ where $A \sim \chi^2(n-1)$.

- For QotD, $k = 3$, so for $A \sim \chi^2(2)$,

$$\mathbb{P}(A \geq 3.92) = 0.1408584,$$

close to what was found by Monte Carlo.

- For the size 200 sample data set,

$$\mathbb{P}(A \geq 2 \cdot 3.92) = 0.01984109,$$

again well within the bounds given by Monte Carlo.

28.2. General contingency tables

Consider a data set from the following paper:

> Chase, M.A and Dummer, G.M. (1992), "The Role of Sports as a Social Determinant for Children," *Research Quarterly for Exercise and Sport*, 63, 418-424

Students in grades 4-6 were asked whether good grades, popularity, or sports was the most important to them. The results:

	4	5	6	Total
Grades	49	50	69	168
Popular	24	36	38	98
Sports	19	22	28	69
Total	92	108	135	335

- Null hypothesis: the grade level and choice of Grades, Popular, Sports are independent of each other.
- We can view this as one big contingency table with $k = 9$ entries:

	4	5	6	Total
Grades	x_{11}	x_{12}	x_{13}	r_1
Popular	x_{21}	x_{22}	x_{23}	r_2
Sports	x_{31}	x_{32}	x_{33}	r_3
Total	c_1	c_2	c_3	

- Linear constraints include things like $x_{11} + x_{12} + x_{13} = r_1$ and $x_{12} + x_{22} + x_{32} = c_2$.
- Let's say grade level and choice of grades, popularity, or sports were independent. Then for instance, the chance a student was both grade 5 and said popular would be $(108/335)(98/335)$. Then the expected number of entires in the cell is $335(108/335)(98/335)$ That makes the expected table:

	4	5	6	Total
Grades	46.13	54.16	67.70	
Popular	26.91	31.59	39.49	
Sports	5.21	22.24	27.80	

- Next calculate χ^2 value:
$$\frac{(49 - 46.13)^2}{46.13} + \cdots + \frac{(28 - 27.80)^2}{27.80} = 1.51$$
- Next degrees of freedom. It's not $rc = 3 \cdot 3 = 9$ because the rows and columns have to add up to their row and columns sums. So that would make is $9 - 3 - 3 = 3$. However, the row sums and column sums add up to the same thing, so one of those equations is redundant! Only 5 of the equations are linearly independent, so the degrees of freedom is $rc - r - c + 1 = 9 - 3 - 3 + 1 = 4$. For $A \sim \chi^2(4)$:
$$\mathbb{P}(A \geq 1.51) = 0.8248,$$
so not enough evidence to reject the null hypothesis that the grade level and choices are independent.

Problems

28.1: In a contingency table, data are subject to what kind of contraints?

28.2: Suppose that $(X_1, X_2, X_3) \sim \mathsf{Multinom}(3, 0.2, 0.5, 0.3)$, what is the chance $X_2 = 3$?

28.3: An auditor is checking glucose levels at two hospitals The glucose of each subject can be high (H), medium (M), or low (L). They gathered the following data.

	H	**M**	**L**	Total
Hospital 1	26	29	45	100
Hospital 2	44	26	30	100
Total	70	55	75	200

They want to test whether the glucose level is independent of where the patient is. Describe how you would test this at the 5% level, being sure to state your null hypothesis, test statistic (which you should calculate for this data), and rejection region (which you can write using a cdf or cdf^{-1} function, you do not have to calculate it exactly.)

Chapter 29

Nonparametric ANOVA and Correlation

Question of the Day Suppose that we have a factor with two or more levels. How can we decide if the medians are independent of the type. Recall ad campaign data:

A	B	C
11	14	26
23	17	13
09	16.1	24
10	16.2	19
12	8	

Can we reject the null hypothesis $H_0 : \text{median}(X_A) = \text{median}(X_B) = \text{median}(X_C)$?

In this chapter

- Kruskal-Wallis test

29.1. Nonparametric form of ANOVA

- Before, we used ANOVA to analyze this type of data.
  - Assumed observations independent, normal, and variance all the same.
- New method, do not need normality assumption.
- I've got 14 pieces of data, give each rank 1 (smallest) to rank 14 (largest)

data	11	23	09	10	12	14	17	16.1	16.2	8	26	13	24	19
rank	4	12	2	3	5	7	10	8	9	1	14	6	13	11
group	A	A	A	A	A	B	B	B	B	B	C	C	C	C

- For an average rank for each of the three groups, A, B, and C:

$$r_A = \frac{4+12+2+3+5}{5},\ r_B = \frac{7+10+8+9+1}{5},\ r_C = \frac{14+6+13+11}{4}.$$

Evaluating:

$$r_A = 5.2,\ r_B = 7, r_C = 11.$$

Notice that the total average rank is

$$\frac{1+2+\cdots+14}{14} = \frac{14+1}{2} = 7.500$$

- So now the question is, are $(5.2, 7, 11)$ far enough away from $(7.5, 7.5, 7.5)$ to reject null that the type A, B, C does not matter?
- Need a statistic that measures how far away things are from mean. Use sum of squares approach:

$$H = (N-1)\frac{\sum_{i\in\{A,B,C\}} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{14}(r_i - \bar{r})^2}$$

Note that the denominator is actually independent of the data, it only depends on the number of subjects 14. We can work out the denominator exactly to get the Kruskal-Wallis test statistic.

Definition 72
Kruskal-Wallis The **Kruskal-Wallis test statistic** for a table that has N entries and one factor with k levels, is

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{k} n_i(\bar{r}_i - \bar{r})^2.$$

- Note that $H \geq 0$, and when H is big, data is far from average.
- For qotd

$$\begin{aligned} H &= \frac{12}{(14)(14+1)}\left[5(5.2-7.5)^2 + 5(7-7.5)^2 + 11(7-11)^2\right]. \\ &= 11.64 \end{aligned}$$

- Is that big enough to reject H_0?

Fact 50
Under the null that all entries of the table have the same median, consider $n_{\min} = \min\{n_i\}$. As $n_{\min}$ goes to infinity, H approaches a chi-square with $k-1$ degrees of freedom.

- The probability that $X \sim \chi^2(3)$ has $X > 11.64$ is only 0.002967605, so reject null hypothesis at 5% level.

29.2. Nonparametric correlation

One of the disadvantages of a parametric correlation estimate is that it is not robust in the sense that a single outlier can sharply move the estimate away from the true value. Nonparametric correlation estimates can prevent this from happening.

- Suppose I have data pairs $\{(x_i, y_i)\}_{i=1}^n$.
- Are they positively related? Negatively related? Neither?
- If they are positively related, then it should be that if $x_i < x_j$, it should be more likely that $y_i < y_j$.
- If negatively related, then the reverse should be true.

Definition 73
For points (x_i, y_i) and (x_j, y_j) where $x_i < x_j$, $y_i < y_j$ call the points **concordant**, if $y_i > y_j$ call the points **discordant**.

Then Kendall's tau is the number of concordant point pairs minus the number of discordant point pairs all over the number of point pairs.

Definition 74
Let $\{(x_i, y_i)\}_{i=1}^n$ be a collection of point pairs. For $1 \leq i < j \leq n$, let $p(i,j) = (x_j - x_i)(y_j - y_i)$. Note that for the line that passes through (x_i, y_i) and (x_j, y_j), the slope is positive if $p(i,j) > 0$ and negative if $p(i,j) < 0$. Then **Kendall's Tau** is

$$\tau = \frac{\sum_{1 \leq i < j \leq n} [\mathbb{1}(p(i,j) > 0) - \mathbb{1}(p(i,j) < 0)]}{n(n-1)/2}.$$

Fact 51
Kendall's tau (like the Pearson sample correlation) falls between 1 and -1.

Spearman Rank Order Correlation Coefficient

- There are many ways to create a nonparametric correlation coefficient.
- Another way is to just replace the points with there ranks.
- Example:

$$(2.2, 1.1),\ (3, 4),\ (3.8, 2.5)$$
$$2.2 < 3 < 3.8 \Rightarrow \text{ranks of } x_i \text{ are } 1, 2, 3$$
$$1.1 < 2.5 < 4 \Rightarrow \text{ranks of } y_i \text{ are } 1, 3, 2$$
$$\text{Rank points:} (1,1), (2,3), (3,2).$$

S.R.O.C.C. just finds the Pearson correlation of $(1,1), (2,3), (3,2)$.

$$r((1,1), (2,3), (3,2)) = 0.5.$$

Definition 75
Consider points $(x_i, y_i)_{i=1}^n$. Let r_i be the rank of x_i and s_i the rank of y_i. Then **Spearman's rho** is the Pearson sample correlation coefficient of $(r_i, s_i)_{i=1}^n$.

- For the above data:

Method	Correlation Coefficient
Pearson	0.482663
Kendall	0.333333
Spearman	0.500000

Which is better?

- There is no clear answer which is better, all three are widely used.
- It is easier to generalize Kendall's Tau to data sets with missing data.
- Example: Efron and Petrosian studied Quasar data. They wanted to know if the brightness of quasars was correlated with how far away the quasars were. But a quasar that is both dim and far away will be difficult to see. So this gives truncated data. General relativity considerations actually truncated the data on both ends, but the point is you can still define Kendall's tau for this truncated data, but not Spearman's rho.

Problems

29.1: True or false: Pearson, Kendall, and Spearman correlation coefficients will always be the same for independent data.

29.2: Consider the following three points:

$$(0.4, 0.6), (0.7, 0.5), (1.2, 1.1).$$

(a) Find Pearson's r

(b) Find Spearman's Rho

(c) Find Kendall's Tau

29.3: Consider the following three points:

$$(0.4, 0.7), (0.7, 1.1), (1.2, 0.4).$$

(a) Find Pearson's r

(b) Find Spearman's Rho

(c) Find Kendall's Tau

29.4: Consider the following four data points:

$$(0.3, 1.2), (0.5, 2.4), (0.7, 1.7), (0.9, 2.0).$$

(a) Calculate the Pearson's correlation coefficient for this data.

(b) Calculate Kendall's Tau for this data.

(c) Calculate Spearman's rho for this data.

(d) Now suppose that the last data point $(0.9, 2.0)$ is replaced with $(0.9, 10.0)$. Repeat the calculation for Pearson's r, Kendall's tau and Spearman's rho.

Chapter 30

Multiterm ANOVA

Question of the Day I have data on the ages of 49 men and women married in Mobile County, Alabama. In general, are the men getting married older than the woman?

In this chapter

- Hypothesis Testing on Parts of Models

Recall

- Have response variables and explanatory variables.
- Some explanatory variables are nuisances: don't want to

Definition 76
An explanatory variable in a model that we are not interested in is called a **covariate**.

- What is a covariate depends on the discretion of modeler.
- Example: If we are testing lung cancer versus smoking amount, we might consider income a covariate. It might affect the rate of lung cancer, but it is not the relationship we are trying to uncover.
- In ANOVA, need to separate out variables to see what is happening individually with each.

30.1. Adding variables to an ANOVA reduces $SS_{\text{residuals}}$

Marriage records data Consider a set of 48 marriage records from Mobil County in Alabama. This data set was taken from https://roam.probate.mobilecountyal.gov/ late in the year 1996. The first few records in the data set look like

BookpageID	Person	Age
B230p539	Bride	28.7
B230p539	Groom	32.6
B230p677	Bride	52.6
B230p677	Groom	32.3
⋮	⋮	⋮

We can go ahead an fit a simple linear model `Age` $\sim$ `Person`

```
marriage <- read.csv("marriage.csv",header=TRUE)
model1 <- lm(Age ~ Person,data=marriage)
summary(model1)
```

to get from R

	Estimate	Std. Error	t value	$\mathbb{P}(>\|t\|)$
Intercept	33.239	2.060	16.133	$< 2 \cdot 10^{-16}$
PersonGroom	2.546	2.914	0.874	0.384

Note that I asked for `Age` versus `Person` and the variable it returned was `PersonGroom`. That indicates that it turned the binary variable `Person` into 1 if the value was `Groom`, and 0 if it was `Bride`.

- So on average it added 2.546 if the person was the Groom versus a Bride. In other words, the men were on average about 2.5 years older than the women getting married.
- But the standard error is 2.914, bigger than the 2.546! So we don't have strong evidence that the `PersonGroom` coefficient is bigger than 0. [The p-value of 38.4% reflects this ambivalence.]

Analyzing this using an ANOVA table gives a similar result

```
anova(model1)
```

gives

	df	Sum Sq.	Mean Sq.	F value	$\mathbb{P}(>F)$
Person	1	158.8	158.75	0.7633	0.3845
Residuals	96	19967.5	208.00		

Note that both the t statistic $(\hat{\mu}/\hat{\sigma})(n-1)$ and the F statistic $(SS_{\text{person}}/1)/(SS_{\text{residuals}}/96)$ are exact statistics under the model that the residuals are normal and homoscedastic, so both give the same p-value in the end.

We can do better!

- So with this basic analysis, we do not have enough evidence to say that men on average have higher age than the women.
- However, these previous analyses ignore the fact that we know which pairs are getting married!
- Let's use `BookpageID` as part of the model to compare within the married couple what is happening:

```
model2 <- lm(Age ~ Person + BookpageID,data=marriage)
anova(model2)
```

	df	Sum Sq.	Mean Sq.	F value	$\mathbb{P}(>F)$
Person	1	158.8	158.75	9.0699	0.004137
BookpageID	48	19127.3	398.49	22.7661	$< 2.2 \cdot 10^{-16}$
Residuals	48	840.2	17.50		

So what changed?

- Well, the `BookpageID` has 48 degrees of freedom! So it soaks up a huge amount of variance! The sum of squares and mean squares for `Person` is unchanged from before, but the residual sum of squares is much smaller, making the
$$F = \frac{MS_{\text{Person}}}{MS_{\text{residuals}}}$$
much larger. That makes the p-value much smaller.
- Note: this ANOVA analysis was what earlier we called a t-test for paired data. So we could have used that approach rather than ANOVA here.

30.2. Order matters in multiterm ANOVA

Now one thing to be aware of when doing an anova with more than one factor is that the order in which you put the terms will make a difference in the analysis.

- Consider records for the 100 meter freestyle from 1905 to 2004.
- Consider model:

$$\text{time} \sim \text{year} + \text{sex} + \text{year:sex}$$

- Can make an ANOVA table using R

```
swim <- read.csv("swim100m.csv",header=TRUE)
mod1 <- lm(time~year+sex+year:sex,data=swim)
anova(mod1)
```

	df	Sum Sq	Mean Sq	F value	Pr(¿F)
year	1	3578.6	3578.6	324.738	$< 2.2 \cdot 10^{-16}$
sex	1	1484.2	1484.2	134.688	$< 2.2 \cdot 10^{-16}$
year:sex	1	26.7	296.7	26.922	$2.826 \cdot 10^{-6}$
Residuals	58	639.2	11.0		

- Suppose that I thought gender was more important than year. Then I could put `sex` before `year` in the model:

```
mod2 <- lm(time~sex+year+year:sex,data=swim)
```

This actually changes the ANOVA! It does not just swap the coefficients

	df	Sum Sq	Mean Sq	F value	$\mathbb{P}(> F)$
sex	1	1720.7	1720.7	156.141	$< 2.2 \cdot 10^{-16}$
year	1	3342.2	3342.2	303.286	$< 2.2 \cdot 10^{-16}$
year:sex	1	26.7	296.7	26.922	$2.826 \cdot 10^{-6}$
Residuals	58	639.2	11.0		

To be clear, the least squares linear model hasn't changed:

```
coef(mod1)
coef(mod2)
```

Both give:

$$697.3 - 0.3240459 \cdot \text{year} - 302.4638388 \cdot \text{sexM} + 0.1499166 \cdot \text{year} \cdot \text{sexM}$$

- But by putting the `sex` variable first, it soaks up more of the sum of squares than it did before.

Summary

- If I start with variable `a` in ANOVA and add variable `b`, the sum of squares for `a` stays the same, sum of squares for `b` rises from 0 to some positive number, which means the sum of squares of the residuals goes down. So that can lead to saying `a` is statistically significant where it wasn't before.
- If I have variable `a` and `b`, and `a` comes before `b` in the model, then if I switch the order of `a` and `b`, that can raise the sum of squares of `a` and lower the sum of squares of `b`, again possibly changing the p-value associated with each. Recommendation: put the variable you are actually interested in (the explanatory variable) in the model first, then put covariates.

Problems

30.1: True or false?

(a) Changing the order of terms in the model can change the least squares fit.

(b) Changing the order of terms in the model can change the p-value for the terms in the ANOVA.

Chapter 31

Causality

Question of the Day How can we design experiments to test causality?

In this chapter

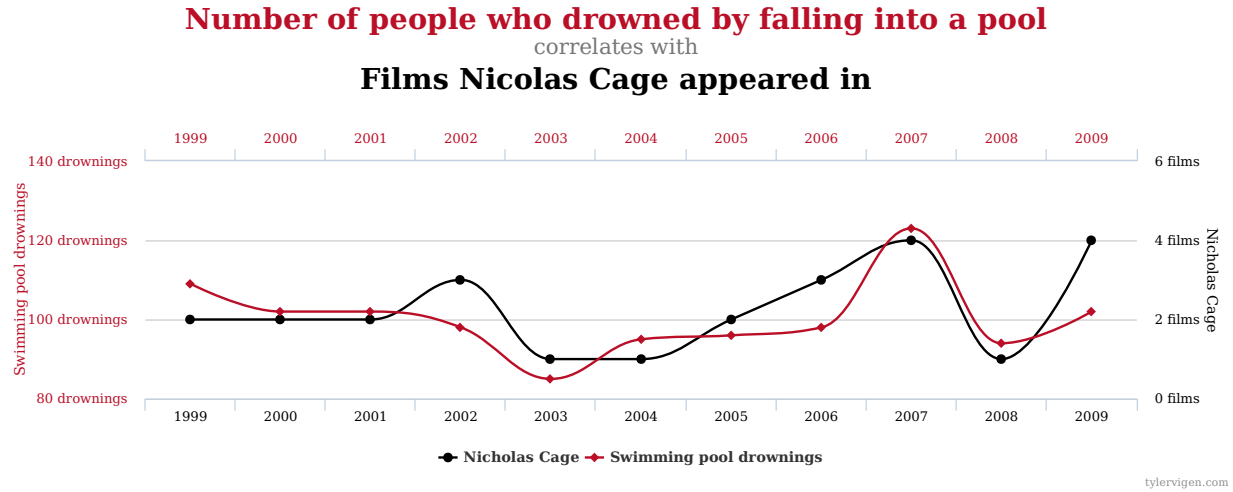
- Determining causal relationships from data.

Definition 77
Causal inference is the study of how to determine if effect A causes B.

Fact 52
If A and B are uncorrelated, then A does not cause B.

Correlation does not prove causation!

- Children with big feet spell better
  - Older children have bigger feet
  - Older children spell better
- The number of people who drowned by falling into a pool correlates very well with number of films Nicolas Cage appeared in from 1999 to 2009. [This picture came from Tyler Vigen's excellent website: Spurious Correlations.]



- So how can we prove correlation?
  - Does smoking cause lung cancer?
  - Does high cholesterol cause heart disease?
  - Do CO_2 emissions drive Climate Change?

31.1. Showing causality through experimental design

Completely randomized design

- Example: Medical study. 500 patients enrolled to see if a drug lowers the rate of heart disease. Each patient will be given either the drug or a placebo.
  - For each patient, flip a fair coin. Heads they go into treatment, tails they get placebo.
  - Random assignment eliminates other factors.
- Problem: suppose that the drug works better for men than for women.
  - If more women than men are assigned to the drug, the results might be affected.

	Drug	Placebo
Men	125	125
Women	125	125

Randomized block design

- Men/Women about 50%/50% of population.
- Try to replicate that 1:1 ratio
  - In choice of patients to enroll in the study
  - In who gets the drug versus the placebo.

	Drug	Placebo
Men	125	125
Women	125	125

- Could add extra factors like family history of heart disease, etcetera.

Definition 78
In a **randomized block design**, the experiment divides the participants into **blocks** based on one or more properties so that the percentage of participants in each block reflects as close as possible the percentage of the population as a whole. Each member of a block is then randomly assigned to a treatment so that the number receiving each level of treatment is as equal as possible.

31.2. Proving causality when you can't control treatment

- Can't inject people with cholesterol to see if they get heart disease.
- Can't control CO_2 levels to test climate change.
- Still there are things to look for.
- Throughout, assume A and B are positively correlated.

Mechanism for causation

- Most important: is there a plausible mechanism by which A could cause B?
  - Smoking causes lung cancer needed cellular pathology experiments to verify that carcinogen's caused cellular change.
  - Climate change requires global modeling to argue how 400 parts per million can cause global effects.
- Second: is there no plausible mechanism by which B could cause A?
- Third: is there no plausible C which could cause both A and B?
  - Ex: are low grades caused by malnutrition? Or does poverty cause both low grades and malnutrition? To answer, must include C in model:

$$A \sim 1 + B + C.$$

Hill Criteria Applies to epidemiology studies

1: Temporal Relationship: Does A always precede B in time?

2: Strength of correlation

3: Dose-response relationship

4: Consistency (replicated across studies and settings)

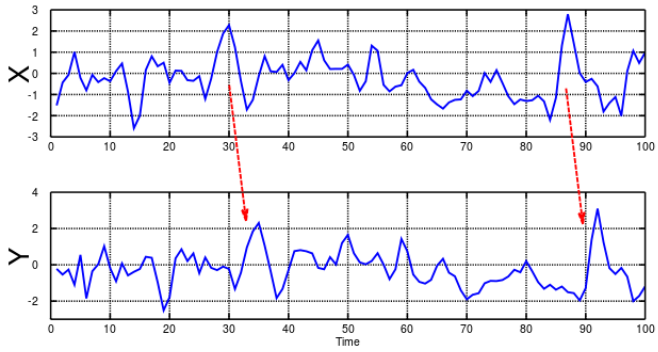
5: Plausibility (some theoretical basis for cause)

6: Consideration of Alternate Explanations

7: Experiment: B can be stopped by just stopping A

Granger

- Economist
- Concentrated on Temporal relationship.
- Compared two time series offset by time:



- Won the Nobel for his work in 2003.

Example: What causes achievement at elite schools?

The following paper tried to discover causal relationships between achievement and attendance at a school.

> A. Abdulkadiroglu, J. Angrist, and P. Pathak, The Elite Illusion: Achievement effects at Boston and New York Exam Schools, *Econometrica*, Vol. 82, No. 1 (January, 2014), 137–196

The authors realized that a controlled experiment would be best...

> An ideal experiment designed to reveal causal effects of peer characteristics would randomly assign the opportunity to attend schools with high-achieving peers and fewer minority classmates. The subjects of such a study should be a set of families likely to take advantage of the opportunity to attend schools that differ from their default options. Imagine sampling parents found in suburban Boston real estate offices, as they choose between homes in Newton and Waltham.We might randomly offer a subset of those who settle for Waltham a voucher that entitles them to send their children to Newton schools in spite of their choice of a Waltham address. This manipulation bears some resemblance to the Moving to Opportunity (MTO) experiment, which randomly allocated housing vouchers valid only in low-poverty neighborhoods. MTO was a complicated intervention, however, that did not manipulate the school environment in isolation (see Kling, Liebman, and Katz (2007) and Sanbonmatsu, Ludwig, Katz, Gennetian, Duncan, Kessler, McDade, and Lindau (2011)).

So instead, the authors used exam schools, selective schools where students are bused in, which require high performance on an admission test. Not totally random assignation, but better than nothing.

Problems

31.1: When we wish to show that one effect causes another, and we have complete control of the experimental design, we use what method?

Chapter 32

Sufficient statistics

Question of the Day Suppose $X_1, \ldots, X_5 \overset{\text{iid}}{\sim} X$, where $[X|p] \sim \mathsf{Bern}(p)$. What is the minimum information about $(X_1, \ldots, X_n)$ that I need to estimate p?

