Engineering Statistics

Matias Shedden

December 2023

Contents

0 Introduction

The definition of inference is "a conclusion reached on the basis of evidence and reasoning." In statistics, we are generally concerned with evidence that has been *quantified* in some way. Quantification allows us to claim that the evidence we gather has some deeper mathematical interpretation. Starting from an experiment, the job of the statistician is to assign a mathematical framework to the data that was collected, and then make probabilistic conclusions within this framework. These probabilistic conclusions are then interpreted within the context of the experiment.

We interpret the data as being *random*, or *randomly sampled*. Formalizing the notion of randomness is the purpose of probability theory, which comprises the first chapter of this book.

Definition: The *sample space* (S) is defined to be the set which contains all possible *outcomes* of an experiment.

Example: A fair coin has a 50/50 chance of landing on heads or tails. Let 0 represent tails, and 1 represent heads. The sample space (denoted S) for the experiment of tossing a singular fair coin is given by:

$$S = \{0, 1\}$$

If we know that the coin is fair, we can completely describe the probability of observing each result (we have a probability of 0.5 associated with each outcome, 0 or 1). This complete description of the probabilities of observing each outcome in our experiment is the end goal of statistics.

This "complete description" is referred to by many names. In probability theory, this is referred to most commonly as the *distribution*. This is an intuitive name; if we know how the probability is "distributed", we know how likely it is to observe a certain outcome. In the example of the coin, I could distribute the probabilities differently, say 0.3 to heads and 0.7 tails. This still serves as a complete description of the probability of observing each outcome of the coin.

Other times, this "complete description" is referred to as the *population*. This is also an appeal to intuition, this time appealing to the idea that there is some infinite population from which we observe a sample. We

could also call it the *data generating process*, since knowing the probability of observing each outcome would allow you to generate new data.

In this book, and in most elementary treatments of statistics, however, the central probabilistic concept we consider is called a *random variable*. For now, we will start with an intuitive understanding of what a random variable is, before we move on to the more formal mathematical definition.

Suppose that A is some subset of the real numbers, denoted $\mathbb{R}$. Then a random variable X can be thought of as a variable (i.e. a "non-random" variable x, that you might consider in algebra or calculus class), that has a certain *probability* of falling in the set A. We will denote the probability of X being in A as:

$$P(X \in A)$$

When we make a probabilistic statement, we are essentially saying: "what is the probability of a certain event occurring". When we consider random variables, we make this statement more specific: "what is the probability of the random variable falling within a certain range of values". Note of course that $X \in A$ is, in itself, a particular kind of event.

The general process we consider is this: we take reality, turn it into quantifiable observations through experimentation, and then use the appropriate statistical analysis to estimate probabilities.

0.1 Population and Sample

The population is not an "infinite set" from which we draw observations, though it is often purposefully *interpreted* this way in the context of certain experiments. What the *population* refers to is simply the *data generating process*. It is a function which describes the probability of observing all elements of the sample space.

Example: Suppose a pollster draws a sample of opinion from a group of 100 interviewees. One may interpret the "population", in this case, to be the entire voting population of the country.

Example: Consider the example of a fair coin. We can describe the probabilities completely using a *probability mass function*, which assigns a probability to each event in the sample space:

$$f_X(x) = \left\{ \begin{array}{ll} 1/2, & \text{if } x = 1 \\ 1/2, & \text{if } x = 0 \end{array} \right\} = 1/2$$

This true model completely describes the *population*. It completely describes all of the "randomness" (the probability of observing each outcome) in the coin tossing experiment.

Note that we have certain analogies between "population" and "sample" *quantities*. Consider the case of a continuous observation (an observation that can values on some subset of the real numbers). The population mean is the *expected value* of the probability density function, which we will define later. It is equivalent to finding the center of mass of a continuous 1-dimensional density. The sample mean is simply the familiar arithmetic mean. As our sample size grows, we should expect the sample mean to approach the population mean (the law of large numbers), as our individual observations better represent the true *continuous* density.

0.2 Uncertainty in Conclusions

The goal of *statistical inference* is to take observations, i.e. a *random sample*, and then use this to make conclusions about the *population*. Since we cannot directly observe the properties of the population, these conclusions are *probabilistic* in that they are not certain. Rather, they involve some level of uncertainty. The mathematical underpinnings of statistics are what allow us to quantify this uncertainty, under the assumptions of our model.