In this chapter

- Sufficient statistics

Consider two data sets

- $(1, 1, 0, 1, 0)$ and $(0, 0, 1, 1, 1)$. Does one contain more information about p than the other?
- Since $(X_1, \ldots, X_n)$ are iid X, any fixed permutation σ of data $(\sigma(X_1, \ldots, X_n)$ will also be iid X.
- Under permutations, the statistic S defined as:

$$S = X_1 + X_2 + \cdots + X_n$$

  does not change

Recall

- A model is *parametric* if the distribution that the data $(X_1, \ldots, X_n)$ comes from depends only on parameter $\theta \in \mathbb{R}^d$.
- Ex: $X_i \overset{\text{iid}}{\sim} X$, $[X|p] \sim \mathsf{Bern}(p)$ is a parametric model, where the parameter is p
  - Claim: for this parametric model, the statistic $S = X_1 + \cdots + X_n$ contains all the information in the data about p.

Definition 79
Say $(X_1, \ldots, X_n) \sim f_\theta(x_1, \ldots, x_n)$. A statistic S is **sufficient** for θ if the conditional distribution $[X_1, \ldots, X_n|S]$ does not depend on p.

- Intuition: $S(X_1, \ldots, X_n)$ contains all the information about θ that the entire data set $(X_1, \ldots, X_n)$ did.

QotD

- For $X_1, \ldots, X_n \overset{\text{iid}}{\sim} X$, $[X|p] \sim \mathsf{Bern}(p)$, show that $S = X_1 + \cdots + X_n$ is a sufficient statistic for p
- Let $\Omega = \{0,1\}^n = \{(x_1, \ldots, x_n) : (\forall i)(x_i \in \{0,1\}\}$ be the set of possible outcomes for the data.

Fact 53
For $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Bern}(p)$, $S = X_1 + \cdots + X_n$, let Ω_S be those points $(x_1, \ldots, x_n)$ in $\{0,1\}^n$ such that $x_1 + \cdots + x_n = S$. Then

$$[(X_1, \ldots, X_n)|S] \sim \mathsf{Unif}(\Omega_S).$$

Proof. Let $\vec{X} = (X_1, \ldots, X_n)$, and $\vec{x} = (x_1, \ldots, x_n) \in \Omega_S$. Fix $S \in \{0, 1, \ldots, n\}$. Then

$$\begin{aligned}
\mathbb{P}(\vec{X} = \vec{x}|S = s) &= \frac{\mathbb{P}(\vec{X} = \vec{x}, X_1 + \cdots + X_n = s)}{\mathbb{P}(S = s)} \\
&= \frac{p^s(1-p)^{n-s}\mathbb{1}(x_1 + \cdots + x_n = s)}{\binom{n}{s}p^s(1-p)^{n-s}} \\
&= \mathbb{1}(\vec{x} \in \Omega_s)/\binom{n}{s}.
\end{aligned}$$

□

- Ex: $T = X_1$ is *not* sufficient for p (as long as $n > 1$):

$$\mathbb{P}(\vec{X} = (1, 0, \ldots, 0)|X_1 = 1) = (1-p)^{n-1}$$

  which is a function of p.

- Note that to show that T is not a sufficient statistic just requires a single counterexample, but to show that it is requires that we understand $\mathbb{P}(\vec{X} = \vec{x}|S)$ for all possible $\vec{x}$.

Theorem 8 (Factorization Theorem)
Suppose the data X has density $f_\theta(x)$. Then $S(X)$ is a sufficient statistic if and only if there exists nonnegative functions g and h such that

$$f_\theta(x) = h(x)g_\theta(S(x)).$$

- So if the density of the data at a value x only depends upon:
  - A piece which depends only on the data x but not on the parameter, and...
  - a piece which depends only on the statistic $S(x)$ and the parameter,

  then the effect of the parameter on the data only comes through $S(x)$, so $S(X)$ is the only thing we need to know about the data to understand the parameter.

Fact 54
For $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Bern}(p)$, $S = X_1 + \cdots + X_n$ is a sufficient statistic for p.

Proof. Let $S(x_1, \ldots, x_n) = \sum_i x_i$. Note $f_p(\vec{x}) = p^{S(\vec{x})}(1-p)^{n-S(\vec{(x)})}$. Use the Factorization Theorem with $h(\vec{x}) = 1$ and $g_p(S(\vec{(x)})) = p^{S(\vec{x})}(1-p)^{n-S(\vec{(x)})}$. □

32.1. Minimal Sufficient Statistics

- We'd like the "smallest" sufficient statistic in some sense.
- Note that when you run a number through a function, if it is 1-1 then you don't lose information about the number.
  - Ex: if $y = x^3$, and $y = -27$, then $x = -3$
- But if it is not 1-1 then you can lose information about the number:
  - Ex: if $y = x^2$ and $y = 4$, then either $x = 2$ or $x = -2$.
- So if I can run a sufficient statistic U through a function g and get a new sufficient statistic S, say that $S \preceq U$.

Definition 80
A sufficient statistic S is **minimal** if for U any other sufficient statistic, then $S = g(U)$ for some function g.

- Note that minimal statistics are not unique.
  - If S is a minimal sufficient statistic, so is $k(S)$ for any invertible function k.

How can we be sure if we are at a minimal statistic?

Lemma 4
Consider the ratio of likelihoods between points x and y:

$$R_\theta(x, y) = \frac{f_\theta(y)}{f_\theta(x)}.$$

Suppose S is a statistic with the following property. $R_\theta(x, y)$ does not depend on θ if and only if $S(y) = S(x)$. Then S is a minimal sufficient statistic.

Example, Bernoulli experiments

- For $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Bern}(p)$, $S = X_1 + \cdots + X_n$,

$$\begin{aligned} R_p(\vec{x}, \vec{y}) &= \frac{p^{S(\vec{y})}(1-p)^{n-S(\vec{y})}}{p^{S(\vec{x})}(1-p)^{n-S(\vec{x})}} \\ &= [p/(1-p)]^{S(\vec{y})-S(\vec{x})}. \end{aligned}$$

  If $S(\vec{y}) = S(\vec{x})$, then $R = 0$, but if $S(\vec{y}) \neq S(\vec{x})$, then R depends on p, therefore S is a minimal sufficient statistic.
- Now consider $T(X_1, \ldots, X_n) = (X_1, \ldots, X_n)$. [The identity statistic is always sufficient!]
- Then it is still the case that

$$R_p(\vec{x}, \vec{y}) = [p/(1-p)]^{S(\vec{y})-S(\vec{x})},$$

  but now there exist $T(\vec{y}) \neq T(\vec{x})$ such that $S(\vec{y}) = S(\vec{y})$ so that R_p does not depend on p.

One last way of understanding sufficient statistics...

- Say $f_\theta(x) = h(x)g_\theta(S(x))$.
- Then $\ln(f_\theta(x)) = \ln(h(x)) + \ln(g_\theta(S(x)))$.
- So $\arg\max \ln(f_\theta(x)) = \arg\max \ln(g_\theta(S(x)))$.
- Hence a statistic is sufficient if it is sufficient to calculate the MLE for θ.

Problems

32.1: Suppose that $(X_1, \ldots, X_n)$ given λ are iid $\mathsf{Exp}(\lambda)$. Show that $S(X_1, \ldots, X_n) = X_1 + \cdots + X_n$ is a sufficient statistic for λ.

Chapter 33

Bayesian decision theory

Question of the Day How can we decide between the posterior mean or the posterior median?

In this chapter

- Risk function
- Posterior risk

33.1. Frequentist risk

Definition 81
Given a parameter θ and estimator $\hat{\theta}$, a loss function $\ell(\theta, \hat{\theta})$ is a measure of the discrepancy between the estimate and the true value.

- Common loss functions:
  - $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ (called a *squared loss function*)
  - $\ell(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$
  - $\ell(\theta, \hat{\theta}) = \ln(1 + |\theta - \hat{\theta}|)$

Definition 82
Given a statistical model $[X|\theta]$ and statistic S, the **frequentist risk function** measures the average value of the loss function over the data as a function of θ:

$$R(\theta, S) = \mathbb{E}(\ell(\theta, S(X))|\theta).$$

Notation 8
A subscript on a $\mathbb{P}$ or $\mathbb{E}$ symbol indicates that the random variable is drawn according to the distribution implied by the subscript. So for instance, $\mathbb{P}(Y \in A|\theta) = \mathbb{P}_\theta(Y \in A)$, and $\mathbb{E}[Y|\theta] = \mathbb{E}_\theta[Y]$.

- Now the average distance of an estimator $\hat{\theta}$ from its true value is the *bias* of the estimate.

Definition 83
The **bias** of $\hat{\theta}$ for θ is $\theta - \mathbb{E}[\hat{\theta}|\theta]$.

Fact 55
When $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$, then

$$R(\theta, \hat{\theta}) = (\theta - \mathbb{E}_\theta(\hat{\theta}))^2 + \mathbb{E}[(\hat{\theta} - \mathbb{E}_\theta(\hat{\theta}))^2],$$

that is to say, R is the sum of the square of the bias and the variance of $\hat{\theta}$.

Proof.

$$\begin{aligned} R(\theta, \hat{\theta}) &= \mathbb{E}_\theta(\ell(\theta, \hat{\theta})) \\ &= \mathbb{E}_\theta((\theta - \hat{\theta})^2) \\ &= \mathbb{E}_\theta[((\theta - \mathbb{E}_\theta(\hat{\theta}) + (\mathbb{E}_\theta(\hat{\theta}) - \hat{\theta}))^2] \\ &= \mathbb{E}_\theta[(\theta - \mathbb{E}_\theta(\hat{\theta}))^2] + \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta(\hat{\theta}))^2], \end{aligned}$$

because the cross term is

$$\mathbb{E}_\theta[2(\theta - \mathbb{E}_\theta(\hat{\theta}))(\mathbb{E}_\theta(\hat{\theta}) - \hat{\theta})] = 2(\theta - \mathbb{E}_\theta(\hat{\theta}))\mathbb{E}_\theta[\mathbb{E}_\theta(\hat{\theta}) - \hat{\theta}] = 0.$$

□

- With an unbiased estimator, bias is 0.
- Hence the goal of minimizing the variance in $\hat{\theta}$ using a minimum variance unbiased estimator.
- It could be that a biased estimator has lower overall risk, but can be very difficult to find.

33.2. Bayesian risk

- Bayesians use the fact that they have the posterior distribution of the random variable θ.

Definition 84
The **posterior risk** is $\mathbb{E}[\ell(\theta, \hat{\theta})|X]$.

Compare: risk functions and confidence intervals

	Confidence Interval	Risk Function
Frequentist	$(\forall\theta)(\mathbb{P}(\theta \in [a(X), b(X)]\|\theta) \geq \alpha)$	$\mathbb{E}[\ell(\theta, \hat{\theta})\|\theta]$
Bayesian	$\mathbb{P}(\theta \in [a(X), b(X)]\|X) \geq \alpha$	$\mathbb{E}[\ell(\theta, \hat{\theta})\|X]$

Definition 85
The **Bayes action** is the estimator $\hat{\theta}$ that minimizes the posterior risk.

Fact 56
The choice of $\hat{\theta}$ that minimizes the risk under a squared loss function is to use the posterior mean, that is,

$$\hat{\theta}_2 = \mathbb{E}[\theta|X].$$

Proof. Under squared loss:

$$\mathbb{E}[(\theta - \hat{\theta})^2|X] = \mathbb{E}[\theta^2 - 2\theta\hat{\theta} + \hat{\theta}^2|X].$$

Now $\hat{\theta}$ is a function of the data X, so conditioned on X, you can treat $\hat{\theta}$ as a constant. So

$$\begin{aligned} \mathbb{E}[(\theta - \hat{\theta})^2|X] &= \mathbb{E}[\theta^2|X] - 2\mathbb{E}[\theta|X]\hat{\theta} + \hat{\theta}^2 \\ &= (\mathbb{E}[\theta|X] - \hat{\theta})^2 + \mathbb{E}[\theta^2|X] - \mathbb{E}[\theta|X]^2. \end{aligned}$$

Those last two terms (which happen to add to $\mathbb{V}(\theta|X)$, but that's not important) are independent of $\hat{\theta}$, we have no control over them!

So the minimum of the expression occurs when $\hat{\theta} = \mathbb{E}[\theta|X]$. □

- What's the problem with using squared loss error?
  - It privileges outliers over data close to the mean.

So why did I include the mean here?

Fact 57
The choice of $\hat{\theta}$ that minimizes the risk under an absolute value loss function is to use the posterior median, that is,

$$\hat{\theta}_1 \in \text{median}[\theta|X].$$

- Idea: consider a set of x_i values. What m minimizes $\sum_i |x_i - m|$. If more than half of the x_i are greater than m, then $m + \delta$ moves more than half of the x_i closer by δ, and less than half of the x_i farther by δ, so the total distances between x_i and $m + \delta$ is smaller than that between the x_i and m.

Proof. Let $f(m) = \mathbb{E}[\,|m - \theta||\,X]$. Suppose $m < \text{median}[\theta|X]$. Consider $f(m+h)$. If $\theta \leq m$, then $|m+h-\theta| = |m - \theta| + h$. If $\theta > m + h$, then $|m + h - \theta| = |m - \theta| - h$. If $\theta \in (m, m + h] = g(\theta)$, where $|g(\theta)| \leq h$. So

$$f(m + h) - f(m) = \mathbb{E}[h\mathbb{1}(\theta \leq m) - h\mathbb{1}(\theta > m + h) + g(\theta)\mathbb{1}(\theta \in (m, m + h])|X]$$

Note that for any random variable, $\mathbb{E}[\mathbb{1}(W \in A)] = \mathbb{P}(W \in A)$. So

$$\begin{aligned} f(m + h) - f(m) &= h\mathbb{P}(\theta \leq m) - h\mathbb{P}(\theta > m + h) + \mathbb{E}[g(\theta)\mathbb{1}(\theta \in (m, m + h])|X] \\ &\geq h\mathbb{P}(\theta \leq m) - h\mathbb{P}(\theta > m + h) - h\mathbb{E}[\mathbb{1}(\theta \in (m, m + h])|X] \\ &= h\left[\mathbb{P}(\theta \leq m) - \mathbb{P}(\theta > m)\right]. \end{aligned}$$

Therefore, if $m \geq \text{median}(\theta|X)$, then $f(m + h) \geq f(m)$, and f is an increasing function at m.

A similar argument gives that if $m \leq \text{median}(\theta|X)$, then $f(m + h) \leq f(m)$, then f is decreasing at m.

Therefore, if m minimizes $\mathbb{E}[|m - \theta||X]$, then $m \in \text{median}(\theta|X)$.

□

Some thoughts on decision theory

- When it is possible to use an absolute loss function, do it.
  - Mitigates the effects of outliers on your estimate.
- There are even more complicated loss functions

$$\ell(\theta, \hat{\theta}) = \ln(1 + |\theta - \hat{\theta}|).$$

Recall

$$\ln(1 + \delta) = \delta - \delta^2/2 + \delta^3/3 - \cdots,$$

so for δ small loss is about δ, but it discounts outliers even more than median does.

  - Appropriate for very heavy tailed distributions.

– Could be difficult to calculate optimal value.

Problems

33.1: Consider the function

$$f(s) = |s - 13| + |s - 14| + |s - 17|.$$

Find $\min f(s)$ and $\arg\min f(s)$ for $s \in \mathbb{R}$.

Part III

Statistics Laboratory Experiments

Chapter 34

Stats Lab: An Introduction to R

Instructions

This lab will introduce you to using R for statistical computations. If you have used R before, great! However, this lab does not assume that you have used statistical software before. You have the full period to complete this lab, but if you finish early you are welcome to work on homework problems. For any and all of these questions, if you run into difficulty, please ask me for help, that's what I'm here for!

Begin by starting up R or RStudio. In this lab we will learn how to assign variables. We also learn about the central variable type in R, the data frame.

- The assignment operator in R is a less than sign followed by a hyphen. So for instance, typing `a <- 3` will assign the value 3 to the variable `a`. After typing this command, try just typing `a`. What does R return?

- The [1] 3 means that `a` is a variable that holds a single number. The [1] means that row position starts with 1. If I wanted to assign the numbers from 100 to 200 to `a`, type `a <- 100:200`. Now type `a` again. The result is a *vector* or *array* of numerical values. What is the 19th number in the array?

- Notice that you can bring back commands that you typed in the R console previously by using the up arrow on the keyboard. Try bringing back the command `a <- 100:200` and modifying it to be `a <- 100:250`.

- Suppose that we have a set of ages as data, and they are 24, 27, 43, and 12. This can be put into a vector using the `c` command, which stands for concatenate (or combine). Try typing `ages <- c(24,27,43,12)`. Once you have the data in place, you can use standard commands in R to estimate the sample mean and variance. Use `mean(ages)` and `sd(ages)` to estimate the mean and standard deviation of this data set. [Note: you can separate commands in R using a semicolon ;. So the command `mean(ages);sd(ages)` would do both calculations with one line of input.]

- In R, you can get help for a command by putting a `?` before the command name. Try `?c` to get help on the `c` command. What is the first sentence of the Description for `c`?

- Suppose now that the subjects have names, Alice, Bob, Charlie, and Dorothy. We can create *string* variables by using quotation marks. Try typing

```
names <- c("Alice","Bob","Charlie","Dorothy")
```

What is the result of typing `names[3]`?

- So far there are two types of data, there is the names, and there are the ages. In statistics, types of data are called *factors*. Let's combine the two into a *data frame* which is used for storing different factors together. Type `df <- data.frame(names,ages)` what does typing `df` return?

- You can see that our data frame has automatically labeled the columns with names of the original variables, and the rows are just numbered 1 through 4. To access a particular row of a data frame, use the `$` symbol. What does `df$ages` return?

- Typing something like `mean(df$ages)` then would return the mean of the ages. What does typing `mean(df$names)` return?

- Of course that didn't work because the names are strings, they are not numerical values. R has some data frames built in. For instance, if you type `mtcars` you get a data set of car specifications taken from a 1974 Motor Trend article. Estimate the average mpg (miles per gallon) from this data set.

- You can use the `[data frame name]$[factor name]` construction to draw data from the table, or you can directly draw from rows and columns. What does `mtcars[3,4]` return?

- In statistical notation, the comma (,) is used as a wildcard. So `mtcars[,2]` means any row and the second column. What command to R would return the second row and any column?

- Now usually, your data is often in a computer file. Download the `beer.txt` file from the course website to the Documents folder on your computer. Open this file up with a text editor. How many factors are there in this data set?

- Before we can load `beer.txt` into R, we have to tell R where to look for it. The Document folder on my computer is C:/Users/Mark/Documents So I used

  ```
  setwd("C:/Users/Mark/Documents")
  ```

  in R to tell it the directory where I wished to work. (Here `setwd` stands for "Set working directory".)

  Okay, now that I'm in the right directory, type `beer <- read.delim("beer.txt",header=TRUE)` to load the file into the data frame called `beer`. The `header=TRUE` part tells R that the first line of the file isn't actually data, it's the names of the factors. Estimate the average percent alcohol in beer from the data provided.

- Use `?read.delim` to find a command that is related to the `read.delim` command, and write it below

- Of course, R can do much more than just estimate the mean and standard deviation. Estimate the median of Carbohydrates in beer for this data set using R.

- You can get quantiles as well with the `quantile` command. Or you can get everything at once with the `summary` command. What is the result of typing `summary(beer$Carbohydrates)`?

- While point estimates are useful, visualizations of data can have a greater impact Visualization of data is also very important. Create a histogram by using `hist(beer$Carbohydrates)` and sketch the results.

- How many beers fall into the leftmost bar?

- You can add a tick mark for each actual value by using `rug(beer$Carbohydrates)`. This creates what is known as a rug plot. These marks show the exact location of the data values within the histogram bars. How many distinct Carbohydrate values fall in the leftmost bar? (That is, how many tick marks are under the leftmost bar.)

- Can you think of why there are fewer tick marks under the leftmost bar than the height of the bar?

- One final note: you can save your plot to a file using `dev.copy(pdf,"mygraph.pdf");dev.off()`. This version would save it as a .pdf file, other formats are of course available as well. Try saving your plot and opening it with Acrobat to make sure that you saved it correctly.

Chapter 35

Stats Lab: Working with different distributions

- R has a lot of built in commands to work with distributions. Distributions that R has include: Beta (beta), Binomial (binom), Cauchy (cauchy), Chisquare (chisq), Exponential (exp), F (f), Gamma (gamma), Geometric (geom), Hypergeometric (hyper), Logistic (logis), Lognormal (lnorm), Negative Binomial (nbinom), Normal (norm), Poisson (pois), Student t (t), Uniform (unif), Tukey (tukey), Weibull (weib), and Wilcoxon (wilcox). For each distribution, there are 4 commands associated with it, each formed by adding a letter to the name of the distribution. For example, for the Binomial distribution the four commands are `dbinom`, `pbinom`, `qbinom`, and `rbinom`. The first, `dbinom` evaluates the density of the distribution. For instance, the command `dbinom(x=3,size=10,prob=0.4)` will give $\mathbb{P}(X = 3)$ where $X \sim \mathsf{Bin}(10, 0.4)$.

- Use R to find $\mathbb{P}(A = 4)$ where $A \sim \mathsf{Bin}(20, 0.5)$.

- We can use this to plot the density of values at various values. First let's get our x values. If you use `x <- seq(0,1,length=101)`, what are the first few values of `x`?

- Now let's get the values of the density. Try a $\mathsf{Beta}(3, 2)$ density by using `hx <- dbeta(x,3,2)`. Now let's plot it with `plot(x,hx,type="l")`. [Note that the symbol inside the quotation marks is the letter l, not the number 1. The l here stands for line, and means that there is a line connecting all the points of the plot.

- Let's find the maximum value of `hx` using `max(hx)`. What is it?

- Unfortunately, if $f_X(s) \sim \mathsf{Beta}(3, 2)$ was our posterior density in a Bayesian computation, we don't care about $\max f_X(s)$, we want $\arg\max f_X(s)$. Try using `which.max(hx)` to find out the index of the maximum value. But that still isn't $\arg\max f_X(s)$. To get the true answer, use `x[which.max(hx)]` to find the x value that maximizes the density. What is it?

- The true value of the maximum is $2/3$, and so it is close to the true answer, but since the x values change by 0.01, that is as close as it can get. Now let's numerically estimate the mean of a $\beta(3, 2)$ random variable, by estimating $\int_0^1 x f_X(s)\, ds$ using the trapezoidal rule:

```
(x[2]-x[1])*(sum(x*hx)-0.5*x[1]*hx[1]-0.5*x[length(x)]*hx[length(hx)])
```

Note that a construction like `x[length(x)]` will pick out the last element of the vector `x`. What do you get from using the trapezoidal rule?

- The true answer is $3/(3+2) = 0.6$, so again it is close, but not quite correct.
- Now let's suppose that $p \sim \mathsf{Beta}(3,2)$ is the prior distribution for a Bayesian analysis of the outcome of a drug trial. The statistical model of the observed data is $[X|p] \sim \mathsf{Bin}(n,p)$. Suppose that $n = 20$, and $X = 11$ successes were observed in the experiements. Then we know from Bayes' Rule that the posterior distribution will be beta distributed with parameters $3+11 = 14$ and $2+9 = 11$. You can add to an existing plot using the `lines` command in R. So use `post <- dbeta(x,14,11)` to get the density of the posterior, and then `lines(x,post,col=''red'')` to add it to the plot. Sketch the result.

- Okay, so the idea was good, but the new plot is too tall for the existing one! Let's find out how tall it should have been by using `m <- max(post)`. What is the value of `m`?

- Now we will go back to our `plot` command using the uparrow and add limits to the y values with `plot(x,hx,type="l",ylim=c(0,4.5))`. Note that as soon as you use the `plot` command it erases all extra lines that you added. So add back the plot of the $\mathsf{Beta}(14,11)$. Now sketch the result.