Example: Confidence Intervals. We may estimate some value and then claim that we are 95 confident it lies within a particular range. For example, I could report an estimate $\hat{x}$ as:

$$\hat{x} = 5 \pm 1$$

We say that 5 is a *point estimate* for x, which we report along with some uncertainty. The interval $[4, 6]$ as a whole is known as an *interval estimate* for x.

0.3 Models

Definition: [1] A **statistical model** is a mathematical model that embodies a set of statistical assumptions concerning the generation of sample data (and similar data from a larger population). A statistical model represents, often in considerably idealized form, the data-generating process.

The "data-generating process" which the above definition refers to is why we use probability theory and the idea of a random variable throughout statistics. Essentially, we interpret each data point or observation as having had a particular probability of occurring (a probability which is specified by the statistical model). Using this, we can then infer, *from the data*, the properties of the data-generating process.

Example: (Fitting a model) Suppose we have an unfair coin, and we don't know the probability of obtaining heads. If we can flip the coin, we can estimate the probability of obtaining a heads from a random sample of flips. The most intuitive way to do this is to simply divide the number of heads we observed by the total number of flips. This is an attempt to describe the population. In particular, it is an attempt to estimate the density function. It could also be seen as a point estimate for the probability of obtaining a heads.

1 Probability Theory

1.1 Naive Set Theory

1.1.1 Set builder notation

Sets are denoted by curly brackets ({}). *Elements* of a set are contained within the brackets, separated by commas.

Example: $A = \{a, b, c\}$ We use the symbol $\in$ to denote that an element is "in" a certain set. For example, we would write $b \in A$.

Common sets which have a specific symbol associated with them: the real numbers $\mathbb{R}$, the integers $\mathbb{Z}$, the natural numbers $\mathbb{N}$, and the empty set $\emptyset$.

A vertical bar (|) or a colon (:) means "such that", i.e. all elements "such that" a specified condition is met.

Example: All even integers: $\{x \in \mathbb{Z} | x/2 \in \mathbb{Z}\}$ or $\{x = 2m : m = \cdots, -2, -1, 0, 1, 2, \cdots\}$

Definition: Set containment. We say that a set A is *contained* in a set B if $x \in A \implies x \in B$. We denote this as $A \subset B$.

Definition: Set equivalence We say that a set A is *equal* to another set B ($A = B$) if $A \subset B$ and $B \subset A$.

Definition: The *Cartesian Product* of two sets A and B is defined to be

$$A \times B := \{(a, b) : a \in A \text{ and } b \in B\}$$

Suppose we take the Cartesian product of A with itself, then we can write:

$$A \times A = A^2$$

If we take the Cartesian product of A with itself n times, then we write:

$$A \times \cdots \times A = A^n$$

The most common Cartesian products are those of the real numbers with itself. For example, the usual coordinate plane:

$$\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$$

1.1.2 Set operations

Definition: Complementation. Consider S to be the sample space of some experiment. The complement of a subset $A \in S$ is $A^c := \{\omega : \omega \notin A\}$

Definition: Intersection. $A \cap B = \{\omega : \omega \in A \text{ and } \omega \in B\}$

Definition: Union. $A \cup B = \{\omega : \omega \in A \text{ or } \omega \in B\}$

Example: The above three operations can be visualized using venn diagrams. They're a lot like the logical operations "not", "and", "or".

Note Let A_i be sets for each $1 \leq i \leq n$. Then for intersection, for example, we can write

$$\bigcap_{i=1}^{n} A_i := \{\omega : \omega \in A_i, \text{ for all } 1 \leq i \leq n\}$$

Definition: The power set (denoted $\mathcal{P}(A)$) is the set which contains all possible subsets of A.

Note: The power set of A contains A itself, as well as the empty set $\emptyset$. Also note that the power set is a *set of sets*, that is, an set whose elements are themselves sets.

1.1.3 Laws of set operations

Theorem: Union and intersection are commutative and associative operations.

Theorem: Distributive Laws:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

and

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

Theorem: (DeMorgans Laws).

$$(A \cup B)^c = A^c \cap B^c$$

and

$$(A \cap B)^c = A^c \cup B^c$$

Definition Disjoint Sets. Sets A and B are disjoint there exists no ω such that $\omega \in A$ and $\omega \in B$.

1.2 The Sample Space

Recall the following definition:

Definition: The *sample space* (S) is defined to be the set which contains all possible *outcomes* of an experiment.