- Note that the $\mathsf{Beta}(14,11)$ is more concentrated than the original $\mathsf{Beta}(3,2)$. As you take more and more data, the results will be closer and closer to the true parameter.
- The `p(distribution name)` commands find the cdf of a random variable with that distribution. So for instance, `pbinom(q,n,p)` command will find $\mathbb{P}(X \leq q)$ where $X \sim \mathsf{Bin}(n,p)$. `pbeta` is similar. Using this command, find the probability that $p \sim \mathsf{Beta}(14,11)$ is at most 0.7.

- Find the probability that $p \sim \mathsf{Beta}(14,11)$ is in $[0.4, 0.7]$.

- The `q(distribution name)` commands find the inverse of the cdf. For example, using the command `qbinom(0.3,10,0.4)` will tell R to find the value of a such that $\mathbb{P}(X \leq a) = 0.3$. This can be useful in finding the median of a distribution. Recall that a value m is a *median* for a random variable X if $\mathbb{P}(X \leq m) \geq 1/2$ and $\mathbb{P}(X \geq m) \geq 1/2$.

  Use `qbeta` to find the median of $p \sim \mathsf{Beta}(14,11)$.

- Now use the `pbeta` command to verify that the value that you found is actually the median value.
- Again for $p \sim \mathsf{Beta}(14, 11)$, find a and b such that $\mathbb{P}(p \leq a) = 0.025$ and $\mathbb{P}(p \geq b) = 0.025$ so that $\mathbb{P}(p \in [a, b]) = 95\%$. [In Bayesian statistics, this is called a credible interval.]

- The last set of commands, `r(distribution name)`, draws iid random draws from the distribution requested. For instance, try `s <- rbeta(10000,14,11)`. Look at the first couple values using `head(s)`. Note that they all lie between 0 and 1, like beta values should. Use a histogram to look at the data with `hist(s)`. We can approximate the density of the results using `plot(density(s))`. Sketch the result.

- Overlay the actual density plot of the $\mathsf{Beta}(14, 11)$ using the `lines` command as before. Is it a good fit?

- Of course R has built in commands to do maximum likelihood estimation of parameters. In order to do that, we have to first load in a new package called MASS. Use the command `library(MASS)` to do that.
- We need to know the parameters used by the `dbeta` function in R. By typing `?dbeta` you can see that these parameters are labeled "shape1" and "shape2". Now try the command

  ```
  fitdistr(s,"beta",start=list(shape1=1,shape2=1)
  ```

  to run a numerical optimization to get the MLE for the shape parameters for this random data. What are the results?

- Of course, what if we didn't know ahead of time that our data was beta? We might have used a different distribution. Try

  ```
  fitdistr(s,"normal")
  ```

  to obtain the MLE treating this as normal data. What are the best fit parameters?

- Reset your plot of your data values with `plot(density(s))`. Now plot the density of a normal distribution with your MLE values found above over this plot in the same way you did earlier with the beta distribution.

Chapter 36

Stats Lab: Confidence Intervals

Instructions

If you have time in the period, complete both the main and extended portions of the lab. If you run out of time, you do not have to complete the extended lab.

Goals

- Learn how to build confidence intervals.
- Introduction to scripts in R.

Main Lab

- First, try typing `# This is a comment` into R. This returns nothing: R ignores any command that begins with the number sign `#`. This is useful to know when you learn about scripts and functions later and want to make comments on what you are doing.
- Now let's do some statistics. We will begin by building a 95% z-value interval for some ficticious data. Try typing

```
x <- c(10,15,7,22,14,7,8)
a <- mean(x)
s <- sd(x)
n <- length(x)
```

  into R. Here a is your sample mean, s is your sample standard deviation, and n is the number of data points in the sample. Then type

```
error <- qnorm(0.975)*s/sqrt(n)
left <- a - error
right <- a + error
print(left)
print(right)
```

  (Note the `qnorm(0.975)` command finds the location where 97.5% of the probability is to the left for a standard normal random variable.) What is the 95% z-value confidence interval for the data?

- If you had to type all eight lines after the assignment to `x` each time you wanted a confidence interval, that would get tiresome very quickly. So instead we are going to create what is called a *script*. Click

on the menu option "File" move the cursor down to new file, and then right to "R Script". That brings up a new tab in the editor, which by default is above the console in R Studio. Now type the same eight commands

```
a <- mean(x)
s <- sd(x)
n <- length(x)
error <- qnorm(0.975)*s/sqrt(n)
left <- a - error
right <- a + error
print(left)
print(right)
```

into the editor. Note that you can go back and forth between the editor and console by clicking on the appropriate window.

When you are ready, click the checkmark box Source on Save in the script window, and then the little floppy disk "Save" icon to the left of the checkbox. Save your script as ci95.R.

Note that in the console, the command `source('~/ci95z.R')` has executed. The result should be the same as before. Now try the command `x <- c(14,22,13,8)`, and then execute your script using the same source command. What is the 95% confidence interval to 4 sig figs for the new data?

- In building the z-value CI, we assumed that $(\mu - \hat{\mu})/(\hat{\sigma}/\sqrt{n}))$ had a normal distribution. Since $\hat{\sigma}$ does not quite equal σ, this is not quite right. Instead, $(\mu - \hat{\mu})/(\hat{\sigma}/\sqrt{n})$ has what is called a t distribution with $n-1$ degrees of freedom. To visualize what a t distribution looks like, try

  ```
  x <- seq(-3,3,by=0.1)
  plot(x,dt(x,df=3),type="l")
  ```

  To compare to a normal distribution, type

  ```
  lines(x,dnorm(x),col="red")
  ```

  Sketch the result.

- Try doing the same plot and sketch the result for a t distribution with 19 degrees of freedom. Do you think there will be much difference between a z and a t CI with 19 degrees of freedom?

- Modify your script `ci95zvalue.R` to `ci95t.R` by changing `qnorm(0.975)` to `qt(0.975,df=n-1)` and resaving with the new filename. Use your new script to find the 95% t CI for the data $(14, 22, 13, 8)$.

- Note that the t CI is slightly wider than that z CI. That is because the t CI is acknowledging that our estimate $\hat{\sigma}^2$ is not exactly σ^2, and so that extra uncertainty widens the interval slightly.
- Now let's try this on a real data set. `CO2` is a data set that built into R. Type `head(CO2)` to get a look at the first few data points. It records the uptake of CO2 by plants in various locations under various conditions. Let's pick out the tests where the Treatment was chilled and the Type was Quebec by using:

  ```
  x <- CO2$uptake[(CO2$Treatment == "chilled" & CO2$Type=="Quebec")]
  ```

  Find a 95% t CI for the Quebec chilled plants.

- Unsurprisingly, R does have a built in command for finding t confidence intervals. Try `t.test` and verify that the the 95% t CI for the Quebec chilled plants is what you found above.

- Now look at the help for `t.test` and change your command so that it delivers a 99% confidence interval. Is your confidence interval narrower or wider than the 95% confidence interval?

- Now find a 95% t CI for the Mississippi chilled plants.

- Assuming the statistical model is true, there is at most a 5% chance that the Quebec CI does not cover the true answer, and at most a 5% chance that the Mississippi CI does not cover the true answer. Therefore there is at least a 90% chance that both CI cover the true answers for each. Does the Quebec CI and Missippi CI overlap?

- Since the intervals do not overlap and there was (before we took the data) at least a 90% chance that the intervals would both contain their true parameter, we can say that the hypothesis that the two averages should be rejected at the 10% level. Now suppose that we want a stronger statement. Instead of rejecting at the 10% level, we want to reject at the 5% level. What level should the two CI for the Quebec and Mississippi data have been to get this level for the hypothesis test?

- Should we reject the hypothesis that the Quebec and Mississippi average CO2 uptake are the same at the 95% level?

Extended lab

In this optional part of the lab we'll explore the notion that a confidence interval is itself a random variable.

- Suppose $X_1, \ldots, X_{10} \sim \mathsf{N}(3, 5^2)$. Then for each draw of the data, we will get a different confidence interval! To see this in action, create a new script `montecarloci.R` that contains the following commands

```
x <- rnorm(n=10,mean=3,sd=5)
source("~/ci95t.R")
```

- Try sourcing the `montecarloci.R` script several times. Each time you source it you will see that you get a different confidence interval. Sometimes this confidence interval will contain the true mean 3, and sometimes it will not.
- To keep track of how often this occurs, we need to learn some new commands. First, by using `rep(0,20)` we can create a vector of length 20 where each entry is 0. Give it a try in R.
- Next, to repeat the same set of commands over and over again, a `for` loop can be used. For instance, try the following command:

  ```
  s <- 0; for (i in 1:10) s <- s + i;print(s)
  ```

  That will sum the integers from 1 to 10 and print the result. Verify that the comand returns

  $$\sum_{i=1}^{10} i = 55.$$

  Modify and use to find $\sum_{i=1}^{100} i^2$.

- Now we are ready for our next script, `montecarloci2.R`. [Note that we put several commands inside brackets { and } in order to execute several commands for each value of `i` from 1 to `trials`.

```
trials <- 10000
results <- rep(0,trials)
for (i in 1:trials) {
  x <- rnorm(n=10,mean=3,sd=5)
  a <- mean(x)
  s <- sd(x)
  n <- length(x)
  error <- qt(0.975,df=n-1)*s/sqrt(n)
  left <- a - error
  right <- a + error
  results[i] <- (left <=3)*(right>=3)
}
print(mean(results))
```

  What is your result from running this script?

- Repeat with a 99% confidence interval.

Chapter 37

Stats Lab: Nonparametric confidence intervals

Instructions

So far we have seen how to build nonparametric confidence intervals for the median of a distribution. Now you will learn a way to build nonparametric confidence intervals for the mean (or any other statistic of the data). The method is known as the *bootstrap*, and it was developed in the 1970's to take advantage of the rise in computing power. It is an example of a *Monte Carlo method*, where randomness is intentionally introduced in order to understand the behavior of a distribution.

The nice thing about this method is that it can turn any point estimate into a confidence interval, without the need to build a pivot!

If you have time in the period, complete both the main and extended portions of the lab. If you run out of time, you do not have to complete the extended lab.

Main Lab

- The more complicated the model, the more difficult it becomes to build reasonable confidence intervals for the results. A simple (nonparametric) way to solve this problem is to use the idea of the *bootstrap*. In the bootstrap method, the analysis is performed on random subsets of the data in order to see what variation can be expected in the answer based on the random choices made. In order to utilize the bootstrap, we will have to be able to draw random subsets of a vector. Try the following

```
x <- c('a','b','c','d')
sample(x,10,replace=TRUE)
```

  and report your results.

- You can also sample without replacement using `sample(x,2,replace=FALSE)`.

- What happens if you replace 2 in the above command with 10?

Heavy tailed distributions

- Now let's look at the difference between light tailed and heavy tailed distributions. First let's compare normals and cauchys:

```
z <- rnorm(1000)
x <- rcauchy(1000)
```

Find the minimum and maximum of the `z` values and the `x` values

```
min(z)
max(z)
min(x)
max(x)
```

- Those very large and very small values make it difficult to look at histograms as well. Try sketching the following two histograms:

```
hist(z)
hist(x)
```

- The sample average of `z` will be very close to 0 (try `mean(z)` to see,) while that of `x` might be very close to 0 or might be very far away. Both distributions have a mean of 0. Report what you find with

```
median(x)
median(z)
```

- These nonparametric estimates converge no matter how heavy tailed the distribution. Moreover, we can quickly generate nonparametric confidence intervals for the median. Let $N = \#\{i : X_i < \text{median}(X)\}$. Then $N \sim \mathsf{Bin}(n, 1/2)$. For $n = 1000$, $\mathbb{P}(N \leq 468) = 0.0231456$ and $\mathbb{P}(N \leq 468) = 0.0231456$). So $[X_{(469)}, X_{(532)}]$ is a 95% nonparametric confidence interval. Try

```
x <- sort(x)
x[469]
x[532]
```

and report the 95% confidence interval.

The bootstrap: nonparametric confidence intervals for the mean

- Now let's bring in our data for the week by putting the file `flavors_of_cacao.csv` into the variable `cacao` using the command

```
cacao <- read.csv("flavors_of_cacao.csv",header=TRUE)
```

Use `names(cacao)` to find the names of the factors. What is the name of the last factor?

- Use `nrow` to find out how many data points there are. How many chocolate bars are rated in this data set?

- The ratings for the chocolate bars are under the factor `Rating`. Using `cacao$Rating` returns the ratings for the bars. You can use `head(cacao$Rating)` to see that the ratings use a quarter point scale in their rankings. What is the mean rating? The minimum rating? The maximum rating?

- Now I want to try to get an idea of the variation in the mean rating just from the data alone, that is, without assuming that the ratings follow a particular pattern. Start by randomly sampling 1795 ratings from the data set with replacement.

```
samp <- sample(cacao$Rating,1795,replace=TRUE)]
```

  Find the mean of your random subsample.

- Most of the time, the mean of `samp` will be different from that of the original ratings. Now let's try repeating this experiment over and over again. The way to do this in R is to use the `replicate` command. The first argument gives the number of times that you wish to do the replication, then the second argument is the command that you wish to replicate. Try this:

```
results <- replicate(1000,mean(sample(cacao$Rating,1795,replace=TRUE)))
```

  What are the first five entries of the variable `results`?

- Now sketch a plot of the results with

```
plot(density(results), lwd=3, col="steelblue") \\
abline(v=mean(cacao$Rating), lwd=3, col='gold')
```

- Note how much variation there is in the estimate of the mean! And this was just from random samples from the same data! Now let's build a 95% confidence interval from the bootstrap. First we will sort the random results that we found, then we take the order statistics $\hat{\mu}_{(26)}$ and $\hat{\mu}_{(975)}$ that contain 95% of the results:

```
results <- sort(results)
results[26]
results[975]
```

  What is the 95% bootstrap confidence interval for the mean of the rating data?

- Now find the 95% t-value confidence interval for the mean of the ratings data using `t.test`. Is this close to your bootstrap confidence interval?

Extended Lab

- The command we used for the bootstrap was

```
results <- replicate(1000,mean(sample(cacao$Rating,1795,replace=TRUE)))
```

  Modify the command to get a bootstrap confidence interval for the standard deviation of the ratings data.

- So far we've found nonparametric bootstrap confidence intervals for the mean and standard deviation. The nice fact about this type of interval estimate is that it can be found for any statistic, without the need for a pivot. For instance, the mean absolute deviation (MAD) of a random variable is $\mathbb{E}[|X - \bar{X}|]$. MAD has the advantage over the standard deviation is that any random variable with finite mean will also have a finite MAD. To estimate the value of the MAD for data in vector `x`, we can use

```
mean(abs(x - mean(x)))
```

  in R. Use this to get a bootstrap confidence interval for the MAD of the ratings.

Chapter 38

Stats Lab: Regression and models

Instructions

If you have time in the period, complete both the main and extended portions of the lab. If you run out of time, you do not have to complete the extended lab.

Main Lab

- Recall that a linear model is of the form

$$Y = X\beta + \epsilon,$$

where Y is an $n \times 1$ matrix (aka a length n column vector), X is an $n \times k$ matrix (n pieces of data, and k predictor variables), β is a $k \times 1$ matrix of coefficients, and ϵ is an $n \times 1$ matrix (length n vector) of random effects. You won't be surprised to learn that R has built in commands to estimate β.

To illustrate these commands, we will start with the daily returns of IBM and the S&P 500 index to use one value to predict the other. This data is stored in a text file where the data is separated by tabs.

Load this data into a data frame in R

```
spxibm <- read.table("spx_ibm2006.txt",
  header=TRUE, sep='\t', row.names=1))
```

Now, consider the simple model

$$\text{ibmret} \sim 1 + \text{spxret}.$$

To put this model into R, try the command

```
lm(ibm ~ spx,data=spxibm)
```

Two notes:

1: We don't have to put in the constant term, R just assumes that you always want that.

2: Here `lm` stands for linear model. It is the letter ell and the letter em, not the number 1 followed by an m!

What are the results?

- This gives a coefficient for the (Intercept) which is the constant, and a coefficient associated with the value of the S&P 500 in columb spx. In mathematical finance, the coefficient of the stock price versus a stock index like the S&P 500 is considered the part of the stock's returns that are explained by the underlying economy rather than the particular properties of the stock. This is called the *beta* of the stock. What is the IBM beta versus the S&P 500 to 4 sig figs?

- The `lm` command returns a new type of object called a linear model that you can store in a variable name for further analysis. Commands that take linear models as arguments can be used to obtain the fits and residuals. Try the following:

```
ibm.lm <- lm(ibm ~ spx,data=ibmspx)
coef(ibm.lm)
```

  This just gives the coefficients of the fitted model. That is, the prediction is of the form:

$$y_i\beta_0 + \beta_1 x_i$$

  where the (Intercept) coefficient is β_0 and the `spx` coefficient is β_1.

  Now let's get $X\beta$ and ϵ and put these into variables:

```
ibm.fit <- fitted(ibm.lm)
ibm.resid <- resid(ibm.lm)
```

  This last command gives the values of the ϵ, that is the difference between the true response Y and the predicted response $X\beta$. Using `head(ibm.resid)`, what was the residual from 2006-01-05 for the least squares linear model?

- Because this data is low dimensional, we can plot it to get an idea of what it looks like. Start by making a scatterplot of the data.

```
plot(spxibm$spx,spxibm$ibm,xlab="S&P 500",ylab="IBM")
```

  Its kind of a mess! Now lets put the least squares regression line onto this plot using the `abline` command:

```
abline(ibm.lm,lwd=3,col='blue')
```

  Note that this line would be very difficult to draw by hand without fitting the least squares solution.

- Of course, even if the data fit the linear model perfectly, there would still be variance in the choice of the β coefficients. To understand how big that variance is, we can use bootstrapping.

  First we will create a simple function in R. Start by opening up the text editor using the "File" menu command and then "New Script". Now try entering the following commands.

```
test <- function(x) {
  return(x^2)
}
```

  Save your script as `test.R`. Source your script with `source("test.R")`. Type `test`, what does R return with?

- When you just type the name of a function, you get the lines that define the function. To actually use the function to calculate something, you follow the function name by one or more arguments in parenthesis. Try `test(4)` and `test(16)`. What does this return?

- Now try `test()`. What does R say?

- Here R was expecting an input for the function, and not finding one, it returned an error. It helps to set a default value for inputs, so that the function has a base value to use even if the use does not give one. That way we can have many parameters that the user might or might not set. Change the first line of `test.R` to read

  ```
  test <- function(x = 2) {
  ```

  and then use `source("test.R")` to read in the function again, and then try `test()`. What does R return now?

- Of course, you can override the default by giving an argument to the function. If you call `test(4)`, what does R return?

- Now let's add more lines to `test.R` to illustrate a function with more than one argument.

  ```
  test.multiple <- function(x = 0,y = 0) {
    return(2*x + 3*y)
  }
  ```

  Source the file again, and then try the commands `test.multiple(5,4)`, `test.multiple(x = 5,y = 4)`, and `test.multiple(y = 4,x = 5)` and record the results.

- The moral of the last task is that you can always use the variable names in a function call to make sure that you are correctly assigning the proper name to the right argument. That way you don't have to remember the order the function needs its variables in.

  Okay, now let's get back to bootstraping the beta for the IBM stock returns.

  What we want to do for the bootstrap is to write a function that given a response and explantory variable, gives us a beta for a randomly chosen set of (spxret,ibmret) ordered pairs. Add the following lines to your `test.R` file:

  ```
  beta.random <- function(data) {
    n <- nrow(data)
    indices <- sample(n,n,replace=TRUE)
    sampled.data <- data[indices,]
    return(coef(lm(ibm ~ spx,data = sampled.data))[2])
  }
  ```

  Now source `test.R` again, and try `beta.random(spxibm)` three times. What do you get?

- Now we are ready to do the bootstrap. Use

  ```
  beta.boot <- replicate(1000,beta.random(spxibm))
  ```

  to get 1000 replications of the bootstrap experiment. Use

  ```
  plot(density(beta.boot), lwd=3, col="steelblue")
  ```

to get an idea of the different values of β_1 found, and then

```
abline(v=coef(ibm.lm)[2],lwd=3,col='gold')
```

to add a line indicating the β_1 value for the data. Sketch the result.

- Sort the bootstrap observations and look at the 26th and 975th entries to obtain a 95% confidence interval for the beta for the IBM stock price.

- A beta of 0 indicates that the stock price is unrelated to the S&P 500 returns. Based upon your 95% confidence interval, do you think it is likely that the IBM daily returns and the S&P 500 returns are unrelated?

Extended Lab

- Let's learn a little bit more about what a linear model object is like in R. First try

```
class(ibm.lm)
```

What is the class of a linear model object in R?

- Now try

```
names(ibm.lm)
```

to see the variety of different parts of the linear model. Try

```
ibm.lm$rank
```

What is the rank of the matrix X? Does it have full column rank?

- What command would you use to get the beta coefficients without using the `coef` function?

- In deriving the MLE for least squares we assumed that the residuals were normally distributed random variables that were independent with the same variance. Such variables are called *homoscedastic* (here homo is the Greek prefix for same, so this literaly means "same variance".) Let's take a look at the residuals with

```
plot(imb.lm$residuals)
```

Okay, they certainly aren't as heavy tailed as a Cauchy, but are they normal? One way to test is with a *QQ-plot* which plots the empirical cumulative distribution function of the data against the cdf of a standard normal. If the data comes from the distribution, it should lie on a straight line. Try it out for the data residuals with:

```
qqnorm(ibm.lm$residuals)
qqline(ibm.lm$residuals,lwd=3,col="gold")
```

Interpreting QQ-plots is a bit of an art, but in this case you can say that the normal model is reasonable, at least from about -1.5 to 1.5.

Chapter 39

Stats Lab: p-values

Instructions

If you have time in the period, complete both the main and extended portions of the lab. If you run out of time, you do not have to complete the extended lab.

For some statistical models, it is very difficult to get p-values exactly. Monte Carlo methods can be used to obtain an exact p-value for such a data set. The test statistic for this data set will be

$$T = \sum_i \min_{j \neq i} \text{dist}(v_i, v_j),$$

so T is the sum over all points of the distance of the point to the next closest point.

Main Lab

- Begin by loading the file `spanish_cities_plain5.csv` into the variable `st`. This is a classic data set of the locations of towns in an area of Spain taking by the United States Geologic Service right after World War II ended.

  Use `nrow(st)` to answer: How many points are there in this data set?

- The Euclidean distance between two vectors `v1` and `v2` can be found by using `sqrt(sum((v1-v2)^2))`. Recall that you can get row `i` of `st` using `st[i,]`. Find the Euclidean distance between the first and second points in `st`.

- Now let's build a function that calculates T for any set of data. To do this we'll need to use a for loop. In programming, a for loop is a set of commands that are repeated over and over again for different values of a variable. For instance, in R if I use `for (i in 1:5)`, the command following the for command will be executed five times, once using `i <-1`, once with `i <- 2`, and so on up to `i <- 5`. Find $\sum_{i=1}^{100} i$ by using

```
s <0 0; for (i in 1:100) s \assignop{} s + i;
 cat(s,'\textbackslash{}n')}
```

- Now let's write code to find T. Open up a new script, and enter:

```
spatialt <- function(points) {
  n <- nrow(points)
  d <- rep(0,n)
  s <- rep(0,n)
  for (i in 1:n) {
```

```
    v <- points[i,]  # the row to compare to every other row
    # now find the distance from v to every row
    d <- sapply(1:n,function(j) sqrt(sum((v-points[j,])^2)))
    # now find the second smallest entry of d
    s[i] <- sort(d,FALSE)[2]
  }
  return(sum(s))
}
```

What is the value of T for `st`?

- Now let's see how long that took to calculate. Try the command `system.time(t <- spatialt(st))` to determine how long it took to run the command once. What is the elasped time?

- One of the reasons that the calculation is so slow is because R has to first convert the data frame to numerical values. A matrix of values can be stored much more efficiently, and calculations done much more efficiently as well. A data frame can be converted to a matrix with the `data.matrix` command. Try the following.

```
stm <- data.matrix(st)
system.time(spatialt(stm))
```

  How much elapsed time did this take?

- Now the question becomes: is the value T(`st`) that you found earlier large? small? medium? In order to understand that, suppose that our statistical model is that the data X is all independently drawn from the unit square. Then our goal is to understand the distribution of the test statistic $T(X)$.

  To draw 68 points from the unit square, we can draw $(2)(68) = 136$ uniform random variables from $[0, 1]$, and form them into a matrix with two columns. Commands in R of the form `r<distribution name>` generate one or more random draws from that distribution. So `rnorm(1)` gives a random standard normal, `rexp` gives a random exponential, and so on. Try

```
matrix(runif(136),ncol=2)
```

  to see this in action. Now try

```
spatialt(matrix(runif(136),ncol=2))
```

  to get a random draw from T. What value of T did you get?

- Now we are going to replicate the result several times. Try

```
replicate(n=3,spatialt(matrix(runif(136),ncol=2)))
```

  and report the result.

- Use `t <- replicate(n=100,spatialt(matrix(runif(136),ncol=2)))` to get a set of 100 values. Using `system.time` to measure, how much elasped time did this take?

- Given this timing, about how long do you think it would take to get 1000 replications?

- Now let's try to make an estimate of $\mathbb{P}(T(X) \geq T(\texttt{st})$. Start with `tst <- spatialt(stm)`. Let's generate a 1000 values from the $T(X)$ distribution.

```
tvalues <- replicate(n=1000,spatialt(matrix(runif(136),ncol=2)))
```

  Take a look at this density, and also plot the value of $T(\texttt{st})$ with

```
plot(density(tvalues))
abline(v=tst,lwd=3,col="gold")
```

- You can see that very few of the T values generated are to the right of the data value. Estimate this probability by turning the values into Bernoulli random variables with

```
results <- tvalues >= tst
```

  What are `mean(results)` and `sd(results)`?

- Note that `sd(results)` here is giving an estimate for a single data point R_i from `results`. But we are interested in the standard deviation of the mean, that is

$$\mathrm{SD}\left(\frac{R_1 + \cdots + R_{1000}}{1000}\right) = \frac{1}{\sqrt{1000}} \cdot \mathrm{SD}(R_1).$$

  We saw this as well with the confidence intervals for z-score as well. Remember that for a γ-level confidence interval, the endpoints are

$$\left[\hat{\mu} - \frac{\hat{\sigma}}{\sqrt{n}} \mathrm{cdf}^{-1}_{N(0,1)}(1/2 - \gamma/2), \hat{\mu} - \frac{\hat{\sigma}}{\sqrt{n}} \mathrm{cdf}^{-1}_{N(0,1)}(1/2 + \gamma/2)\right],$$

  so the width of the interval is proportional to $\hat{\sigma}/\sqrt{n}$.

  The consequence is that to get the estimate of the standard deviation in the results variable, use

```
sd(results)/length(results)}
```

- Do you think that you have enough evidence to reject the null hypothesis that the locations of the cities in the plain are uniformly drawn from the unit square?

- In lecture, we said that the p-values are uniformly distributed over $[0, 1]$. To see why is true, suppose we pick one of our `tvalues` uniformly at random. In R this can be done with the `sample(tvalues,1)` command. Then we ask what is our estimate of the probability that the other data values are less than our sample. In other words, a single draw from the distribution of the p-value can be done with `mean(sample(tvalues,1) >= tvalues)`. Try executing this command three times and report your results.

- Now let's try this 10, 000 times:

```
pvalues <- replicate(10000,mean(sample(tvalues,1) >= tvalues))
```

Check to see that these are roughly uniform with `plot(density(pvalues))`.

- In the first part of the lab we learned how you can use Monte Carlo to estimate any p-value for any statistical test. Of course, for more common tests, R has built in commands to give you the p-value. Let's start by loading in a data set that simply consists of recording the month whenever a cat falls from a particular high rise building.

```
cats <- read.csv("FallingCatsByMonth.csv")
```

Type `cats` to see what the data consists of. How many cats fell in December?

- Currently the data is not in a great state, as we are interested not in having multiple lines each with a month, but instead want a count of the data by each month. Here the `table` command is helpful for organizing the data. Try `table(cats)` and see how many cats fell in August.

- In total there were 119 cats falling during the year. Suppose that our hypothesis is that each of the 119 cats are equally likely to fall in any month. Then in order to test this, we need a test statistic that indicates that the data is far away from this ideal.

  Assume the data came from a non leap year, then we would expect a particular cat to have fallen in June with probability $30/365$. So on average $119(30/365) = 9.780$ cats would have fallen in June. The farther away the data is from that, then the more evidence we have that the data is not randomly distributed.

  Here 14 cats actually fell in June. To measure how far away this was from the average, we will use the square of the difference between the data and the average (and to make it have the same units as the data) we then divide by the average. So the cats in June contribute $(14 - 9.780)^2/9.780$ to the statistic. The overall statistic is the sum of the statistics for each month.

  Then is called the *chi-squared* statistic, because as $n \to \infty$, this statistic (if the null hypothesis of uniformity is true) will converge to a χ^2 distribution with $n - 1$ degrees of freedom.

$$\chi^2(\text{data}) = \sum_{i=1}^{n} (X(i) - \mathbb{E}[X(i)])^2 / \mathbb{E}[X(i)].$$

  Now let's do this in R. First we need our probability vector for the months. Since the months are in alphabetical order, we have to put the number of days in each month in alphabetical order as well:

```
pmonths <- (1/365)*c(30,31,31,28,31,31,30,31,31,30,31,30)
avcats <- pmonths*119
csqstat <- sum((table(cats)-avcats)^2/avcats)
```

  What is the chi-squared statistic `csqstat`?

- So now we have a number. But is it a large enough number to give strong evidence that the data was not uniformly generated? Well, if the data actually was uniformly generated, then the p-value will be

the probability that a $\chi^2(12-1)$ is at least the number from the last question. Find this probability with the `pchisq` command.

- Now, R has a built in command to calculate these things directly as well. Try

  ```
  chisq.test(table(cats),p=pmonths)
  ```

  to check your answer from the previous two parts.

Extended Lab

- Now consider how p-values work in the presence of an alternate hypothesis. Consider a statistical model that has $X \sim \mathsf{Exp}(\lambda)$. The null hypothesis is that $\lambda = 1$ while the alternate is that $\lambda = 1.4$.

  Then if we draw $X_1, \ldots, X_n$, we have our estimate $\hat{\lambda} = n/(X_1 + \cdots + X_n)$. If $\hat{\lambda}$ is large, that gives more weight to the alternate, which also has a larger spread of λ values. So let's use test statistic $T = X_1 + \cdots + X_n$, and reject the null hypothesis when T is too small (so $\hat{\lambda}$ is too big).

  Suppose our data gave for $n = 4$, $\hat{\lambda} = 1.2$ so $T = 4/1.2 = 3.33333$. What is the p-value? (That is, what is the probability that $T \leq 3.33333$ given $\lambda = 1$. Recall that the sum of n exponential random variables of rate λ has a gamma distribution with parameters n and λ.)

- The density of the p-value if the null hypothesis is true is uniform over $[0, 1]$. To see this in action, let's simulate a bunch of $T \sim \mathsf{Gamma}(4, 0.7)$ values and see what that chance that a new draw T_{new} is greater than each of them. This can be done by using the `rgamma` command to draw the values, then the `pgamma` command to calculate the p-values. For instance, try

  ```
  pvalues.null <- pgamma(rgamma(1000000,4,rate=1),4,rate=1))
  ```

  Now estimate the density by looking at a histogram of the values:

  ```
  plot(density(pvalues.null))
  ```

- Now suppose what the density of the p-values would be if the alternate is true. The p-value is the probability that a draw from the null will be worse than the value of T, so the `pgamma` of the above expression remains the same. However, if the alternate is true, then the random generation of the gamma random variables should be using rate 1.4 rather than 1. So try

  ```
  pvalues.alt <- pgamma(rgamma(1000000,4,rate=1.4),4,rate=1)
  plot(density(pvalues.null))
  abline(h=1,col="blue",lwd=2)
  ```

- Let's throw in a vertical line at 0.05 with `abline(v=1,col="gold",lwd=3)`.

Given that $p = 0.05$, give an eyeball guess at the density d of the p-value under the alternate. Then $d/(1 + d)$ will be the probability that the alternate is true given that initially they were each equally likely to be true.

- Repeat this process for 20 data values and see how d changes. [Note that the change in d is not proportional to the change in the number of data points!]

Chapter 40

Stats Lab: Hypothesis Testing

Instructions:

If you have time in the period, complete both the main and extended portions of the lab. If you run out of time, you do not have to complete the extended lab.

This week...

In this lab we'll look at another classic data set, which comes from Student's 1908 paper in Biometrika illustrating the use of the t test for hypothesis testing. We'll use both t-tests and Bayes Factors to test different hypotheses.

Main Lab

- Begin by loading the data into R. Because this is a built-in data set, you can do this with `data(sleep)`. This loads the data into the variable `sleep`. What are the factors for this data set?

- Look at this data set. You can see that there were ten people in the study, with ID 1 through 10. The first night every person in the group was given drug 1, then the next night every person was given drug 2. Therefore, there are 20 lines total in the data set for the 10 people for two nights. The factor labeled `extra` records the amount of extra sleep each person received as a result of the drug. Calculate the difference between the extra sleep for the two drugs with `ds <- sleep$extra[1:10]-sleep$extra[11:20]`. Let's get some stats on the data.

```
m <- mean(ds)
s <- sd(ds)
n <- length(ds)
```

  What is the mean difference between scores?

- Recall that the t test statistic for a data set x is $t = (\bar{x} - \mu)/(\hat{\sigma}/\sqrt{n})$. This can be used as a pivot to build a γ-level confidence interval for the value of μ:

$$\left[\bar{x} + \frac{\hat{\sigma}}{\sqrt{n}} \operatorname{cdf}^{-1}_{t(n-1)}((1-\gamma)/2), \bar{x} + \frac{\hat{\sigma}}{\sqrt{n}} \operatorname{cdf}^{-1}_{t(n-1)}((1+\gamma)/2)\right]$$

Find the 95%, 99.5%, 99.9% confidence intervals for this data.

- If we think about using these confidence intervals to test the null hypothesis that $\mu = 0$, use your intervals from the last question to get lower and upper bounds on the p-value.

- So far we've used t as a pivot. But we can also use this as a means of testing the hypothesis.

  If we assume that the null hypothesis $\mu = 0$ is true, then what is the value of the test statistic $t = (\bar{x} - \mu)/(\hat{\sigma}/\sqrt{n})$ for the data that has a t distribution?

- Recall that the t test statistic should have a distribution that is t with $n-1$ degrees of freedom. What is the probability that a draw from this distribution has absolute value greater than the t value you found for the differences?

- Now let's check your answers, using R's built in t-test function. Use `t.test(ds)` to get the t value and p-value for the t-test.

- Does this p-value provide strong evidence against the null hypothesis that $\mu = 0$?

- So far so good! If we were testing at the 5% level, we would certainly reject the null hypothesis that $\mu = 0$. Note that the value $t = (\bar{x} - \mu)/(\hat{\sigma}/\sqrt{n})$ is unit free. That means that if we scale the data by a fixed constant, the t value does not change. Verify this by trying `t.test(2*ds)`. What is the new t value?

Using R to calculate Bayes' Factors

- In this section we will again consider the t test, but now we will try hypothesis testing from a Bayesian perspective.

- The Bayes Factor for H_1 versus H_0 tells us the ratio of the probabilities $\mathbb{P}(H_1|X)/\mathbb{P}(H_0|X)$ where X is the data.

  Start with a simple example of calculating Bayes' Factors for the `sleep` data. Suppose we are trying to test $H_0 : \mu = 0$ against the alternative $H_1 : \mu = -1$. Start with a prior that is $\mathbb{P}(H_0) = 0.5$ and $\mathbb{P}(H_1) = 0.5$.

  If we calculate the t-statistic with $\mu = 0$ we get a number $t_0 = -4.062128$. If we calculate the t-statistic with $\mu = -1$ we get a new value t_1. What does t_1 equal?

- Let d_0 be the density of a $t(n-1)$ distribution evaluated at t_0, and d_1 be the density of a $t(n-1)$ distribution evaluated at t_1.

  Then Bayes Rule for continuous random variables says that are probabilities are proportional to the density of the test statistic that results from them, weighted by the prior probabilities. So

  $$\mathbb{P}(H_0|X) = \frac{d_0\mathbb{P}(H_0)}{d_0\mathbb{P}(H_0) + d_1\mathbb{P}(H_1)}, \quad \frac{\mathbb{P}(H_1|X)}{\mathbb{P}(H_0|X)} = \frac{d_1}{d_0} \cdot \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)}.$$

  What is the Bayes factor for H_1 over H_0? This is equivalent to the odds of H_1 versus H_0. [Note under the Jeffries scale Bayes Factors greater than 100 are considered decisive, from 30 to 100 is very strong, 10 to 30 is strong, and 3 to 10 substantial.]

- Unfortunately we don't often know that either $\mu = 0$ or $\mu = -1$ or something similar. Instead, we have a null hypothesis $H_0 : \mu = 0$ and $H_a : \mu \neq 0$. In order to undertake a Bayes Factor analysis for this type of result, we need to make the alternate hypothesis stronger: namely, we need to specify a distribution for μ under the alternate. One way to do so is to use a scaled Cauchy random variable for μ.

  This is a nice alternate because it allows for μ to vary and (unlike a normal distribution) allows for reasonable chances for very small and large values of μ. In other words, it allows the data to inform the posterior.

  To be precise, let $H_a : \mu \sim (\sqrt{2}/2)X$, where $X \sim$ Cauchy. Another way to say this is that the null is $H_0 : \mu = 0$, while the alternate is $H_a : \mu \sim f_a(s)$, where

  $$f_a(s) = \frac{2}{\tau} \cdot \frac{1}{1 + s^2/2}.$$

  When you think of the null and alternate this way, this makes the prior distribution on μ a mixture of the discrete distribution that puts 100% of the probability on 0, and the continuous scaled Cauchy distribution. This weights of the mixture are the prior probabilities for H_0 and H_1. Then the Bayes analysis continues as usual: Given the prior, use the likelihood of the data to build the posterior. For $T = \bar{x}/(\hat{\sigma}/\sqrt{n})$, given μ the density of $T - \mu/(\hat{\sigma}/\sqrt{n})$ will be that of a $t(n-1)$ distribution. Hence for $T = t$, the density is $f_{t(n-1)}(t - \mu/(\hat{\sigma}/\sqrt{n})$.

  Using Bayes' rule is tricky for mixtures of discrete and continuous distributions. The resulting posterior will also be a mixture of a discrete and continuous distribution. The result will have $\mathbb{P}(\mu = 0|T = t) \propto \mathbb{P}(H_0)f_{t(n-1)}(t)$, and $\mathbb{P}(\mu \in ds|T = t) \propto f_{t(n-1)}(t - s/(\hat{\sigma}/\sqrt{n}))f_a(s)$ for $s \neq 0$. To see this posterior and the density, try plotting both with

```
x <- seq(-5,5,by=0.01)
dprior <- dcauchy(x,scale=sqrt(2)/2)
dlike <- dt((m - x)/(s/sqrt(n)),n-1)
plot(x,dprior,type="l")
lines(x,dlike,type="l",col="blue")
```

- Now let's plot the product of the prior and the likelihood:

```
plot(x,dprior*dlike,type="l")
```

(Remember, that just gives something that is proportional to the posterior, not the actual posterior.)

- To find the integral of the continuous part of the prior times the likelihood, just use the left endpoint rule (you could use the trapezoidal rule, but frankly it is overkill for these types of problems where the endpoints are so close to 0.)

```
c <- 0.5*sum(dprior*dlike)*(x[2]-x[1])
```

- Next find the contribution of the discrete part of the prior times the likelihood:

```
d <- 0.5*dt(t,length(ds)-1)
```

- Then the Bayes Factor in favor of the alternate is `d/c`, and the Bayes Factor in favor of the null is `c/d`. What is the Bayes Factor in favor of the alternate?

- One thing that we might worry about is that our choice of scaling factor for the Cauchy unduly influenced our Bayes' Factor. Repeat the analysis for a scale factor of 1 rather than $\sqrt{2}/2$.

- Repeat the analysis for a scale factor of 0.1.

Extended Lab

Using `prop.test` in R

- The `t.test` function isn't the only kind of test in R. For instance, suppose we wanted to estimate the probability p that when the person received drug 2 they got at least 1 more hour of sleep than with drug 1.

```
b <- sum(ds <= -1))
```

  Out of the 10 participants, how many received at least 1 more hour of sleep with drug 2 than drug 1?

- If we view each of the trials/people as being indepenent, then we want to estimate the proportion that get at least one more hour of sleep with drug 2 versus drug 1. In R the test of proportionality is `prop.test`. Using `?prop.test`, we see that we need to provide the test with the counts of success, `b` in our case, the total number of trials `length(ds)`, and optionally a value of p for the null hypothesis. Try

```
prop.test(b,length(ds),p=0.5)
```

That will test to see if we can reject the null hypothesis that $p = 0.5$. It also gives a 95% confidence interval for the value of p so that you can see the effect size. Report your p-value.

- Use `prop.test` to get a 99% confidence interval for p.

A/B Testing

- One type of testing often done in statistics is A/B testing, where a customer is presented with option A or B and we attempt to find out which the customer likes more.

  For example, a company might have two versions of a homepage on their website. The first version is A, the second is B. The idea is to measure the attractiveness of the pages by measuring their click through rate (CTR).

  We will start by generating two sets of binomial data, one for the 500 visitors to the A page, and one for 500 visitors to the B page. You can see from the parameter that it is set up so that customers have a higher CTR for the B page.

```
A <- rbinom(1,500,0.3)
B <- rbinom(1,500,0.35)
```

Frequentist approach: Fisher's exact test

- Our null hypothesis is that there is no difference between choice A and B. In this case, we could make a little table with the possibilities:

	Did click	Did not click
Choice A	A	$500 - A$
Choice B	B	$500 - B$

  A table like this in statistics is called a *contingency table*.

  Fisher studied this type of problem and showed that under the null hypothesis that the rows of the table are independent, then a particular statistic called the χ^2 statistic will allow you to calculate a p-value.

  Fortunately, you don't need to know much about the distribution to test the null. First, we'll put our data into a table.

```
table <- rbind(c(A,500-A),c(B,500-B))
```

  Now run Fisher's Exact Test for Count Data using

```
fisher.test(table)
```

  What is the p-value for your data?

- Draw new data A and B and repeat the Fisher exact test. What was your p-value this time?

- Would you say that the p-value is reliable?

Bayesian A/B testing

- In order to undertake Bayesian testing, we start with prior knowledge about the parameter. Suppose our history of webpages tells us that the prior distribution for parameter concentrates roughly between 0.2 and 0.5. So we decide to use a beta distribution centered at $(0.2 + 0.5)/2 = 0.35$ to make this happen.

```
plot(x,dbeta(x,35,65),type="l")
x<-seq(0,1,by=0.01)
```

  Sketch the prior on p.

- Recall that the beta distribution is a conjugate prior for binomial data. So if $p \sim \mathsf{Beta}(a, b)$ and $[X|p] \sim \mathsf{Bin}(n, p)$, then $[p|X] \sim \mathsf{Beta}(a + X, b + n - X)$. Let's put up the posterior distributions for the data collected so far

```
x <- seq(0.2,0.5,by=0.001)
plot(x,dbeta(x,35+A,65+500-A),type="l",col="blue")
lines(x,dbeta(x,35+B,65+500-B),type="l",col="red")
```

- Let p_a be the probability of click through for choice A and p_b the probability of click through for choice B. Then what we want to know is, given this data, what is $\mathbb{P}(p_a < p_b)$?

  To test this, we will estimate the probability that $p_a < p_b$ by generating a number of draws from A and calculate the chance that the choice B value of p is smaller than it.

```
pa <- mean(pbeta(rbeta(10000,35+A,65+500-A),35+B,65+500-B))
```

  What does your data return as an estimate for p_a?

- As with the Fisher test, generate new data A and B and repeat the process.

- Note: most likely both your p-values and Bayesian p_a changed dramatically with different data. What this is telling us is that we need much more than 500 data points to distinguish a 0.3 CTR from a 0.35 CTR.

Chapter 41

Starts Lab: Testing with two samples

Instructions:

If you have time in the period, complete both the main and extended portions of the lab. If you run out of time, you do not have to complete the extended lab.

Main Lab

Using the two-sample t-test In this lab we'll be looking at a data set from Frisby and Clatworth (1975). In this experiment, they gave random dot stereograms to 78 participants, and then asked each participant to report how long it took them to fuse the steorogram, that is, to "see the hidden figure". Out of the 78 subjects, 35 participants were given a hint as to what the target image was.

- You can look at this data by going to the website http://lib.stat.cmu.edu/DASL/Datafiles/FusionTime.html. This website (hosted by Carnegie Mellon University) contains many data sets, along with the paper they came from and the story behind them. The name of the site DASL, stands for Data and Story Library. What are the names of the two groups of participants in this data set?

- One nice thing about R is that you can load data directly from a website into the console by using the `url` option in `read.table`. Try

```
randDotStereo <- read.table(url("http://lib.stat.cmu.edu/DASL/Datafiles/
    FusionTime.html"), header = FALSE, skip = 33, nrows = 78)
colnames(randDotStereo) <- c("fuseTime", "condition")
```

  to bring it into R, and to label the two columns. Sketch a plot of the density of the fuseTime created with `plot(density(randDotStereo$fuseTime))`.

- Notice that the tail extends far to the right. This characterizes the data as *heavy tailed* or *right-skewed*. One way to turn heavy tailed data into light tailed data is to just take the logarithm of the data values. Use

```
randDotStereo$logFuseTime <- log(randDotStereo$fuseTime)
```

Sketch the density plot of the result.

- The data is bimodal, because the NV and VV groups have been combined into one group. These groups can be separated out with a boxplot. Without going into too much detail, a boxplot tries to capture the center of the distribution of the data in a rectangle known as a box. The middle third of the data lie inside the box, with the median marked by a thick black line. To break down the groups in the boxplot, try

```
boxplot(logFuseTime~condition,data=randDotStereo,col="lightblue")
```

  Sketch the result.

- So the VV group definitely did better in response time. Did they do well enough to state definitively that they are better though? Let's try computing Welch's t-test:

```
t.test(logFuseTime ~ condition,data=randDotStereo)
```

  Give the 95% confidence interval for the difference and the p-value.

- Based upon this, would you say that the difference in the sample means is statistically significant?

- In the test, we cared if the means were either positive or negative. However, suppose the experimenter was only interested if the times for the NV group were greater than the times for the VV group. Then we would want to use a one sided test. Use

```
t.test(logFuseTime ~ condition,data=randDotStereo,alternative="greater")
```

  to test this hypothesis. Again, give the 95% confidence interval as well as the p-value.

- Suppose we believed the VV group would have higher fuse times than the NV group. Use

```
t.test(logFuseTime ~ condition,data=randDotStereo,alternative="less")
```

  to test this hypothesis. Once again, give the 95% confidence interval as well as the p-value.

- Now let's do the Bayes Factor approach. This time, let's use a package `BayesFactor` to accomplish this. First load the BayesFactor library using `library(BayesFactor)`. If the library is not already installed on your computer, you have to use `install.packages("BayesFactor")` before this command to install the package. Then use

```
ttestBF(formula=logFuseTime ~ condition,data=randDotStereo)
```

to give the Bayes equivalent of the t-test. What is the Bayes Factor?

- Do you think this evidence is as strong as the p-value evidence given earlier?

- Just as with the t-test from earlier, it is possible to do a one-sided test on the data:

```
ttestBF(formula=logFuseTime ~ condition,data=randDotStereo,
  nullInterval = c(0,Inf))
```

to find the Bayes Factor of no effect versus positive effect. What is the Bayes Factor now?

- Does it make sense that the Bayes Factor is higher now that the null is more restrictive?