An *experiment* is any process that generates data (flipping a coin, etc.). The data is a set of *observed outcomes* (*observations*) of the experiment. Each outcome in the sample space is referred to as an *element* or *member* of the sample space (the sample space is a set).

Examples: Coin: $S = \{H, T\}$

6-sided Die: $S = \{1, 2, 3, 4, 5, 6\}$

2 6-sided Dice: $S = \{(1,1), (1,2), ...\} = \{(x,y) : 1 \leq x \leq 6, 1 \leq y \leq 6\}$ (there are $6^2 = 36$ elements of this sample space)

3 6-sided Dice: $S = \{(1,1,1), (1,1,2), ...\} = \{(x,y,z) : 1 \leq x, y, z \leq 6$ (there are $6^3 = 216$ elements of this sample space)

1.3 Events

Definition: An *event* is a subset of the sample space.

Examples: If our experiment is to toss a coin three times, an event could be that we obtain two heads. This event is denoted by the set $E = \{(H,H,T), (T,H,H), (H,T,H)\}$
If our experiment is one roll of a 6-sided die, an event could be that we roll a 4. This event is represented by the set $E = \{4\}$

If our experiment is rolling two dice, an event could be that the sum of the dice is equal to 4. This event is denoted by the set $E = \{(1,3),(2,2),(3,1)\}$

1.4 Probability

Definition: A *probability function* is a function P which takes subsets of the sample space S as inputs, and which satisfies the following properties:

- $0 \leq P(A) \leq 1$ for any $A \subset S$
- $P(S) = 1$ and $P(\emptyset) = 0$
- If $A_1, A_2, \cdots$ is a collection of disjoint events, then $P(\bigcup_{n=1}^{\infty} A_n) = \Sigma_{n=1}^{\infty} P(A_n)$

The above properties are known as the "Axioms of Probability", or the "Kolmogorov Axioms". These properties allow the function P to behave in a way which is consistent with our preexisting idea of what "probabilities" should be.

Example: Consider a coin, so that our sample space is $S = \{0,1\}$. Define P as follows:

$$\begin{aligned} P(\emptyset) &= 0 \\ P(\{0\}) &= p \\ P(\{1\}) &= 1-p \\ P(\{0,1\}) &= P(S) = 1 \end{aligned}$$

Where p is some constant between 0 and 1. Then the function P satisfies the axioms of probability, and is hence a probability function.

Exercise: Consider an experiment where we have n (finite) number of elements in the sample space S, each of which have an equal probability of occurring. Define the appropriate probability function P.

Example: Consider the experiment where we flip 2 coins (again, 1 represents a heads, 0 represents a tails). Our sample space is

$$S = \{(0,0),(1,0),(0,1),(1,1)\}$$

Clearly, each event has an equal probability of occurring, so we can refer to the previous example in order to develop a probability function.

Now suppose that we change the experiment: flip 2 coins and count the number of heads that we observe. The sample space is now:

$$S = \{0, 1, 2\}$$

The probability of observing a 1 is twice as large as observing 0 or 2. We know this simply by counting the number of elements of our sample space which contain exactly 1 heads.

Side note: We will see later that this is the *induced* probability function, obtained by defining a *random variable*, which are functions from the sample space to the real numbers. In this case, the random variable we're considering is a function which takes an input of an element in the sample space (the result of two coin flips) and returns the number of heads.

Example: Consider an experiment where you roll 3 six-sided die. The sample space has $6^3 = 216$ elements, and each individual element has equal probability, if we consider fair dice.

Suppose that we want to find the probability of the following event:

$$A = \{\text{roll two 1's}\}$$

aka we wish to find $P(A)$. All the elements of A have the form:

$$(1, 1, (\text{not } 1)), (1, (\text{not } 1), 1), ((\text{not } 1), 1, 1)$$

Since there are 5 possibilities for "not 1", we have $3 * 5 = 15$ outcomes such that there are exactly 2 ones. Therefore, the probability is $15/216 \approx 0.0694$. But what if we rolled 10 six-sided die. Could we still calculate these probabilities? More complicated cases necessitate the use of counting rules.