- You can also change the Bayes factor by changing the prior associated with the alternative. When the difference of means divided by the standard deviation is not zero, it is modeled as a Cauchy. The parameter `rscale` which can take on either `medium`, `wide`, or `ultrawide` gives a sense of how spread out the alternative is. Try

```
ttestBF(formula=logFuseTime ~ condition,data=randDotStereo,
  rscale = "wide")
```

and report the Bayes factor.

- Note that the wider the prior, the less evidence the data provides against it not being 0.

Nonnormal versus normal data Recall that the two sample t-test assumes that the data is normal in finding it's results. Let's take a look at how closely that assumption is needed, and how things go when it is violated.

- Now let's look at how the t-test is affected by nonnormality of the data. Try the following experiment. Create a new script file `script.R` and add the following code

```
normt <- function(n = 5,m = 7) {
  x <- rnorm(n,10,1)
  y <- rnorm(m,10,1)
  return(t.test(x,y)$p.value)
}
```

This essentially draws two random data sets of size n and m from the normal distribution, runs a two sample t-test on them, and returns the p-value results.

Recall that if the null hypothesis is true, then the p-value should have a uniform distribution. Try this out with

```
resultsn <- replicate(100000,normt(5,7))
plot(density(resultsn))
```

Sketch the result.

- Now suppose that our data had the same mean of 10, but was not normally distributed, but instead had an exponential distribution. Try adding

```
expt <- function(n = 5,m = 7) {
  x <- rexp(n,rate=1/10)
  y <- rexp(n,rate=1/10)
  return(t.test(x,y)$p.value)
}
```

to your `script.R` file, re-sourcing it, and then using

```
resultse <- replicate(100000,expt)
plot(density(resultse))
```

Sketch the resulting p-values.

- Note that these p-values are not uniform, but skew left. That means that a resulting p-value less than 0.05 does not occur 5% of the time. To find out how often the p-value from the test is at most 0.05, use

```
mean(resultse < 0.05)
```

in the console. This behavior comes from the fact that the assumptions of the data were not satisfied.

Using nonparametric tests

- Now let's try our nonparametric test. Use

```
wilcox.test(logFuseTime ~ condition,data=randDotStereo)
```

What p-value did it find?

- Is this similar to what you found using the parametric test from earlier?

Extended Lab

Now let's see how the tests do when faced with nonnormal data, and see how that compares to the normal data.

- Start by drawing normal data to test if they come from the same distribution:

```
z1 <- rnorm(78,mean=0)
z2 <- rnorm(71,mean=0.9)
```

  Using `t.test`, what is the p-value associated with them?

- Now test the same `z1` and `z2` data using `wilcox.test`. What is the p-value?

- This time let's give them something more difficult to discern, Cauchy (heavy-tailed) data that has the same medians as above.

```
c1 <- rcauchy(78,location=0)
c2 <- rcauchy(71,location=0.9)
```

  What does `t.test` give as the p-value? [Important note: it is completely inappropriate to run `t.test` on this data as it is not normally distributed! This is just to see what it reports when the data is not normally distributed.]

- Now analyze the same data with `wilcox.test`. What is the p-value?

- The point is that the nonparametric test is easily able to tell the two heavy-tailed data sets apart, while the t-test (that assumes the data is normal) is not.

Fisher information: A non-regular score function

- Suppose that $[X|\theta] \sim \mathsf{Unif}([0,\theta])$. Then given $X_1, X_2, \ldots, X$ iid from $[X|\theta]$, a simple unbiased estimate for θ is

$$\hat{\theta} = \frac{n+1}{n} \max_i \{X_i\}$$

- Let's check this is unbiased by creating a function:

```
test.thetahat <- function(n,theta) {
  return((n+1)/n*max(runif(n,min=0,max=theta)))
  }
```

  Try out the function with `test.thetahat(10,2)`. What was the result?

- We'll check that the estimate is unbiased by repeating the estimate multiple times.

```
results <- replicate(10000,test.thetahat(10,2))
print(mean(results))
print(sd(results)/sqrt(length(results)))
```

Does the estimate $\hat{\theta}$ appear to be unbiased?

- What is the density of $[X|\theta] \sim \mathsf{Unif}([0, \theta])$?
- What is the natural logarithm of the density?
- What is the score function $S(s) = (\partial \ln(f_{X|\theta}(s)))/\partial\theta)$?
- Is there any way that the mean of this score function could be 0? Explain.
- Is there any way that this score function could be regular? Explain.
- Note that for score functions that are non regular, an estimator (such as $\hat{\theta}$) does not have to obey the Cramér-Rao inquality!

Chapter 42

Stats Lab: ANOVA

Instructions:

If you have time in the period, complete both the main and extended portions of the lab. If you run out of time, you do not have to complete the extended lab.

In this lab you will be learning the building blocks of Analysis of Variance, which has the weird abbreviation ANOVA, pronounced Ah-nova. [As a piece of grammar trivia, since ANOVA is an abbreviation pronouced as a word, it is an acronym. However, unlike many acronyms such as laser and scuba, the abbreviation is not formed solely from the initial letters in the phrase it is abbreviating, so it is not an acrostic.]

Main Lab

- Open a spreadsheet and enter the table of values

A	B	C
11	14	26
23	17	13
9	16	24
10	16	19
12	8	

  Save the spreadsheet as a comma separated value (csv) file and then load it into R into the variable `adcampaign` using `read.csv`. Look at the data. Are all the entries numeric values?

- Since missing data is a pretty common thing in statistics, R has a special value for indicating that data is not there. The NA entry stands for not available. Other nonnumerical values in R include NaN for not a number, which you'll get if you try to divide by zero. Try to get the means of the columns using `colMeans(data)`. What is returned as the mean of the column with the NA entry?

- Using `?colMeans`, we see that the function has a parameter `na.rm`, which if you set to `TRUE`, removes the NA entries when computing the means. Try using `colMeans(data,na.rm=TRUE)` and report the column means.

- Use `barx <- sum(data,na.rm=TRUE)/14` to get the overall mean of the data. What is this mean?

- Now let's get the total sum of squares with:

```
sst <- sum(sum((data-barx)^2,na.rm=TRUE))
```

  What is SS_T for this data set?

- The next step is to find the sum of squares between blocks, which remember is

$$SS_B = \sum_{j=1}^{k} \sum_{i=1}^{n_j} n_j(\bar{x}_{\cdot j} - \bar{x})^2$$

  where n_j represents the number of data points in column j. At this point, you can calculate this in R with

```
n <- c(5,5,4)
ssb <- sum(n*(colMeans(data,na.rm=TRUE)-barx)^2)
```

  What is SS_B for this data set?

- Remember, to find the mean square, divide the sum of squares by the degrees of freedom, which is $k-1$. Here k refers to the number of different treatment levels, so in this case $k = \#(\{A, B, C\})$. (Remember the degrees of freedom of the population variance for a k dimensional vector is only $k-1$.) So what's MS_B, the mean square for between blocks?

- Recall that $SS_T = SS_W + SS_B$. Also, SS_W has $N-k$ degrees of freedom. Use this to find MS_W.

- Under the null hypothesis, the statistic $F = MS_B/MS_w$ will have the F distribution with parameters 2 and 11 (the degrees of freedom of MS_B and MS_W respectively.) Verify that the F statistic has value 2.706 for this data.

- Now sketch a plot of the density of the F distribution using

```
x <- seq(0,10,length=100)
plot(x,df(x,df1=2,df2=11),type="l")
abline(v=2.706,lwd=3,col='gold')
```

- Given this plot, do you think "as or more extreme" means that the statistic is greater than 2.706, or less than 2.706?

- Find the probability that an F statistic with parameters 2 and 11 is "as or more extreme" than 2.706 using the `pf` function.

- The F distribution comes from $(X/d_1)/(Y/d_2)$, where X and Y are chi-squared distributed with d_1 and d_2 degrees of freedom respectively. Because this is division, the order of d_1 and d_2 is very important. Try graphing the density of an F distribution with parameters 11 and 2 (rather than 2 and 11 as earlier) to see what a difference it makes. Sketch the plot.

- Okay, so we did all of this by hand, but surely R has built-in functionality to create an ANOVA table, right? Of course it does! In order to use it, however, we need to put the data as a single vector, with a label for each element of the vector indicating if the treatment was `A`, `B`, or `C`.

  First, put the data rows in `data` into a single row vector `r`:

  ```
  r <- c(t(as.matrix(data)))[1:14]
  ```

  What is `r[4]`?

- Next, create labels for the treatments and values for k and n:

  ```
  f <- c("A","B","C")
  k <- 3
  n <- 6
  ```

  Now we want to create a vector that tell the treatments for each entry in `r`. For instance, the fourth entry in `r` should correspond to treatment A. The `gl` (generate factor levels) can be used to accomplish this automatically rather than by hand. Try

  ```
  tm <- gl(k,1,n*k,factor(f))
  tm <- tm[1:14]
  tm
  ```

  As you can see, this has created a single factor (that's the 1 in the second argument to **gl** that takes on level values A, B, and C. It also created a list of 14 factors corresponding to our values in `r`. Just to check that you got it right, how many positions in `tm` received the label of level C?

- Now that everything is labeled, create a linear model where `r` depends on `tm`.

  ```
  lm1 <- lm(r ~ tm)
  ```

Use `anova(lm1)` to print out the ANOVA table. Write down this table below.

- So, recall that we used SS_B to denote the sum of squares between blocks, and it should have $k-1$ degrees of freedom. SS_W is the sum of squares within blocks, and it has $N-k$ degrees of freedom. So from your table above, what are SS_W and SS_B?

- Does this match what you calculated by hand earlier?

- While the p-value of 11% is not statistically significant at the 95% level, it perhaps does warrant additional investigation. Suppose that there was in fact no relationship between the treatment and the data values. This can be simulated by taking a random permutation of the vector r. Try

```
r2 <- sample(r,length(r))
lm2 <- lm(r2 ~ tm)
anova(lm2)
```

What is your new p-value for this data?

Extended Lab

- In this part of the lab we'll look at how to output our results from an ANOVA table for use in a report. Start by loading in the `knitr` package. As always, if it isn't already installed, you will need to use `install.packages("knitr")` to do so.

```
library(knitr)
```

- For this part of the lab we'll use another built in data set in R. This one records the number of times yarn broke under varying conditions. Two type of wool were used, A and B, and tension was either low (L), medium (M), or high (H). Just type `warpbreaks` to see the data.

  To see the data organized for ANOVA, use the structure command `str`. Try

```
str(warpbreaks)
```

  How many factors are there in this data set?

- Before you run a formal analysis on data, it is useful to run some preliminary graphics. For instance, we can look at histograms of the data to get an idea where they might lie. For that we'll use the `ggplot command`.

  This command is far more powerful than the simple plot, and allows us to add pieces to a plot using commands separated by a `+` sign.

```
library(ggplot2)
ggplot(warpbreaks, aes(x=breaks)) +
  geom_histogram(bins=10) +
  facet_grid(wool ~ tension) +
  theme_classic()
```

Under low tension, what does the histogram indicate is the type of wool that breaks more easily?

Next we'll do a comparative boxplot. Without going into too many details now, a boxplot puts a thick line at the median of data, the top of the center rectangle is the 75% quantile, and the bottom of the center rectangle is the 25% quantile. The following commands makes boxplots for the different tensions, and fills the boxes with different colors for different wools.

```
ggplot(warpbreaks, aes(y=breaks, x=tension, fill = wool)) +
  geom_boxplot() +
  theme_classic()
```

Under which tensions (if any) is wool A better? What the ANOVA can hopefully do for us is tell us

if we have gathered enough evidence to find the differences in the tables. Let's give it a shot. Here `wool * tension` means our model will include the factors wool, tension, and the "product" of wool and tension.

```
model <- lm(breaks ~ wool * tension, data = warpbreaks)
```

Now let's run an ANOVA on the table, and save the result in variable `sstable`.

```
sstable <- anova(model)
sstable
```

For which effect are the p-value below 5%?

- Now suppose that we needed to put that table `sstable` into a report. The `kable` command does this nicely. For instance, rounding to the second decimal digit (*not* to two signifcant digits) can be done with

  ```
  kable(sstable,digits=2)
  ```

  This just uses ASCII text characters to create the graphics. Now let's try creating the table for use in LaTeX:

  ```
  kable(sstable,format='latex',digits=2)
  ```

  If you know LaTeX, you will see that it has created a `tabular` environment and put everything in place. By the way, from a formatting perspective, people (and the default option in `kable`) tend to way overuse the verticle bar option in `tabular`. Using `lrrrrr` instead of `l|r|r|r|r|r` will create a label in LaTeXthat is much easier on the eyes.

- Next let's try making it into an HTML formatted table.

  ```
  kable(sstable,format='html',digits=2)
  ```

  Try taking this output, copying it to a file `anovatable.html` and opening the result in a browser to see the final effect.

Chapter 43

Stats Lab: Correlation

Instructions:

If you have time in the period, complete both the main and extended portions of the lab. If you run out of time, you do not have to complete the extended lab.

In this lab you will learning how to build scatter plots, perform linear regression and locally weighted scatterplot smoothing, and get a close look at Simpson's paradox.

- Start by generating some random data with a strong correlation.

```
x <- seq(0,4,by=0.1)
y <- 0.7*x+rnorm(length(x))
plot(x,y)
```

Even though your simulated data will (with high probability) have a high correlation, it is difficult to see from the graph. Start with a basic correlation statistic. We'll put the `x` and `y` variables into a data frame, and then use the `cor` function.

```
simdata <- data.frame(x,y)
cor(simdata)
```

What is the correlation between x and y? Is this correlation positive or negative?

- Repeat the last step but this time use `y <- -0.7*x+rnorm(length(x))`.

LOWESS Unfortunately, not all data is nearly linear the way our simulated data was. Locally weighted scatterplot smoothing (LOWESS) is a way of handling data that doesn't just lie on a straight line. To see how this works, let's generate some data from a sine curve plus some normal noise. Try the following.

```
x <- seq(0,4,by=0.1)
y <- sin(x)+0.5*rnorm(length(x))
plot(x,y)
```

Sketch the result.

- The thing about linear regression, is that you can fit a least squares line through *any* set of points, even for data which (as here), a straight line seems like a bad option. Add the least squares line to your plot above using

```
abline(lm(y~x),col="red",lwd=3)
```

  You can see it tries its best to "average" the plot, but just doesn't fit something that is changing very well.

- In the late 1970's, Cleveland introduced locally weighted polynomial regression. In this technique, near each point a low order polynomial (determined by the neighboring points in the area) are introduced. Use

```
lowess(x,y)
```

  to get the (x, y) points for a variant of this idea called LOWESS, and use

```
lines(lowess(x,y),col='blue',lwd=3)
```

  to add this fit to your plot. Which method, least squares or LOWESS, would you say fits your data better?

- The `lowess` function has a smoothness parameter `f` that controls how smooth the function is. Its default value is 2/3. Try

```
lines(lowess(x,y,f=1),col='blue',lwd=3)
```

  to see what happens when the curve is lower smoothness. Try adding a line with $f = 0.1$ to see what happens when you allow the curve to wriggle too much. As with all fitting methods in statistics, there is a tradeoff between trying to fit too tightly versus fitting too loosely.

Visualizing data Graphing the data before you conduct a regression of any kind is very important. When the data is already close to a line, using LOWESS doesn't gain you much. You can see this by looking at a data set `faithful` that is built into R. This data set consists of 272 data points, each of which records the eruption time in minutes of Old Faithful geyser in Yellowstone, together with the waiting time (also in minutes) until the next eruption.

```
erupt <- faithful$eruptions
wait <- faithful$waiting
plot(erupt,wait,main="Geyser data",xlab="Eruption times (minutes)",
  ylab="Time until next eruption (minutes)")
abline(lm(wait~erupt),col="red",lwd=3)
lines(lowess(erupt,wait),col="blue",lwd=3)
```

Sketch the result.

- Let's use R to estimate the correlation using Pearson's sample correlation coefficient.

```
cor(erupt,wait,method="pearson")
```

  [Note if you do not specify the method, the default value is `"pearson"`.] What is the correlation?

- You can find the R^2 value using `summary(lm(wait$\sim$erupt))`. Verify that this is the square of the Pearson correlation coefficient.
- Now find the Kendall's Tau value for the data and the Spearman's Rho.

- They vary pretty widely. Which is correct?

  As usual in statistics, there is no one right answer as to which is right. But notice in the graph of the data, the points really live in two clumps, one at the lower left, and one in the upper right. Let's separate out these two clumps by sorting the data frame by eruption time:

```
newdata <- faithful[order(erupt),]
```

  Now plot the first 100 points in newdata (with eruption times 3.333 or smaller) and find their correlation.

```
plot(newdata$eruptions[1:100],newdata$waiting[1:100])
cor(newdata$eruptions[1:100],newdata$waiting[1:100])
```

  What is the correlation among this clump in the lower left corner?

- Find the correlation in the upper right clump using data points number 101 through 272.

- No longer the 0.9 correlation from earlier. In fact, let's build an example that is even more pronounced!

```
x1 <- 1:50; x2 <- 51:100
y1 <- 51-x1; y2 <- 150-x2
plot(c(x1,x2),c(y1,y2)).
```

  Sketch the plot.

- What is the Pearson correlation coefficient for `c(x1,x2)` and `c(y1,y2)`?

- What is the Pearson correlation coefficient for `x1` versus `y1`?

- What is the Pearson correlation coefficient for `x2` versus `y2`?

- D'oh! This is another example of what is called Simpson's Paradox. You can have clumps of data that independently have a low or even negative correlation, but when combined, they give a positive correlation! The only remedy to overcoming this paradox is to graph your data, so you can see what's going on with it directly!

Extended Lab

- So far we have only used the `cor` function to calculate the correlation between two sets of data points, but it can actually find the correlation between multiple vectors simulataneously. The result is an estimate of the *correlation matrix* between the different entries.

  To illustrate this, let's look at the built-in R data set `mtcars`.

  ```
  data("mtcars")
  cor(mtcars)
  ```

  Because there are 11 different columns of numbers in `mtcars`, the correlation matrix is an 11 by 11 matrix. What is the correlation bewteen the mpg (miles per gallon) and cyl (the numbers of cylinders) in the car?

- Repeat the last part using Kendall's Tau rather than Pearson's R.

- It is often helpful to use summary visualization tools to get a "big picture" look at a matrix. For instance, try

  ```
  symnum(cor(mtcars))
  ```

  Note that this breaks down the values in the matrix based on their absolute value. What is the upper left 3 by 3 matrix produced by this command?

- The heatmap command (`heatmap`) can also be used to visualize the matrix. Note that by default it reorders the rows and columns in an attempt to bring similar factors together based on their correlation. Which factor is more correlated with the carburetor?

- Another way to visualize the correlation matrix is through a correlogram. To use this in R, first load in the library corrgram and then the corrgram command will be available. As always when loading libraries, it might be necessary to use `install.packages("corrgram")` first to download before using.

  ```
  library(corrgram)
  corrgram(cor(mtcars))
  ```

Note that since the correlation of any variable with itself is 1, the `corrgram` command using the diagonal entries to label the rows and columns to save space. Does the color Blue correspond to positive entries, or negative entries? (Note that there is a visual hint within each red square and blue square in case you forget which is which!)

- Now if you really want to dump everything into a matrix like format, install the package PerformanceAnalytics, and use

```
library(PerformanceAnalytics)
chart.Correlation(mtcars)
```

Information galore! Because it knows the correlation matrix is symmetric, it takes advantage to put different info in the upper half and the lower half of the display. Sketch the histogram for mpg.

Chapter 44

Stats Lab: Logistic Regression

Instructions:

If you have time in the period, complete both the main and extended portions of the lab. If you run out of time, you do not have to complete the extended lab.

With some models, the question is whether a binary choice leads to variance in a response variable. For instance, the classic question of whether a smoker is more likely to get lung cancer, or whether or not a child participates in preschool programs lead to higher SAT scores down the road. In this lab you will be learning how to deal with yes-no data, and logistic regression.

Main Lab

Often, we are interested in predicting the probability of an event. Does smoking change the probability of cancer? Does pre-school increase the chance of high-school graduation? The response variable tends to be Yes/No for these situations. Yes the subject did get cancer; No the subject did not graduate high school.

Linear regression is not well-suited for these tasks. Lines are unbounded, and do not lie in $[0, 1]$ like probabilities do. To solve this problem, in 1958 David Cox introduced the idea of Logistic regression (aka logit regression or the logit model) and his simple idea is becoming more widely used every day in machine learning.

The idea is to predict the probability that an event occurs given the value of x as

$$y = \frac{\exp(c_0 + c_1 x)}{1 + \exp(c_0 + c_1 x)}.$$

Since the exponential function is always nonnegative, the ratio that is y will always lie between 0 and 1.

This is called *logistic regression* since the function y is known as the logistic function. That is because y is a solution to a differential equation with exponential growth restrained by limited resources, and logistics is the study of moving resources to where they are needed.

- In this lab we will learn about several extensions to R that give extra ways to manipulate data known as the tidyverse. Use `install.packages("tidyverse")` if the tidyverse is not already installed on your system. Now let's bring in three libraries.

  ```
  library(tidyverse)
  library(modelr)
  library(broom)
  ```

  The data set we will be using is the `Default` data that is part of the `ISLR` package. To use ISLR, we have to give the command `install.packages("ISLR")` if it is not already installed. A tibble is much like a data.frame in the tidyverse. The data that we are using is simulated data for the rate of defaults dependent on various factors such as the balance of the loan and the income of the person taking the loan.

```
default <- as_tibble(ISLR::Default)
head(default)
```

The `default` data records Yes/No data for "did the customer default" along with the balance of the loan, the income of the customer and Yes/No for "is the customer a student."

- What we will do next is break our data into a *training sample* which we use to fit coefficients and a *testing sample* so we can check how well the fit works. First we will set the random number seed: this means that the "random" numbers generated for each person will be the same, and so everyone doing the lab will get the same answers.

```
set.seed(323)
trainrows <- sample(1:nrow(default),floor(nrow(default)*0.6))
train <- default[trainrows,]
test <- default[-trainrows,]
```

Verify that the head of `train` and `test` are different rows of data.

- Now let's visualize our default versus balance of the loan

```
plot(train$balance,train$default)
```

Notice in the plot that it translated Yes on default to a value of 2, and No on default to a value of 1. Sketch the result.

- This is the point of logistic regression: a line through these data points concentrated as $y = 1$ and $y = 2$ is not going to get near a lot of the points. So we fit $y = \exp(c_0 + c_1 x)/(1 + \exp(c_0 + c_1 x))$ with a *generalized linear model*. In this case, because there are two values for the default, the model is a Binomial Logistic Regression.

```
blr <- glm(default ~ balance, family = "binomial", data = train)
```

(Just a detail: the `glm` command uses an optimization program to find the MLE for the coefficients c_0 and c_1.) Now let's see what it came up with using `ggplot` to create a nicer graph than we can get with the `plot` command.

A couple things about the next set of commands. First, we are using a *pipe* designated by `%>%`. What this does is pipe values from one variable to another variable or command. So the first command pipes the values of `default` to the `mutate command`, which then pipes its output to the `ggplot` command for visualization. Pipes are a nice way of keeping track what is being fed into what without the need to define a bunch of extra variables.

What the `mutate` command is change (mutate) the data and turn a default value of Yes into a 1, and No into a 0.

```
default %>%
  mutate(prob = ifelse(default=="Yes",1,0)) %>%
  ggplot(aes(balance,prob)) +
  geom_point(alpha = 0.15) +
  geom_smooth(method = "glm",method.args=list(family="binomial")) +
  ggtitle("Binomial Logistic Regression") +
  xlab("Balance") + ylab("Probability of default")
```

Sketch the resulting graph.

- Estimate from your graph, when the balance is \$1500, what is the probability of default?