1.5 Counting Rules

Counting is useful for performing some intuitive examples of probability. Consider the situation in which the sample space is finite and contains n total elements, e.g.

$$S = \{s_1, s_2, \cdots, s_n\}$$

Suppose further that

$$P(\{s_i\}) = \frac{1}{n} \qquad i = 1 \cdots, n$$

That is, each individual outcome of the experiment occurs with equal probability. If this is the case,

$$P(A) = \frac{|A|}{|S|}$$

where $|A|$ is the *cardinality* of the set A, i.e. the number of elements in the set A. Likewise $|S|$ is the cardinality of the sample space, which in this example, means $|S| = n$.

Rule: (Multiplication Rule) If an operation can be performed in n_1 ways, and for each of those ways, a second operation can be performed n_2 ways, then the two operations can be performed together in $n_1 n_2$ ways

Definition: Factorial: The number of ways to arrange n objects (the number of permutations) is $n!$ (read as "n factorial", where $n! = n * (n - 1) * \cdots * 2 * 1$)

Definition: Permutations: The number of ways that r objects can be chosen from n total objects where order matters is $nPr = \dfrac{n!}{(n-r)!}$

Definition: Combinations: The number of ways that r objects can be placed from n total objects where order doesn't matter is $nCr = \dfrac{n!}{(n-r)!r!}$

Definition: Tho *choose* function is defined to be the number of combinations of n distinct objects taken r at a time. It is defined as:

$$\binom{n}{r} = \frac{n!}{(n-r)!r!}$$

Rule: Sampling with replacement: The number of ways to arrange r objects from n total objects where order matters, with replacement, is n^r

Remark: There is one final case where we can arrange objects where order doesn't matter, with replacement. This sometimes called "stars and bars", which we won't cover here.

Example: Recall the example from above: Consider an experiment where you roll 3 six-sided die and count the number of 1's we obtain.

The sample space of dice rolls has $6^3 = 216$ elements.

The set of outcomes for which there are 2 ones looks like this:

$$(1, 1, (\text{not } 1)), (1, (\text{not } 1), 1), ((\text{not } 1), 1, 1)$$

Using counting rules, we can explain this further. There are three possible locations where we can put the fixed "1" values, from which we must choose two. Then, there are five possible options for the final value, from which we must choose one. Therefore, the total number of outcomes such that there are exactly two 1s is

$$\binom{3}{2}\binom{5}{1} = \frac{3!}{2!(3-2)!}\frac{5!}{1!(4-1)!} = 3 * 5 = 15$$

Counting rules can allow us to solve even more complicated examples. Consider rolling 10 six-sided die. What is the probability of rolling exactly 6 ones? Using the same logic, we will have ten possible locations where we can put the fixed "1" values, from which we must choose six. Then we have five options for each of the four remaining locations. Hence the number of observations such that there are exactly 6 ones is:

$$\binom{10}{6}5^4 = 131250$$

Note that in the experiment with 10 die, the sample space has $6^{10} = 60466176$ elements. Therefore the probability of observing exactly 6 ones is:

$$\frac{131250}{60466176} = 0.00217$$

1.6 Properties of the Probability Function

Proposition: $A \subset B \implies P(A) \leq P(B)$

Proof: Note that we can write

$$B = A \cup (B \cap A^c)$$

where A and $B \cap A^c$ are disjoint sets. Therefore, by the third axiom of probability,

$$P(B) = P(A) + P(B \cap A^c) \geq P(A)$$

since $P(B \cap A^c) \geq 0$ by the first axiom of probability.

Proposition: If $A \subset B$, then $P(B \cap A^c) = P(B) - P(A)$

Proof: Recall from the above proof that:

$$P(B) = P(A) + P(B \cap A^c)$$

Rearranging gives the desired result.

Proposition: $P(A^c) = 1 - P(A)$

Proof: Note that $S = A^c \cup A$ is a disjoint union of sets (A^c and A are disjoint, recall the definition from earlier).

Therefore, we can apply the third axiom of probability and say:

$$P(S) = P(A^c \cup A) = P(A^c) + P(A)$$

Furthermore, we know that $P(S) = 1$. Therefore

$$1 = P(A^c) + P(A)$$

and hence

$$P(A^c) = 1 - P(A)$$

Proposition: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Proof: First note this is easy to see if you consider a Venn diagram. Here is a formal proof:

$$P(A \cup B) = P(A) + P(B \cap A^c)$$

Furthermore, we have that

$$P(B) = P((B \cap A) \cup (B \cap A^c)) = P(B \cap A) + P(B \cap A^c)$$

which, by subtracting $P(B \cap A)$ from both sides, yields

$$P(B \cap A^c) = P(B) - P(B \cap A)$$

By substituting into our first relationship, we have the result:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B \cap A^c) \\ &= P(A) + P(B) - P(B \cap A) \end{aligned}$$

Example: The Birthday Problem.