- To get a more detailed look at the regression, use

  ```
  summary(blr)
  ```

  To focus in on just the coefficients, try

  ```
  tidy(blr)
  ```

  What that means is that the probability of default is predicted to be:

  $$y = r/(1+r),\ r = \exp(-10.35727 + 0.005370624x)$$

  where x is the balance. Using this, compute the predicted probability of failure when the balance is \$1500.

- We can get confidence intervals on the coefficients using

  ```
  confint(blr)
  ```

  Use these to get a 95% confidence interval on the prediction of failure when the balance is \$1500.

- Of course, R has a command to give the predictions based on the balance values so you don't have to calculate them yourself.

  ```
  predict(blr,data.frame(balance=c(1400,1500,1600)),type="response")
  ```

  Suppose the bank wants to know the largest balance amount (to the nearest dollar) where the probability of default is predicted to be at most 7%. Find this value.

- So far we have done our prediction using a continuous variable `balance`, but we could also create a prediction using the binary variable `student`

```
blr2 <- glm(default ~ student, family = "binomial", data = train)
tidy(blr2)
```

  What is the coefficient for Yes values for student?

- What does the sign of this coefficient tell us about the likelihood of students defaulting versus nonstudents?

Extended Lab

So far we've been doing our regression based on a single variable, but there is no limit to how many variables that we can use. Let's try including all the factors to predict defaults.

```
blr3 <- glm(default ~ student + balance + income, family="binomial",data=
    train)
tidy(blr3)
```

What is the coefficient of balance now?

- Okay, so the coefficient is positive, which means as the balance goes up, the chance of default goes up. What happens as income increases?

- When we just looked at student, being a student increased the chance of default. But now with our bigger model, being a student decreases the chance of default? What's going on? What's going on is that we have data which is correlated. Let's look at the balance of loans taken by students versus the balance of loans taken by nonstudents using the `tidyverse` version of the boxplot.

```
ggplot(train,aes(student,balance))+geom_boxplot(aes(colour=student))
```

  Sketch this plot.

- What this plot shows is that on average, students take out loans with higher balances. And loans with higher balances are more likely to default. So if we just look at the student status, we are likely to see being a student as a negative predictor.

  In fact, the balance of the loan being high makes the borrower both more likely to be a student and less likely to pay back the loan. The balance is an example of what is called a *confounding variable.*

To eliminate the effect of the confounding variable, let's go back to our balance of $1500 example from earlier, and pick an income of $40 000. Now we will predict the rate of default for both when the person is a student and when they are not.

```
predict.value <- tibble(balance = 1500,income = 40, student = c("Yes","No"))
predict.value
predict(blr3,predict.value,type="response")
```

What are the results?

- The only way to be absolutely sure that you have eliminated confounding variables is to randomly assign your subjects of your experiment to different treatments. Unfortunately, like in the loan example, often the data collected cannot be assigned randomly, you are stuck with what actually happened out in the world.

  Especially when dealing with social, economic, and racial issues, it is important to look for any confounding variables that might explain your results, as otherwise you can end up blaming a particular factor value for a result that is not actually causing it at all.

References

This lab was based on the following blog post.

UC Business Analytics R Programming Guide, Retrieved 15 April, 2018, http://uc-r.github.io/logistic_regression.

Chapter 45

Stats Lab: QQ/Box/Violin Plots

Instructions:

If you have time in the period, complete both the main and extended portions of the lab. If you run out of time, you do not have to complete the extended lab.

Main Lab

In this lab you will be learning several more ways of visualizing data, and how to visually check whether or not data comes from specific distributions, such as normal and gamma.

Air quality data We will make use of a built in data set in R that measures air quality in New York.

- We'll start by looking at data from the R built-in data set `airquality`. Take a look at this data set with `head(airquality)` and `tail(airquality)` to get an idea of what the data is like. You can see just from the first few lines that some of the row contain missing Ozone level data. So let's begin by checking out the ozone levels:

  ```
  ozone <- airquality$Ozone
  solar <- airquality$Solar.R
  length(ozone)
  ```

  What is the original length of the ozone vector?

- Some of the values in ozone vector are `NA`, which is R's way of telling us that the data was not recorded. Typing `is.na(ozone)` will tell you which of the entries have the `NA` designation. The `!` is logical not in R, so typing `!is.na(ozone)` will tell you which of the entries do not have the `NA` designation. This data shouldnt't contribute to the vector, so let's only keep the non-NA entries.

  ```
  ozone <- ozone[!is.na(ozone)]
  n <- length(ozone)
  ```

  What is the new length of the ozone vector without the missing values?

- Now let's find the population mean, variance, and standard deviation:

  ```
  mu <- mean(ozone)
  sigmasq <- var(ozone)
  sigma <- sd(ozone)
  ```

What is the mean of the ozone values?

- Suppose that we hadn't removed the `NA` values from earlier. What if we left them in? Try it with

```
mean(airquality$Ozone)
```

What is the result?

- We can adjust the `mean` function so that it automatically strips out `NA` values and gives the correct mean. Try `mean(airquality$Ozone,na.rm=TRUE)`. What is the result?

Box Plots

- One way to get an idea of the middle and spread of the data is through a visual tool known as a *box plot*, or more generally a *box-and-whisker plot*. Try the following in R:

```
boxplot(ozone)
```

Sketch this plot.

- There's a lot to unpack here. The thick line in the middle of the box is the median line, you can find it's exact value using `median(ozone)`. The upper line of the "box", the rectangle in the middle is the 75% quantile line. That is, it is the value such that 75% of the data points are below that value. The bottom edge of the box is height equal to the 25% quantile line. You can find these numbers exactly in R with the `quantile` command:

```
quantile(ozone,c(0.25,0.5,0.75))
```

What are these quantile values?

- Call the value of the 75% quantile minus the 25% quantile the *Interquantile distance* or IQR.

  Now consider the "whiskers" or the dotted vertical lines leaving the box cappped by a shorter horizontal line. The top whisker lies at either the maximimum of the data, or at the 75% quantile plus 1.5 time the IQR, whichever is smaller. If there exist data points bigger than the 75% quantile plus 1.5 times the IQR, those points are represented by circle. The circles are unfilled if the points are at most the 75% quantile plus 3 times the IQR, otherwise they are filled in.

  How many data values are bigger than the 75% quantile plus 1.5 times the IQR?

- The bottom whisker is similar, and lies either at the minimum of the data, or at the 25% quantile minus 1.5 times the IQR, whichever is larger.

  How many data values are smaller than the 25% quantile minus 1.5 times the IQR?

QQ Plots

- Suppose that I have two data sets, and I want to know if they come from the same distribution. One way of visually inspecting this is to make a *quantile-quantile plot*, or qq plot for short. In these plots, the empirical quantiles of one data set are plotted against the empirical quantiles of the other data set.

  Another problem that arises frequently is to determine if a given data set comes from a fixed distribution (such as normal). Here the empirical quantiles of the data set are plotted against the theoretical quantiles of the distribution being considered. This is sometimes also called a qq plot, or sometimes a *probability plot*.

  To create this plot, start by assuming that each of the ozone levels is an independent indentically distributed draw from the ozone distribution, then the ith order statistic of the ozone levels will be are about the inverse of the cdf at $i/(n+1)$. For example, in the picture below, the four data points break the real line into five segments. The location of the second order statistic out of four data points will be near the location where the cdf of the distribution is $2/5$.

  So create a vector of probability values using:

```
probabilities <- (1:n)/(n+1)
```

  What are the first few values of `probabilities`?

- Now the vector of ozone values could be modeled in many different ways. Suppose we wanted to know if the distribution was X. Then the first order statistic should be near $\text{cdf}_X^{-1}(0.008547)$, the second order statistic should be near $\text{cdf}_X^{-1}(0.017094)$, and so on. These values, for a normal with mean `mu` and standard devation `sigma`, can be found using:

```
normal.quantiles <- qnorm(probabilities, mean=mu,sd=sigma)
```

  Now if the ozone values actually came from a normal distribution with this mean and standard deviation, then the sorted values of the ozone and the `normal.quantiles` should be approximately the same. So plotting one against the other should give us a straight line. Test this visually using:

```
plot(normal.quantiles,sort(ozone),
  xlab='Theoretical Quantiles from Normal',
  ylab='Sample Quantiles of Ozone',
  main='Quantile-Quantile plot')
abline(0,1,col='gold',lwd=3)
```

  Does the qqplot fit the line?

- The fact that the qq plot curves upward instead of making a straight line indicates the the data is right skewed. A right-skewed data set has a long tail on the right. An example of such a distribution is the gamma distribution. Now, the mean of a gamma with shape parameter α and scale parameter β is $\alpha\beta$, and the variance is $\alpha\beta^2$. So solve

$$\hat{\mu} = \alpha\beta,\ \hat{\sigma}^2 = \alpha\beta^2$$

  for α and β in terms of $\hat{\mu}$ and $\hat{\sigma}^2$.

- To implement the solution I used, try

```
gamma.quantiles <- qgamma(probabilities,shape=mu^2/sigmasq,
 scale=sigmasq/mu)
```

  Now create a qqplot similar to the one made for normal quantiles, and sketch the results.

- That looks much better! Just because it fits a line doesn't guarantee that the data comes from the gamma distribution, but it does provide strong evidence that using gamma to fit the data is a better idea than using a normal fit.

  Of course, R has its own built in functions for creating qqplots. Unlike our plot that we created earlier, R doesn't initially scale the units, so the line will not have slope one. Try

```
qqnorm(ozone)
```

  to create an initial qqplot against the normal density. To add a fit line to it, try

```
qqline(ozone)
```

  Finally, add

```
abline(mean(ozone),sd(ozone),col='gold',lwd=3)
```

  to get a line close to the R line.

  The default for qqline is a normal distribution, but can alter the density used with the `distribution` parameter. First we create α and β, then use `qqline` to make the plot

```
alpha <- mu^2/sigmasq
beta <- sigmasq/mu
qqline(ozone,distribution=function(s) qgamma(s,shape=alpha,scale=beta))
```

- Last, let's try testing data that actually comes from the normal distribution using a qq-plot. Try

```
qqnorm(rnorm(100));abline(0,1,col='gold',lwd=3)
```

  Try this command a few times to get an idea of how the results vary. Note that the fit is far from perfect; so even when you start with perfectly normal data, the qq-plot will not fit the line perfectly, especially in the tails where probabilities are low.

- Now try it again with a distribution where the tails are much wider than a normal, a Cauchy distribution:

```
qqplot(rcauchy(100),qcauchy((1:100)/101));abline(0,1,col='gold',lwd=3)
```

  The line should fit well in the middle, but the tails are a bad loss. So in analyzing a qq plot, look first at the middle values near the center, when those don't fit well (like in our initial fit of the ozone to a normal), that's when you want to try a different distribution.

Extended Lab

Violin plots A violin plot is similar to a box plot, but has a density kernel plot rotated and placed on each side.

- Try the following (as always, before using a library package, if it is not already installed then can install the package first with `install.packages("vioplot")`:

  ```
  library(vioplot)
  vioplot(ozone)
  ```

  Sketch the resulting plot.

- So why use violin plots rather than box plots? In short, they are more informative. A violin plot (kernel density) can show when the data is multimodel, but a box plot cannot.

Multiple Plots Box and Violin plots can give a good way of comparing multiple levels of a factor against each other visually. For instance, there is a built in data set for R called `InsectSprays`.

- Typing `InsectSprays` into Rwill show you the data, which consists of two columns. Count is the number of insect bites in a given time, while spray tells you the type of spray, either A, B, C, D, E, or F.

  Our goal is to get five box plots, one for each type of spray. To do that we want to model the counts versus the sprays. So our model is `count ~ spray`. Let's put that in a box plot with

  ```
  boxplot(count ~ spray, data = InsectSprays)
  ```

  Sketch your result

- Unfortunately, the command `vioplot` can't handle the model format that boxplot can. So let's first separate out the data we want, then plot it.

  ```
  ca <- InsectSprays$count[InsectSprays$spray=='A']
  cb <- InsectSprays$count[InsectSprays$spray=='B']
  cc <- InsectSprays$count[InsectSprays$spray=='C']
  vioplot(ca,cb,cc)
  ```

  Which spray would you pick from A, B, C?

Part IV

Problem Solutions

Chapter 46

Worked problems

1.1: Go to the website `www.wolframalpha.com` and type in

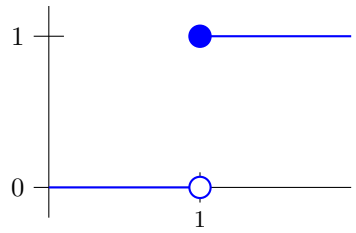
```
sum(1/2)^i for i from 1 to infinity
```

What is $\sum_{i=1}^{\infty}(1/2)^i$?

Solution From the website: $\boxed{1}$

1.2: Graph $f(s) = \mathbb{1}(s \geq 0)$

Solution This graph looks like



1.3: Solve $\int_{-\infty}^{\infty} 2s\mathbb{1}(s \in [0,1])\, ds$

Solution First use the indicator function to change the limits of integration, then use the power rule for antidifferentiation and the Fundamental Theorem of Calculus.

$$\begin{aligned}\int_{-\infty}^{\infty} 2s\mathbb{1}(s \in [0,1])\, ds &= \int_0^1 2s\, ds \\ &= s^2|_0^1 = \boxed{1}.\end{aligned}$$

1.4: What is $\sqrt{\tau}$?

Solution To four significant digits, this is about $\sqrt{6.283185307} = \boxed{2.506}$.

2.1: Let X have density $f_X(1) = 0.2$, $f_X(5) = 0.7$, and $f_X(6) = 0.1$.

(a) What is $\mathbb{P}(X = 5)$?

(b) What is $\mathbb{P}(X = 2)$?

(c) What is $\mathbb{E}[X]$?

(d) What is $\mathbb{V}(X)$?

Solution

(a) This is given by the density of X at 5, or $\boxed{0.7000}$.

(b) This is given by the density of X at 2, which by default unless otherwise stated is $\boxed{0}$.

(c) Because $\mathbb{P}(X \in \{1,5,6\}) = 1$, this can be found as a sum:

$$\mathbb{E}[X] = (0.2)(1) + (0.7)(5) + (0.1)(6) = \boxed{4.300}$$

(d) The variance of a random variable is $\mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. We have $\mathbb{E}[X] = 4.3$ from the last part. To find $\mathbb{E}[X^2]$, just apply the square function to the values that X takes on.

$$\mathbb{E}[X^2] = (0.2)(1)^2 + (0.7)(5)^2 + (0.1)(6)^2 = \boxed{21.3}.$$

This makes

$$\mathbb{V}(X) = 21.3 - 4.3^2 = \boxed{2.810}.$$

2.2: Let X have density $f_X(i) = (2/3)^{i-1}(1/3)\mathbb{1}(i \in \{1,2,\ldots\})$ with respect to counting measure.

(a) Find $\mathbb{P}(X \in \{1,2,3\})$.

(b) Find $\mathbb{E}(X)$.

Solution

(a) Let ν be counting measure, then this is

$$\begin{aligned}\mathbb{P}(X \in \{1,2,3\}) &= \int_{a \in \{1,2,3\}} (2/3)^{a-1}(1/3)\ d\nu \\ &= \sum_{a \in \{1,2,3\}} (2/3)^{a-1}(1/3) = (1/3)[1 + (2/3) + (4/9)] \\ &= 19/27 \approx \boxed{0.7037}.\end{aligned}$$

(b) Then

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} i(2/3)^{i-1}(1/3).$$

Putting `sum i*(2/3)^(i-1)*1/3 for i from 1 to infinity` into Wolfram Alpha then gives $\boxed{3}$.

2.3: Let T have density $f_T(s) = 2\exp(-2s)\mathbb{1}(s \geq 0)$.

(a) Find $\mathbb{P}(X \in [1,3])$.

(b) Find $\mathbb{E}[X]$.

Solution Both of these problems are about setting up the proper integrals.

(a)

$$\mathbb{P}(X \in [1,3]) = \int_{s \in [1,3]} 2\exp(-2s)\mathbb{1}(s \geq 0)\ ds.$$

Notice that since $s \in [1,3]$ already in the integral, the indicator function $\mathbb{1}(s \geq 0)$ always evaluates to 1, and disappears.

$$\begin{aligned}\mathbb{P}(X \in [1,3]) &= \int_{s \in [1,3]} 2\exp(-2s)\ ds \\ &= 2\exp(-2s)/(-2)|_1^3 = 2[\exp(-2) - \exp(-6)] \approx \boxed{0.2657}.\end{aligned}$$

(b) For this integral, we need integration by parts, which you might recall allows us to "slide" a derivative over from one factor to another:

$$\int_A f(x)g'(x)\ dx = \int_A [f(x)g(x)]' - f'(x)g(x)\ dx.$$

In our case

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} s[2\exp(-2s)]\mathbb{1}(s \geq 0)\, ds = \int_{0}^{\infty} s[2\exp(-2s)]\, ds \\
&= \int_{0}^{\infty} s[2\exp(-2s)/(-2)]'\, ds \\
&= \int_{0}^{\infty} s[-\exp(-2s)]'\, ds \\
&= \int_{0}^{\infty} [-s\exp(-2s)]' - [s]'[-\exp(-2s)]\, ds \\
&= \int_{0}^{\infty} [-s\exp(-2s)]' + \exp(-2s)\, ds \\
&= \exp(-2s)/(-2) + s\exp(-2s)|_{0}^{\infty} \\
&= 1/2 = \boxed{0.5000}.
\end{aligned}$$

3.1: For $X \sim \mathsf{Unif}([3,4])$ find

(a) $\mathbb{E}[X]$.

(b) $\mathbb{V}(X)$.

Solution

(a) The mean of X is $(3+4)/2 = \boxed{3.500}$.

(b) The variance of X is $(4-3)^2/12 = 1/12 \approx \boxed{0.08333}$.

3.2: Suppose that I have 10 subjects in an experiment. For each subject, either a drug is effective in lowering blood sugar or it is not. Assuming that the probability the drug is effective is 0.3, and that each subject behaves independently from the rest, what is the distribution of N, the number of subjects where the drug was effective?

Solution Each subject is either a success (counts as 1) or a failure (counts as 0). Adding up this 1 or 0 for each subject gives the total number of successes. Since adding Bernoulli random variables gives a binomial, we have that N is $\boxed{\mathsf{Bin}(10, 0.3)}$.

4.1: Suppose Y is equally likely to be 1, 2, or 3. Let X_1, X_2, X_3 be independent draws of a random variable with density $f(1) = 0.3$, $f(2) = 0.3$, $f(3) = 0.4$ with respect to counting measure.

(a) What is $\mathbb{E}[X_i]$?

(b) What is

$$\mathbb{E}\left[\sum_{i=1}^{Y} X_i\right]?$$

Solution

(a) Since the X_i are discrete:

$$\mathbb{E}[X_i] = \sum_{i \in \{1,2,3\}} i f_X(i) = 0.3(1) + 0.3(2) + 0.4(3) = \boxed{2.100}.$$

5.1: True or false: If $\max_\theta f(\theta)$ exists for $f(\theta) \geq 0$, then $\max_\theta f(\theta) = \max_\theta \ln(f(\theta))$.

Solution $\boxed{\text{False.}}$ Because natural log (ln) is a strictly increasing function, $\arg\max_\theta f(\theta) = \arg\max_\theta \ln(f(\theta))$, but so the place where the maximum occurs is unchanged. But the maximum value itself will be different.

5.2: Find $\arg\max \exp(-(x-4)^2/2)$.

Solution Note that

$$\begin{aligned}\arg\max \exp(-(x-4)^2/2) &= \arg\max \ln(\exp(-(x-4)^2/2)) \\ &= \arg\max -(x-4)^2/2.\end{aligned}$$

Since anything squared is nonnegative,

$$(x-4)^2 \geq 0 \Rightarrow -(x-4)^2/2 \leq 0,$$

and $x = 4$ gives 0, so $\boxed{x=4}$ is the argument maximizer.

5.3: Find $\arg\max_{\lambda>0} \lambda^3 \exp(-2.1\lambda)$

Solution It is easier to work with the log of the function:

$$\begin{aligned}\arg\max_{\lambda>0} \lambda^3 \exp(-2.1\lambda) &= \arg\max_{\lambda>0} \ln(\lambda^3 \exp(-2.1\lambda)) \\ &= \arg\max_{\lambda>0} [3\ln(\lambda) - 2.1\lambda].\end{aligned}$$

Since $[3\ln(\lambda) - 2.1\lambda]' = 3/\lambda - 2.1$, and

$$\begin{aligned}3/\lambda - 2.1 \geq 0 \Leftrightarrow \lambda \leq 2.1/3 \\ 3/\lambda - 2.1 \leq 0 \Leftrightarrow \lambda \geq 2.1/3.\end{aligned}$$

Hence the function is increasing over $[0, 0.7]$ and decreasing over $[0.7, \infty)$. Therefore the argument maximum must be at $\boxed{0.7000}$.

6.1: True or false: if an experimenter is careful, they will always get the same result for their data.

Solution $\boxed{\text{False.}}$ No matter how careful an experimenter is, random effects outside of their control can change the data collected. That is why we use probabilistic models in statistics so often.

6.2: Fill in the blank: For data $X_1, X_2, \ldots$, $(X_1 + \cdots + X_{15})/15$ and $\max_i X_i$ are examples of _____.

Solution $\boxed{\text{Statistics.}}$ In general, any function of the data is a statistic of the data.

7.1: Suppose that $X_1, \ldots, X_n$ given θ are iid $\mathsf{Unif}([0, \theta])$. Find the Method of Moments estimate of θ.

Solution Each X_i has mean $\theta/2$. Hence we set

$$\bar{X}_i = \frac{\hat{\theta}_{\text{MOM}}}{2},$$

and solve to obtain $\boxed{\hat{\theta}_{\text{MOM}} = 2\bar{X}_i.}$

7.2: Suppose I model X given θ as being $\mathsf{Unif}([\theta, 2\theta])$. Say $X_1, \ldots, X_n$ are iid draws from X.

(a) What is the likelihood function $L_{x_1,\ldots,x_n}(\theta)$ given $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$?

(b) Derive the MLE for θ given data $x_1, \ldots, x_n$.

(c) Evaluate your MLE at data $1.3, 2.1, 1.7$.

(d) Derive the MOM for θ given data $x_1, \ldots, x_n$.

(e) Evaluate your MOM at data $1.3, 2.1, 1.7$.

Solution

(a) The likelihood is the density of statistical model viewed as a function of the parameter θ. Since the data is independent, the joint density is the product of the individual densities. So

$$L(\theta) = \prod_{i=1}^{n} f_{X_i}(x_i) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}(x_i \in [\theta, 2\theta])$$
$$= \boxed{\theta^{-n} \left[\prod_{i=1}^{n} \mathbb{1}(x_i \in [\theta, 2\theta]) \right]}.$$

(b) The function θ^{-n} decreases as θ increase. Therefore to maximize the function we make θ as small as possible. But $x_i \leq 2\theta$ for all x_i, so the smallest choice of θ is

$$\boxed{\theta_{\text{MLE}} = (1/2) \max_i x_i}.$$

(c) For this data $(1/2) \max x_i = (1/2)(2.1) = \boxed{1.050}$.

(d) The average is $\mathbb{E}[X|\theta] = (2\theta + \theta)/2 = 1.5\theta$. Hence the MOM satisfies $1.5\hat{\theta}_{\text{MOM}} = \bar{x}$, so

$$\boxed{\hat{\theta}_{\text{MOM}} = \bar{x}/1.5}.$$

(e) For this data, that is $[(1.3 + 2.1 + 1.7)/3]/1.5 \approx \boxed{1.133}$.

8.1: Given data $(1.7, 1.6, 2.4, 3.1)$,

(a) Give an unbiased estimate of the mean of the distribution.

(b) Give an unbiased estimate of the variance of the distribution.

Solution

(a) This is $(1.7 + 1.6 + 2.4 + 3.1)/4 = \boxed{2.200}$.

(b) This is

$$\frac{(1.7 - 2.2)^2 + (1.6 - 2.2)^2 + (2.4 - 2.2)^2 + (3.1 - 2.2)^2}{4 - 1} \approx \boxed{0.4866}.$$

9.1: True or false: The maximum likelihood estimator is always unbiased.

Solution $\boxed{\text{False.}}$ For instance, if $[X_1, \ldots, X_n|\theta] \sim \mathsf{Unif}([0, \theta]^n)$, then $\hat{\theta}_{\text{MLE}} = \max_i X_i$, which has mean $[n/(n+1)]\theta < \theta$.

9.2: Suppose that an experimenter runs a sequence of trials that are each independently a success with parameter p.

(a) Let T be the number of trials needed for one success. So if the sequence was fail, fail, success, then $T = 3$. Find the MLE of p as a function of T.

(b) Find the Method of Moments estimate of p as a function of T.

Solution

(a) In order for $T = i$, there must be $i - 1$ failures and one success. (T is a geometric random variable with mean $1/p$.) So the density of T is

$$f_T(i) = (1 - p)^{i-1} p$$

So $L_T(p) = (1 - p)^{T-1} p$, which means $\ln(L_T(p)) = (T - 1) \ln(1 - p) + \ln(p)$, $[\ln(L_T(p))]' = -(T - 1)/(1 - p) + 1/p$, and $[\ln(L_T(p))]'' = -(T - 1)[1/(1 - p)^2 + 1/p^2]$.

The second derivative is nonpositive, which means there is a unique maximum value at the value of p where $[\ln(L_T(p))]' = 0$, which is

$$\begin{aligned} -(T-1)/(1-\hat{p}_{\text{MLE}}) + 1/\hat{p}_{\text{MLE}} &= 0 \\ (1-\hat{p}_{\text{MLE}})/\hat{p}_{\text{MLE}} &= T-1 \\ (1-p_{\text{MLE}}) &= (T-1)p_{\text{MLE}} \\ p_{\text{MLE}} &= \boxed{1/\text{T}}. \end{aligned}$$

(b) The expected value of a geometric random variable is $1/T$. Hence we set $1/\hat{p}_{\text{MOM}} = T$, which makes $\hat{p}_{\text{MOM}} = \boxed{1/T}$.

10.1: Fill in the blank: A Beta prior and Binomial likelihood gives an example of _____ priors.

Solution Conjugate.

10.2: A rate of typos in a series of plays by an author is modeled as having a prior $\mu \sim \mathsf{Exp}(0.1)$, so $f_\mu(s) = 0.1\exp(-0.1s)\mathbb{1}(s \geq 0)$. Given μ, the number of typos found in a given play is modeled as Poisson distributed with mean μ, so if T denotes the number of typos, for $i \in \{0, 1, 2, \ldots\}$

$$\mathbb{P}(T = i | \mu = s) = \frac{\exp(-s)s^i}{i!}.$$

(a) What is the posterior distribution of μ given T?

(b) If $T = 5$, what is the posterior mean?

Solution

(a) The density of the posterior is proportional to the the density of the prior times the density of the likelihood. That is,

$$f_{\mu|T=i}(s) = C f_\mu(s) f_{T|\mu=m}(i).$$

In this case,

$$\begin{aligned} f_{\mu|T=i}(s) &= C(0.1)\exp(-0.1s)\mathbb{1}(s \geq 0)\frac{\exp(-s)s^i}{i!} \\ &= C'\exp(-1.1s)s^i\mathbb{1}(s \geq 0). \end{aligned}$$

This is the density of a $\boxed{\text{gamma distribution with shape parameter } T \text{ and rate } 1.1.}$

(b) The mean of this will be $T/1.1 \approx \boxed{4.545}$.

10.3: Suppose I have statistical model $[X|\theta] \sim \mathsf{Exp}(\lambda)$, and a prior on λ of $\lambda \sim \mathsf{Unif}([1, 3])$.

(a) Find the density

$$f_{\lambda|X_1,\ldots,X_n=x_1,\ldots,x_n}(t)$$

of the posterior up to an unknown normalizing constant.

(b) For data $1.3, 2.1, 1.7$, what is the posterior mode?

(c) For general data $x_1, \ldots, x_n$, what is the posterior mode?

Solution

(a) The posterior is proportional to the prior times the likelihood:

$$\begin{aligned} f_{\lambda|X_1,\ldots,X_n=x_1,\ldots,x_n}(t) &\propto f_\lambda(t) f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) \\ &= \frac{1}{3-1}\mathbb{1}(t \in [1,3])\prod_{i=1}^n t\exp(-tx_i)\prod_i \mathbb{1}(x_i \geq 0) \\ &\propto t^n\exp(-t(x_1+\cdots+x_n))\mathbb{1}(t \in [1,3]). \end{aligned}$$

Hence

$$\boxed{f_{\lambda|X_1,\ldots,X_n=x_1,\ldots,x_n}(t) = Ct^n \exp(-t(x_1+\cdots+x_n))\mathbb{1}(t\in[1,3]).}$$

where C is an unknown normalizing constant.

(b) Note

$$\begin{aligned}[f_{\lambda|X_1,\ldots,X_n=x_1,\ldots,x_n}(t)]' &= C\exp(-t(x_1+\cdots+x_n))[nt^{n-1} - t^n[x_1+\cdots+x_n]]\mathbb{1}(t\in[1,3]) \\ &= C\exp(-t(x_1+\cdots+x_n))t^{n-1}[n - t(x_1+\cdots+x_n)]\mathbb{1}(t\in[1,3]).\end{aligned}$$

For $t\in[1,3]$, this derivative is nonnegative if and only if $n-t(x_1+\cdots+x_n)\geq 0$ which is equivalent to

$$t \leq \frac{n}{x_1+\cdots+x_n}.$$

The derivative is nonpostive if and only if $n-t(x_1+\cdots+x_n)\leq 0$ which is equivalent to

$$t \geq \frac{n}{x_1+\cdots+x_n}.$$

So for data $1.3, 2.1, 1.7$, $3/(1.3+2.1+1.7) = 0.5882\ldots$, so the derivative is nonpositive for all $t\in[1,3]$. Hence the mode occurs at $\boxed{t=1.}$

(c) For general data, the only extra thing to remember is the $t\in[1,3]$ for the indicator function to be 1. Therefore, the posterior mode is

$$\boxed{\begin{array}{ll} \bar{x} & \text{if } \bar{x}\in[1,3] \\ 3 & \text{if } \bar{x}>3 \\ 1 & \text{if } \bar{x}<1 \end{array}}$$

11.1: Suppose that $X_1, X_2, \ldots, X_{10} \overset{\text{iid}}{\sim} X$, where $[X|\theta]\sim \mathsf{Unif}([0,\theta])$. What is

$$\mathbb{P}(2\min_i X_i \leq \theta \leq 2\max_i X_i)?$$

Solution The only way that $2\min_i X_i > \theta$ is if all the X_i are greater than $\theta/2$. This happens with probability $(1/2)^{10}$. Similarly, the chance that $2\max_i X_i < \theta$ is also $(1/2)^{10}$. The event that we want $\theta\in[2\min X_i, 2\max X_i]$ is the complement of the union of these events, therefore,

$$\mathbb{P}(2\min_i X_i \leq \theta \leq 2\max_i X_i) = 1-(1/2)^9 \approx \boxed{0.9980}.$$

This means that $[2\min X_i, 2\max X_i]$ is a 99.80% level confidence interval for θ for 10 samples.

11.2: Dr. Pamela Isley measures the height of four plant samples, and finds them to be (in centimeters)

$$4.5, 3.7, 1.2, 6.2.$$

(a) Give an unbiased estimate of the mean height of the plants (including units).

(b) Give an unbiased estimate of the variance of the height of the plants (including units).

(c) Give a 90% z-value confidence interval for the mean plant height, using $\Phi(0.95) = 1.644854$.

Solution

(a) An unbiased estimate is the sample average, which yields

$$\frac{1}{n}\sum_{i=1}^n X_i = \boxed{3.900 \text{ cm}}.$$

(b) An unbiased estimate of the variance is the population sample variance

$$\frac{1}{n-1}\sum_{i=1}^n (X_i-\bar{X})^2 = \boxed{4.326 \text{ cm}^2}.$$

(c) Using our pivot for normal random variables, the interval is $[\hat{\mu}-1.644854\hat{\sigma}/\sqrt{4}, \hat{\mu}+1.644854\hat{\sigma}/\sqrt{4}]$. Numerically, this gives

$$\boxed{[2.189 \text{ cm}, 5.611 \text{ cm}]}$$

11.3: Let $X_1, \ldots, X_n$ be modeled as iid draws from the uniform distribution on $[\theta, \theta+1]$.

(a) What is the distribution of $X_i - \theta$? [You do not have to prove the result, simply give the distribution.]

(b) Show that $W = \bar{X} - \theta$ is a pivot.

Solution

(a) Shifting a uniform random variable by θ reduces the endpoints of the interval by θ, so $\mathsf{Unif}([0,1])$.

(b) Note

$$\bar{X} - \theta = \frac{X_1 + \cdots + X_n}{n} - \theta = \frac{(X_1 - \theta) + \cdots + (X_n - \theta)}{n}.$$

For the previous part, we know that $X_i - \theta \sim U_i$, so

$$\bar{X} - \theta \sim \frac{U_1 + \cdots + U_n}{n},$$

where the U_i are iid $\mathsf{Unif}([0,1])$. So the distribution of $W = \bar{X} - \theta$ does not depend on θ in any way, which makes the random variable a pivot.

12.1: Suppose $X_1, \ldots, X_{10}$ are modeled as normal random variables with unknown mean μ and variance σ^2. What is the chance that the relative error in $\hat{\sigma}^2$ is greater than 10%? In other words, what is $\mathbb{P}(\hat{\sigma}^2 \geq 1.1\sigma)$?

Solution Note that

$$\mathbb{P}(\hat{\sigma}^2 \geq 1.1\sigma) = \mathbb{P}(9 \cdot \hat{\sigma}^2/\sigma^2 \geq 9.9),$$

and since $(n-1)\hat{\sigma}^2/\sigma^2 = \chi^2(n-1)$, that means

$$\mathbb{P}(\hat{\sigma}^2 \geq 1.1\sigma) = \mathbb{P}(C \geq 9.9),$$

where $C \sim \chi^2(9)$. Using `1-pchisq(9.9,df=9)` then gives $\boxed{0.3586}$.

13.1: Suppose a drug works with a probability p that is modeled as $\mathsf{Beta}(1,9)$.

(a) What is the prior mean that the drug works?

(b) Suppose that 40 independent trials are run, in 13 of which the drug is a success. What is the posterior distribution of p given this data?

(c) Give a balanced two-tailed 95% credible interval for this data.

(d) Prove that your balanced interval is *not* the narrowest interval.

Solution

(a) For $p \sim \mathsf{Beta}(1,9)$, $\mathbb{E}[p] = 1/(1+9) = \boxed{0.1000}$.

(b) It seems reasonable to model the data X given p as $[X|p] \sim \mathsf{Bin}(40, p)$. Hence $[p|X = 13] \sim \mathsf{Beta}(1+13, 9+31) = \mathsf{Beta}(14, 40)$.

(c) Using `qbeta(0.025,14,40)` and `qbeta(0.975,14,40)`, to 4 sig figs the credible interval is

$$\boxed{[0.1525, 0.3829]}.$$

(d) When I evaluate the density of a $\mathsf{Beta}(14,40)$ at the two endpoints of the interval, I get

$$f(0.1525684) = 1.282862, \;\; f(0.3828247) = 0.8550493.$$

Since these two numbers are different, they cannot possibly form the narrowest interval.

14.1: For the distribution $\mathsf{Unif}([0,\theta])$, find the median as a function of θ.

Solution The density is $f(s) = [1/(\theta-0)]\mathbb{1}(s \in [0,\theta])$, so we wish to find m such that

$$\begin{aligned} 1/2 &= \int_{-\infty}^{m} f(s)\, ds \\ &= \int_{-\infty}^{m} \theta^{-1}\mathbb{1}(s \in [0,1])\, ds \\ &= \int_{0}^{m} \theta^{-1}\, ds \qquad \text{for } m \in [0,1] \\ &= m/\theta, \end{aligned}$$

hence the median is $\boxed{m = \theta/2}$.

14.2: (a) Find the sample median of $\{1.2, 7.6, 5.2\}$.

(b) Find the sample median of $\{3.4, 2.3, 7.3, 5.0\}$.

Solution

(a) Sort the values: $1.2 \le 5.2 \le 7.6$. Then the sample median is the middle value $\boxed{5.200}$.

(b) Again sort the values: $2.3 \le 3.4 \le 5.0 \le 7.3$. Since there is no middle value, average the two values surrounding the middle to get $(3.4+5.0)/2 = \boxed{4.200}$.

15.1: Fill in the blank: $Y = X\beta + \epsilon$ where X is an m by k matrix, β is a k by 1 column vector, is a _____ model.

Solution $\boxed{\text{Linear.}}$ Here the mean of the predictor relates to the unknown coefficients β through multiplication by a matrix, and that makes this a linear model.

16.1: The form $Y = X\beta + \epsilon$ is what kind of model?

Solution This is a $\boxed{\text{linear model.}}$

16.2: Consider some data from Old Faithful geyser showing the length of the eruption together with the waiting time until the next eruption (both measured in minutes.)

3.600	79
1.800	54
3.333	74
2.283	62
4.533	85
2.883	55

We wish to fit a model where the waiting times y_i are predicted by the eruption length x_i using constant, linear, and quadratic terms. So

$$y_i = c_0 + c_1 x_i + c_2 x_i^2 + \epsilon_i$$

(a) What is the vector Y in $Y = X\beta + \epsilon$?

(b) What is the matrix X in $Y = X\beta + \epsilon$?

(c) Using numerical software, find the pseudoinverse of X.

(d) What is the vector β in $Y = X\beta + \epsilon$?

(e) What is the maximum likelihood estimate $\hat{\beta}$ for β?

(f) What is the estimate of the residuals $Y - X\hat{\beta}$?

Solution

(a) The vector Y is the values we wish to come close to with our model, so

$$\boxed{\begin{pmatrix} 79 \\ 54 \\ 74 \\ 62 \\ 85 \\ 55 \end{pmatrix}}$$

(b) Because we are using three predictors (a constant, a linear term, and a quadratic term) there will be three columns of X. The first column is just all 1's, corresponding to the constant term. The second column is the x_i values, corresponding to the linear term. The third column is the x_i^2 values, corresponding to the quadratic term. That makes the X matrix (to four significant figures):

$$\boxed{\begin{pmatrix} 1 & 3.600 & 12.96 \\ 1 & 1.800 & 3.240 \\ 1 & 3.333 & 11.10 \\ 1 & 2.283 & 5.212 \\ 1 & 4.533 & 20.54 \\ 1 & 2.883 & 8.311 \end{pmatrix}}$$

(c) The pseudoinverse of X is

$$(X^T X)^{-1} X^T = \boxed{\begin{pmatrix} -1.615 & 3.458 & -1.854 & 0.5555 & 1.932 & -1.476 \\ 1.106 & -1.974 & 1.335 & -0.08251 & -1.564 & 1.179 \\ -0.1581 & 0.2710 & -0.2033 & -0.01323 & 0.2972 & -0.1936 \end{pmatrix}}$$

(d) The vector β is just our constants in the model:

$$\boxed{\beta = \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix}}$$

(e) We get the maximum likelihood estimate $\hat{\beta}$ by multiplying the pseudoinverse times the Y values to get:

$$\boxed{\hat{\beta} = \begin{pmatrix} 39.39 \\ 6.383 \\ 0.8958 \end{pmatrix}}$$

(f) This model leaves residuals of

$$\boxed{\hat{\epsilon} = \begin{pmatrix} 5.017 \\ 0.2155 \\ 3.379 \\ 3.365 \\ -1.736 \\ -10.24 \end{pmatrix}}$$

17.1: A hypothesis containing only a single parameter value is called what?

Solution Simple.

17.2: Suppose that $T(X) \in R$ where X is our data, T is our test statistic and R is our rejection region. What does that mean for the null hypothesis?

Solution This means that we reject the null hypothesis.

17.3: True or false: t statistics under the null hypothesis have a t distribution.

Solution $\boxed{\text{True.}}$ This is a case where the name of the statistic and the name of the distribution it comes from if the null hypothesis is true are the same.

17.4: Say if the following hypothesis are simple or compound.

(a) $H_0 : \mu = 0$.

(b) $H_0 : \mu < 0$.

(c) $H_a : \mu \geq 0$.

(d) $H_0 : \mu \in \{0, 1\}$.

Solution

(a) This is the classic $\boxed{\text{simple}}$ null hypothesis that the mean does not change.

(b) This is $\boxed{\text{compound}}$ because H_0 has more than one possibility.

(c) There are still more than one value of μ that fits the condition so $\boxed{\text{compound}}$.

(d) There are not an infinite number of values of μ acceptable here, but even just two makes this a $\boxed{\text{compound}}$ hypothesis.

17.5: Suppose that a group of students is trying to assess whether or not the mean price of textbooks has risen more than \$20 in the past five years. Let μ_{-5} be the mean price of textbooks 5 years ago, and μ_0 be the current price.

(a) State the null hypothesis in terms of the μ_i.

(b) State the alternate hypothesis in terms of the μ_i.

Solution

(a) The null is typically that the thing that you are testing for did not happen, so $\boxed{H_0 : mu_0 < \mu_{-5} + 20}$ in this case.

(b) Then the alternate includes all other possibilities, so

17.6: A researcher is considering the effects of childhood income on graduation from college. Let μ_0 be the mean graduation rate for children born in poverty, and μ_1 be the mean graduation rate for children not born in poverty.

(a) State the null hypothesis.

(b) If the researchers only cared that being not born into poverty increased the college graduation rate, state the alternative.

(c) If the researchers only care that being not born into poverty increased the college graduation rate by at least 50%, state the alternative.

(d)

Solution

(a) $\boxed{H_0 : \mu_0 = \mu_1}$

(b) $\boxed{H_a : \mu_1 > \mu_0}$

(c) $\boxed{H_A : \mu_1 \geq 1.5\mu_0}$

18.1: Rejecting the null when the null is true is what type of error?

Solution This is a $\boxed{\text{Type I}}$ error.

18.2: Fill in the blank: _____ is usually used to represent an upper bound on Type II error.

Solution This is $\boxed{\beta}$.

18.3: True or false: The power of a test plus the chance of Type II error must add to 1.

Solution $\boxed{\text{True.}}$ In fact, this is how the power is defined: it is exactly one minus the probability of Type II error.

18.4: True or false: We want Type II error to be as low as possible.

Solution $\boxed{\text{True.}}$ Unfortunately, as Jagger and Richards (1969) have pointed out, you can't always get what you want. It is often not possible to make both Type I and Type II error smaller simultaneously.

18.5: When deciding which is the null and which is the alternate, the hypothesis that an intervention does not change the mean is typically which hypothesis?

Solution The hypothesis that no change occurs with an intervation is usually taken to be the $\boxed{\text{null}}$ hypothesis.

19.1: Under the null hypothesis, the chance that a p-statistic is in $[0.3, 0.34]$ is what?

Solution Under the null hypothesis, the p-statistic is $\mathsf{Unif}([0,1])$, so this is $0.4 - 0.3 = \boxed{0.1000}$.

20.1: True or false: Likelihood ratio tests require two possible hypotheses.

Solution $\boxed{\text{True.}}$ You cannot form the ratio of the likelihood under the null and the likelihood under the alternate if you do not have a null and alternate hypothesis.

20.2: Suppose a research groups gathers a data that is summarized by a statistic X. The group forms a hypothesis that X comes from either density f_0 (the null), or it will come from density f_1 (the alternate).

Describe how you would construct a test for the collected dataset s of the null versus the alternate at the 5% significance level.

Solution Use a Neyman-Pearson likelihood ratio test. Before taking data, find the largest value of K such that

$$\mathbb{P}\left(\frac{f_0(X)}{f_1(X)} \leq K\right) \leq 0.05.$$

Then take the data $X = x$, and reject if $f_0(x)/f_1(x) \leq K$.

20.3: Suppose that a researcher models their summary statistic X as coming (null) from a beta with parameters 2 and 1 (so density $2s\mathbb{1}(s \in [0,1])$) or, alternatively, coming from a beta with parameters 3 and 1 (so density $3s^2\mathbb{1}(s \in [0,1])$.)

(a) Construct the uniformly most powerful test at the 5% for testing the null versus the alternate. Be sure to state any theorems that you are using.

(b) Evaluate your test at data $X = 0.8$. Would you reject the null at the 5% level?

Solution

(a) The uniformly most powerful test (by the Neyman-Pearson Lemma) is the likelihood ratio test. Here we need to find K such that before we take data,

$$\begin{aligned} 0.05 &= \mathbb{P}_{H_0}\left(\frac{f_0(X)}{f_1(X)} \leq K\right) \\ &= \mathbb{P}_{H_0}\left(\frac{2X}{3X^2} \leq K\right) \\ &= \mathbb{P}_{H_0}\left(K' \leq X\right). \end{aligned}$$

Now under H_0, the probability that $X \geq K'$ is

$$\int_{K'}^{1} 2s \; ds = 1 - K'^2 = 0.05 \Rightarrow K' = \sqrt{0.95}.$$

Hence the test is reject if $X \geq \sqrt{0.95}$, or since X is positive under the null, equivalently

$$\boxed{\text{Reject the null if } X^2 \geq 0.95.}$$

(b) Since $X^2 = 0.64 < 0.95$, we would $\boxed{\text{not reject the null}}$.

21.1: Suppose that $X_1, \ldots, X_n$ are iid $\mathsf{Unif}([0, \theta])$. Say $H_0 : \theta = 1$ and $H_a : \theta = 1.1$.

(a) Suppose the data drawn is $\{0.47, 0.76, 0.48\}$. Find the Bayes Factor for H_0 versus H_a.

(b) Suppose the data drawn is $\{0.47, 1.01, 0.76, 0.48\}$. Find the Bayes Factor for H_0 versus H_a.

(c) How much data would we need to take to guarantee a Bayes Factor that is either at least 10 or 0?

Solution

(a) The density under the null hypothesis is

$$f_{(X_1,X_2,X_3)|H_0}(x_1, x_2, x_3) = \prod_{i=1}^{3} \frac{1}{1-0} \mathbb{1}(x_i \in [0, 1]),$$

whereas under the alternate it is

$$f_{(X_1,X_2,X_3)|H_a}(x_1, x_2, x_3) = \prod_{i=1}^{3} \frac{1}{1.1-0} \mathbb{1}(x_i \in [0, 1]).$$

This makes the Bayes Factor:

$$\frac{1/1^3}{1/1.1^3} = \frac{1.1^3}{1^3} = \boxed{1.331}$$

(b) With the 1.01 data point, the numerator density becomes 0, and so the Bayes Factor is $\boxed{0}$.

(c) Generalizing from the earlier part, after taking n data points the Bayes Factor will either be 0, or 1.1^n. $1.1^n \geq 10$ means $n \ln(1.1) \geq \ln(10)$, so $n \geq \ln(10)/\ln(1/1) = 24.15....$ Since n is an integer, that means n is at least $\boxed{25}$.

22.1: Suppose that a drug used for decreasing anxiety is tested on ten patients that are randomly divided into two groups. One group $(X_1, \ldots, X_n \sim X)$ receives the drug, while the other group $(Y_1, \ldots, Y_m \sim Y)$ does not.

Each group initially started with 5 participants, but one of the drug receiving patients left the study part way through. Over the next month, the number of anxiety attacks are recorded, and found to be

patients	1	2	3	4	5
X_i	13	14	17	22	
Y_i	24	30	15	23	24

(a) What should the null and alternate hypothesis be if the company is interested in testing if the drug decreases anxiety?

(b) What is the Wilcoxon rank sum for the data?

(c) What is the average of the Wilcoxon statistic given that your null hypothesis is true?

(d) Write the calculation of the p-value for the Wilcoxon test as p is equal to the probability of an event.

(e) If $p \approx 0.032$, would you reject your null hypothesis at the 5% level?

Solution

(a) $\boxed{H_0 : X \sim Y,\ H_a : Y > X.}$

(b) Ordering them gives

$$13 < 14 < 17 < 18 < 21 < 22 < 24 < 28 < 30 < 31,$$

or $XXYXXYYYY$, so the $X_i's$ are in ranks $1, 2, 4, 5$, which sum to $\boxed{12}$.

(c) Each X_i on average has rank $(1 + 9)/2 = 5$, and there are 4 data points, so total of $\boxed{20}$.

(d) Let $(A_1, \ldots, A_4)$ be four numbers uniformly draw without replacement from $\{1, 2, \ldots, 9\}$. Then

$$\boxed{p = \mathbb{P}(A_1 + \cdots + A_4 \leq 12)}.$$

(e) $\boxed{\text{Yes}}$.

23.1: True or false: Fisher information is always nonnegative when it exists.

Solution $\boxed{\text{True.}}$ Fisher information (when it exists) is the mean of the square of something, which is always nonnegative.

23.2: Let $X \sim \mathsf{Gamma}(4, \lambda)$. Then X has density

$$f_{X|\lambda}(s) = \frac{\lambda^4}{6} s^3 \exp(-\lambda s) \mathbb{1}(s \geq 0).$$

This density is regular.

(a) What is the Fisher information in a single draw X about λ, $I_X(\lambda)$?

(b) What is the minimum variance of an unbiased estimator for λ? (Be sure to explain your answer.)

Solution

(a) First we need to calculate the score function

$$\begin{aligned} S(s) &= \frac{\partial \ln(f_{X|\lambda}(s))}{\partial \lambda} \\ &= \frac{\partial[4\ln(\lambda) + 3\ln(s) - \lambda s - \ln(6)]}{\partial \lambda} \\ &= \frac{4}{\lambda} - s. \end{aligned}$$

Next find $I_\lambda(X) = \mathbb{E}[S(X)^2]$:

$$\begin{aligned} I_\lambda(X) = \mathbb{E}[S(X)^2] &= \mathbb{E}\left[(4/\lambda)^2 - 8X/\lambda + X^2\right] \\ &= 16/\lambda^2 - (8/\lambda)(4/\lambda) + [4/\lambda^2 + (4/\lambda)^2] \\ &= \lambda^{-2}[16 - 32 + 4 + 16] = \boxed{4/\lambda^2}. \end{aligned}$$

(b) Therefore, by the Cramér Rao Theorem (since the density is regular), for any unbiased estimator $\hat{\lambda}$,

$$\mathbb{V}(\hat{\lambda}) \geq 1/I_X(\lambda) = \boxed{\lambda^2/4}.$$

24.1: Suppose $\langle x, y \rangle = 4$. What is $\langle 3x, -2y \rangle$?

Solution By the rules of inner products, you can pull out constants. So

$$\langle 3x, -2y \rangle = (3)(-2)\langle x, y \rangle = (3)(-2)(4) = \boxed{\text{-24}}.$$

24.2: Suppose $\text{Cov}(X, Y) = 4$. What is $\text{Cov}(3X, -2Y)$?

Solution Since covariance is an inner product, you can pull out constants. So

$$\text{Cov}(3X, -2Y) = (3)(-2)\,\text{Cov}(X, Y) = (3)(-2)(4) = \boxed{\text{-24}}.$$

24.3: Suppose that an unbiased estimator for parameter θ that uses data $x = (x_1, \ldots, x_n)$, has the form

$$\hat{\theta} = \theta^2 + \bar{x}/\theta.$$

Is the estimator efficient?

Solution $\boxed{\text{Yes}}$. Because the estimate has the form: function of θ plus a function of θ times a statistic of the data, it must be efficient.

25.1: Fill in the blank: A specific choice of level for every factor is called a _____.

Solution $\boxed{\text{Treatment.}}$

25.2: The first factor has two levels, the second factor has 3. How many total possible treatments are there?

Solution The choice is treatment is a choice from 2 levels for the first factor, and 3 for the second. This makes the number of treatments $2 \cdot 3 = \boxed{6}$.

25.3: An experiment for student performance places students into a group given a soda with no caffeine but with sugar, coffee with caffeine but no sugar, or tea with neither sugar nor cafeine. Their scores on the exam are

Soda:	88	93	93	88	93
Coffee:	89	88	79	94	100
Tea:	90	90	88	91	

(a) Find the overall averages of the scores on the exam.

(b) Find the averages for each of Soda, Coffee, and Tea.

(c) Find SS_B, SS_W, and SS_T.

(d) Verify that $SS_T = SS_w + SS_B$.

Solution

(a) The overall average is $\bar{s} = 90.28571$, or approximately $\boxed{90.28}$.

(b) The averages for each of the three drinks is

$$\boxed{\bar{s}_{\cdot 1} = 91,\ \bar{s}_{\cdot 2} = 90,\ \bar{s}_{\cdot 3} = 89.75.}$$

(Not much evidence that the drink does anything since these are so close both to each other and to the overall mean.)

(c) First SS_B:

$$SS_B = 5(91 - \bar{s})^2 + 5(90 - \bar{s})^2 + 4(89.75 - \bar{s})^2 \approx \boxed{4.017}$$

Next SS_W:

$$(88-91)^2+\cdots+(93-91)^2+(89-90)^2+\cdots+(100-90)^2+(90-89.75)^2+\cdots+(91-89.75)^2 \approx \boxed{276.7}.$$

Finally SS_T:

$$(88 - \bar{s})^2 + (93 - \bar{s})^2 + \cdots + (91 - \bar{s})^2 = \boxed{280.8}$$

(d) Check:

$$SS_B + SS_W = 4.107143 + 276.75 = 280.8571 = SS_T.$$

The equation holds!

26.1: What statistics are produced by a one factor ANOVA table?

Solution A one factor ANOVA table returns an $\boxed{\text{F}}$ statistic which can then be turned into a $\boxed{\text{p}}$ statistic if needed.

26.2: When using the F statistic, when do we reject the null hypothesis that the treatment leaves the mean effect unchanged?

Solution Typically we reject when the F statistic is $\boxed{\text{large.}}$

26.3: True or false: In an ANOVA table the F statistics must have an F distribution even if the null hypothesis is not true.

Solution $\boxed{\text{False.}}$ When the null hypothesis does not hold then all bets are off: the distribution of the F statistic could be nearly anything.

26.4: An experiment for student performance places students into a group given a soda with no caffeine but with sugar, coffee with caffeine but no sugar, or tea with neither sugar nor caffeine. Their scores on the exam are

Soda:	88	93	93	88	93
Coffee:	89	88	79	94	100
Tea:	90	90	88	91	

The team decides to do an ANOVA analysis.

(a) For this data set, fill out the following:

$$\begin{aligned} \text{Number of subjects} &= \\ \text{Number of factors} &= \\ \text{Number of treatments} &= \end{aligned}$$

(b) Your research partner starts filling out an ANOVA table. Fill out the rest.

	df	Sum Squares	Mean Squares	F-statistic
drink		4.107		
Residuals		276.750		

(c) Let $\text{cdf}_{F(a,b)}$ denote the cdf of an F distributed random variable. Write the p-statistic for this table using this function.

(d) Calculate the p-statistic.

(e) The ANOVA analysis requires a major assumption about the distribution of residuals. Name the assumption and define what the assumption means.

Solution

(a)

$$\begin{aligned} \text{Number of subjects} &= 14 \\ \text{Number of factors} &= 1 \\ \text{Number of treatments} &= 3 \end{aligned}$$

(b) The complete table looks like:

	df	Sum Squares	Mean Squares	F-statistic
drink		4.107		
Residuals		276.750		

(c) The p-statistic is $\mathbb{P}(F \geq 0.08162)$ where $F \sim F(2, 11)$. Hence it is $\boxed{1 - \text{cdf}_{F(2,11)}(0.0816)}$.

(d) In this case it comes out to be

(e) The assumption is $\boxed{\text{homscedasticity}}$, which means that the residuals all have the same variance.

26.5: A researcher wants to understand how much student belief affects exam scores. Before taking the exam, the students are made to watch a video that attempts to affect their confidence level. Some students watch an affirming video, others a discouraging video, and a third group a video which is neutral.

Their scores on the exam are

Boost:	8.8	9.2	8.1	9.5	
Discouraged:	9.6	4.5	6.0	7.1	
Neutral:	8.1	7.9	8.0	5.2	7.3

The team decides to do an ANOVA analysis.

(a) For this data set, fill out the following:

$$\begin{aligned} \text{Number of subjects} &= \\ \text{Number of factors} &= \\ \text{Number of treatments} &= \end{aligned}$$

(b) Fill out the following ANOVA table.

	df	Sum Squares	Mean Squares	F-statistic	p-statistic
video					
Residuals					

Solution

(a)

$$\begin{aligned} \text{Number of subjects} &= 13 \\ \text{Number of factors} &= 1 \\ \text{Number of treatments} &= 3 \end{aligned}$$

(b) The complete table looks like:

	df	Sum Squares	Mean Squares	F-statistic
video	2	Residuals	10	

27.1: True or false: Two random variables with positive correlation cannot be independent.

Solution This is $\boxed{\text{true.}}$ Two random variables that are independent have zero correlation. The contrapositive of this statement is that if two random variables have nonzero correlation then they are not independent.

27.2: For X with finite second moment, what is $\text{Cor}(X, X)$?

Solution This is $\boxed{1}$, since

$$\text{Cor}(X, X) = \frac{\text{Cov}(X, X)}{\text{SD}(X)\,\text{SD}(X)} = \frac{\mathbb{V}(X)}{\text{SD}(X)^2} = 1.$$

27.3: If $R^2 = 0.36$, what is Pearson's r?

Solution If Pearson's r squared is 0.36, then $\boxed{r \in \{0.6000, -0.6000\}}$.

27.4: True or false: For data $\{X_i\}$ and $\{Y_i\}$ drawn iid from distributions with finite mean, variance, and covariance, Pearson's r converges to the true correlation as the number of sample points goes to infinity.

Solution $\boxed{\text{True.}}$ This is a consequence of the Strong Law of Large Numbers.

27.5: If $Y = 3X + 3$, what is $\text{Cor}(X, Y)$?

Solution Here $\boxed{\text{Cor}(X, Y) = 1}$. Any time $Y = aX + b$, where a is a positive constant, the correlation between the random variables is 1. Note that if I draw the X_i from X and then make each $Y_i = 3X_i + 3$, then the results will lie on a straight line with positive slope. For a more formal derivation:

$$\text{Cor}(X, 3X + 3) = \frac{\text{Cov}(X, 3X + 3)}{\text{SD}(X)\,\text{SD}\,3X + 3} = \frac{3\,\text{Cov}(X, X) + \text{Cov}(X, 3)}{\text{SD}(X)|3|\,\text{SD}(X)} = \frac{\mathbb{V}(X) + 0}{\text{SD}(X)^2} = 1.$$

27.6: True or false.

(a) Covariance is an inner product.

(b) Correlation is an inner product.

Solution

(a) $\boxed{\text{True.}}$ It is an inner product where the vectors consist of random variables where two random variables are equivalent if they differ by a constant.

(b) $\boxed{\text{False.}}$ Intuitively, correlation is like the cosine of the "angle" between the two random variables.

27.7: True or false: If $U_1, \ldots, U_n \sim \mathsf{Unif}([0, 1])$ where the U_i are independent, then $U_1^2 + \cdots + U_n^2 \sim \chi^2(n)$.

Solution $\boxed{\text{False.}}$ The χ^2 distribution arises from the sum of independent uniform random variables.

27.8: Suppose that Z_1 and Z_2 are independent, standard normal random variables. Let $X_1 = (1/\sqrt{2})Z_1 + (1/\sqrt{2})Z_2$ and $X_2 = Z_1$.

(a) What is the distribution of X_1?

(b) What is the distribution of X_2?

(c) True or false: The distribution of $X_1^2 + X_2^2$ is $\chi^2(2)$.

Solution

(a) For a normal random variable $N \sim \mathsf{N}(\mu, \sigma^2)$, $cN \sim \mathsf{N}(c\mu, c^2\sigma^2)$. Hence $(1/\sqrt{2})Z_1 \sim \mathsf{N}(0, \sigma^2/2)$. The sum of independent normal random variables is normal with the sum of the means and variances so

$$(1/\sqrt{2})Z_1 + (1/\sqrt{2})Z_2 \sim \mathsf{N}(0 + 0, 1/2 + 1/2) \sim \boxed{\mathsf{N}(0, 1)}.$$

(b) This is just the same as the distribtuion of Z_2, $\boxed{\mathsf{N}(0, 1)}$.

(c) This is $\boxed{\text{false}}$, because X_1 and X_2 (although they are standard normals) are not independent.

27.9: Find the Pearson's correlation coefficient for

$$(1.1, 0.4), (-3.2, 4.6), (0.1, 5.1).$$

Solution First find the means of the x and y values:

$$\bar{x} = -0.666666\ldots, \quad \bar{y} = 3.366666\ldots.$$

Next find Pearson's r:

$$\begin{aligned} r &= \frac{(1.1 - \bar{x})(0.4 - \bar{y}) + \cdots (0.1 - \bar{x})(5.1 - \bar{y})}{\sqrt{(1.1 - \bar{x})^2 + \cdots + (0.1 - \bar{x})^2}\sqrt{(0.4 - \bar{y})^2 + \cdots + (5.1 - \bar{y})^2}.} \\ &= \frac{-7.0366666\ldots}{\sqrt{10.1266666\ldots}\sqrt{13.326666\ldots}} \\ &= \boxed{\text{-0.6057}} \end{aligned}$$

28.1: In a contingency table, data are subject to what kind of contraints?

Solution Linear constraints.

28.2: Suppose that $(X_1, X_2, X_3) \sim \mathsf{Multinom}(3, 0.2, 0.5, 0.3)$, what is the chance $X_2 = 3$?

Solution This is the chance that of 3 subjects which have (independently) a 0.5 chance to equal 2, that all 3 subjects are 2. This probability is $(1/2)^3 1/8 = \boxed{0.1250}$

28.3: An auditor is checking glucose levels at two hospitals The glucose of each subject can be high (H), medium (M), or low (L). They gathered the following data.

	H	**M**	**L**	Total
Hospital 1	26	29	45	100
Hospital 2	44	26	30	100
Total	70	55	75	200

They want to test whether the glucose level is independent of where the patient is. Describe how you would test this at the 5% level, being sure to state your null hypothesis, test statistic (which you should calculate for this data), and rejection region (which you can write using a cdf or cdf^{-1} function, you do not have to calculate it exactly.)

Solution Let r_1, r_2 denote the row sums, and c_1, c_2, c_3 denote the column sums.

Null hypothesis The null hypothesis is that the table is drawn from a multinomial distribution with 200 subjects, where the distribution of entry x_{ij} is binomial with parameters n, and probability $(r_i/n)(c_j/n)$.

Test Statistic We can use a χ^2 statistic that is sum of the squares of the difference between the table entries x_{ij} and the mean entries. So

$$\chi^2 = \frac{(26 - (100)(70)/200)^2}{(100)(70)/200} + \cdots + \frac{(30 - (100)(75)/200)^2}{(100)(75)/200}. = \boxed{7.792}.$$

Rejection region The degrees of freedom are the six entries minus the two row and minus the three column constrains, but then one of those is redundant so we have to add it back. So we end with $6-2-3+1 = 2$ degrees of freedom. So under the null hypothesis, the χ^2 statistic should have a $\chi^2(2)$ distribution.

Typically we reject when the statistic is large. We want the probability that the statistic falls into the rejection region to be 0.05, so the probability that the statistic is smaller than the lower endpoint of the rejection region is 0.95. The rejection region then becomes $\boxed{[\text{cdf}^{-1}_{\chi^2(6)}(0.95), \infty)}$.

29.1: True or false: Pearson, Kendall, and Spearman correlation coefficients will always be the same for independent data.

Solution False. Changing the values of the points by a little bit will change r, but not change the ranks of the points, leaving Kendall and Spearman unchanged.

29.2: Consider the following three points:

$$(0.4, 0.6), (0.7, 0.5), (1.2, 1.1).$$

(a) Find Pearson's r

(b) Find Spearman's Rho

(c) Find Kendall's Tau

Solution

(a) Note

$$\bar{x} = \frac{0.4 + 0.7 + 1.2}{3} = \frac{2.3}{3},\ \bar{y} = \frac{0.6 + 0.5 + 1.1}{3} = \frac{2.2}{3},$$

so

$$r = \frac{a}{\sqrt{b}\sqrt{c}},$$

where

$$\begin{aligned} a &= (0.4 - 2.3/3)(0.6 - 2.2/3) + (0.7 - 2.3/3)(0.5 - 2.2/3) + (1.2 - 2.3/3)(1.1 - 2.3/3) \\ &= [0.44 + 0.14 + 1.43]/9, \\ b &= \sqrt{(1.2 - 2.3/3)^2 + (2.1 - 2.3/3)^2 + (3.6 - 2.3/3)^2} = \sqrt{2.94/9}, \\ c &= \sqrt{(1.8 - 2.2/3)^2 + (1.5 - 2.2/3)^2 + (3.3 - 2.2/3)^2} = \sqrt{1.86/9}, \end{aligned}$$

so $r = 2.01/\sqrt{2.94 \cdot 1.86} \approx \boxed{0.8595}$.

(b) Since $0.4 < 0.7 < 1.2$, the ranks of the x_i values are 1, 2, 3. Since $0.5 < 0.6 < 1.1$, the ranks of the y_i values are 2, 1, 3. The ranks averge to 2, hence

$$\begin{aligned} \rho &= \frac{(1-2)(2-2) + (2-2)(1-2) + (3-2)(3-2)}{\sqrt{(1-2)^2 + (2-2)^2 + (3-2)^2}\sqrt{(2-2)^2 + (1-2)^2 + (3-2)^2}} \\ &= \frac{1}{2} = \boxed{0.5000}. \end{aligned}$$

(c) The points 1 & 2 are discordant, while 1 & 3 and 2 & 3 are concordant. Hence Kendall's Tau is

$$\tau = \frac{2-1}{\binom{3}{2}} = \frac{1}{3} \approx \boxed{0.3333}.$$

29.3: Consider the following three points:

$$(0.4, 0.7), (0.7, 1.1), (1.2, 0.4).$$

(a) Find Pearson's r

(b) Find Spearman's Rho

(c) Find Kendall's Tau

Solution

(a) $\boxed{\text{-0.5519109}}$

(b) $\boxed{\text{-0.5000}}$

(c) $\boxed{\text{-0.3333}}$

29.4: Consider the following four data points:

$$(0.3, 1.2), (0.5, 2.4), (0.7, 1.7), (0.9, 2.0).$$

(a) Calculate the Pearson's correlation coefficient for this data.

(b) Calculate Kendall's Tau for this data.

(c) Calculate Spearman's rho for this data.

(d) Now suppose that the last data point $(0.9, 2.0)$ is replaced with $(0.9, 10.0)$. Repeat the calculation for Pearson's r, Kendall's tau and Spearman's rho.

Solution

(a) Here $x = (0.3, 0.5, 0.7, 0.9)$ and $y = (1.2, 2.4, 1.7, 2.0)$. Plugging into the formula for r gives $\boxed{r = 0.4339}$.

(b) Notice that the x_i are all in order, so the concordant (i, j) pairs are $(1, 2)$, $(1, 3)$, $(1, 4)$, and $(3, 4)$. The discordant (i, j) pairs are $(2, 3)$, $(2, 4)$. So Kendall's Tau is:

$$\frac{4-2}{4(3)/2} = 1/3 \approx \boxed{0.3333}$$

(c) For Spearman's rho, we calcuate Peason's r for the ranks of the vectors:

$$r_x = (1, 2, 3, 4), \; r_y = (1, 4, 2, 3),$$

which gives $\rho = \boxed{0.4000}$.

(d) Now consider

$$x = (0.3, 0.5, 0.7, 0.9), \quad y = (1.2, 2.4, 1.7, 10).$$

Immediate, r jumps up to 0.8002, nearly double what it was before. The ranks also change:

$$r_x = (1, 2, 3, 4), \quad r_y = (1, 3, 2, 4)$$

This changes one discordant pair $(2, 4)$ to concordant, so Kendall's tau changes to $2/3$. Similarly, Spearman's rho also increases to 0.8. Overall the changes are:

$$\boxed{r = 0.8002, \; \tau = 0.6666, \; \rho = 0.8000}.$$

30.1: True or false?

(a) Changing the order of terms in the model can change the least squares fit.

(b) Changing the order of terms in the model can change the p-value for the terms in the ANOVA.

Solution

(a) $\boxed{\text{False.}}$ Changing the order of variables in a linear model is the same as swapping columns in the X matrix. This merely swaps the associated variables in $\hat{\beta}$.

(b) $\boxed{\text{True.}}$ The order in which the sum of squares is calculated can lead to a variable presented earlier have a lower p-value.

31.1: When we wish to show that one effect causes another, and we have complete control of the experimental design, we use what method?

Solution $\boxed{\text{Randomized block design}}$. Randomly assigning which subjects get which treatments controls for other factors that might be affecting the outcome.

32.1: Suppose that $(X_1, \ldots, X_n)$ given λ are iid $\mathsf{Exp}(\lambda)$. Show that $S(X_1, \ldots, X_n) = X_1 + \cdots + X_n$ is a sufficient statistic for λ.

Solution The density of the data is

$$\begin{aligned} f_\lambda(x_1, \ldots, x_n) &= \prod_{i=1}^{n} \lambda \exp(-\lambda x_i) \mathbb{1}(x_i \geq 0) \\ &= \lambda^n \exp(-\lambda(x_1 + \cdots + x_n)) \mathbb{1}(x_1, \ldots, x_n \geq 0) \\ &= \mathbb{1}(x_1, \ldots, x_n \geq 0) \lambda^n \exp(-\lambda S(x_1, \ldots, x_n)). \end{aligned}$$

Hence setting $h(x) = \mathbb{1}(x_1, \ldots, x_n \geq 0)$ and $g_\lambda(S(x_1, \ldots, x_n)) = \lambda^n \exp(-\lambda S(x_1, \ldots, x_n))$ in the Factorization Theorem completes the proof. $\square$

33.1: Consider the function

$$f(s) = |s - 13| + |s - 14| + |s - 17|.$$

Find $\min f(s)$ and $\arg\min f(s)$ for $s \in \mathbb{R}$.

Solution Suppose $s \leq 13$. Then

$$f(s) = 13 - s + 14 - s + 17 - s = 44 - 3s.$$

Since the coefficient of s is negative, the minimum occurs when s is as large as possible, 13, and $f(13) = 5$.

Next suppose $s \in [13, 14]$. Then

$$f(s) = s - 13 + 14 - s + 17 - s = 18 - s.$$

Again the coefficient on s is negative, so the minimum occurs when s is as large as possible, 14 in this case. And $f(14) = 4$.

Next is when $s \in [14, 17]$. Then

$$f(s) = s - 13 + s - 14 + 17 - s = s - 10.$$

Here the coefficient of s is positive, so the minimum occurs when s is as small as possible, 14. As before $f(14) = 4$.

Finally, for $s \geq 17$,

$$f(s) = s - 13 + s - 14 + s - 17 = 3s - 44.$$

Again the coefficient of s is positive, so the minimum occurs where s is as small as possible, so at 17. Here $f(17) = 7$.

Combining these results gives $\boxed{\min f(s) = 4,\ \arg\min f(s) = 14}$.

Index

About the Text

Statistical Inference: Theory and Labs is a one-semester course text in statistics aimed at students that have already had a semester course in Calculus based probability. The course is split between explanation of theory (roughly two-thirds), and laboratory exercises (roughly one-third) where students tackle hands on problems analyzing real data sets.

About the Author

Mark Huber is the Fletcher Jones Professor of Mathematics and Statistics at Claremont McKenna College. He received his Ph.D. in Operations Research and Industrial Engineering from Cornell University in 1999. He has been an NSF Postdoc in the Department of Statistics at Stanford University, and is a recipient of an NSF CAREER Award.