There are 177 people in the class. What is the probability that at least two people share a birthday?

Define A to be the event that at least two people share a birthday. A much easier way to calculate this probability is to consider the event A^c, the event that no one shares a birthday, and then use the property $P(A) = 1 - P(A^c)$ that we derived above.

Using the multiplication rule, the total number of ways for no one to share a birthday is (permutations):

$$365 \cdot 364 \cdots 189 = \frac{365!}{188!} = \frac{365!}{(365-177)!}$$

The total number of ways for birthdays to be had is 365^{177} (The number of ways to arrange r objects from n total objects where order matters, with replacement).

Therefore,

$$P(A) = 1 - P(A^c) = 1 - \frac{365!/188!}{365^{177}} \approx 1$$

Thus it is almost certain that at least two people share a birthday.

Note that if there were 23 people, we have that:

$$P(A) = 1 - \frac{365!/342!}{365^{23}} \approx 0.5073$$

so if there are 23 people in a room, it is more likely than not that there are two of them who share a birthday.

1.7 Conditional Probability

Definition: Let A and B be events. The *conditional probability* of event A given that event B has occurred is defined as

$$P(B|A) := \frac{P(B \cap A)}{P(A)}$$

Note that we can also write:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Example: Consider the experiment of flipping 3 coins. The sample space is:

$$S = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$$

Consider the event in which we obtain at least two heads, call it A. Then,

$$A = \{HHH, HHT, HTH, THH\}$$

Suppose we want to calculate the probability that we get no tails, given that we have get two heads. Let B be the event that we get no tails, then we are interested in $P(B|A)$. One way to think of this problem is that by conditioning on the event A, we are considering A to be the "new" sample space. Then, we just have to count the outcomes in A which satisfy B. In this case, it is only the outcome HHH. Then we divide by the number of elements in A. In this case, we have $1/4$.

Another way to approach this problem is by simply using the definition. This involves finding the set $B \cap A$:

$$B \cap A = \{\text{outcomes such that there are two heads and no tails}\} = \{HHH\}$$

Clearly this will lead to the same result.

Example: Is $P(A|B)$ a probability function? We need to check that the function $P(\cdot|B)$ satisfies the axioms of probability.

First axiom:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \geq 0$$

This is clearly true since $P(A \cap B)$ and $P(B)$ are both greater than or equal to 0.

Second axiom:

$$P(S|B) = \frac{P(S \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

Third axiom:

$$
\begin{aligned}
P\left(\bigcup_{i=1}^{\infty} A_i \middle| B\right) &= \frac{P((\cup_{i=1}^{\infty} A_i) \cap B)}{P(B)} \\
&= \frac{P(\cup_{i=1}^{\infty} (A_i \cap B))}{P(B)} \\
&= \frac{\sum_{i=1}^{\infty} P(A_i \cap B)}{P(B)} \\
&= \sum_{i=1}^{\infty} P(A_i|B)
\end{aligned}
$$

Therefore we have that conditioning on an arbitrary event B does indeed result in another valid probability function, defined on the sample space S. It will always return 0 when we consider sets which are disjoint from B, which is why we can interpret this as having "reduced" the sample space to just the set B.

1.8 Event Relations

Here we define three ways in which we can describe relationships between events: complementary, independent, and mutually exclusive.

Definition: Let A be an event. The events A and A^c are complementary events (we've already seen this definition)

Definition: Let A and B be two events. A and B are called *independent* if $P(A \cap B) = P(A)P(B)$. We define two events to be *dependent* if they are **not** independent

Note: This is equivalent to saying $P(A|B) = P(A)$.

Definition: Let A and B be two events. A and B are called *mutually exclusive* if $P(A \cap B) = 0$.

Note: Disjoint sets are necessarily mutually exclusive since for A, B, disjoint, $P(A \cap B) = P(\emptyset) = 0$).

Example: Suppose our experiment consists of tossing a 6-sided die. Let A be the event that we roll a 5 or 6 and B be the event that we roll a 2 or 3. Then events A and B are mutually exclusive.

Let event A be the event that we roll a 1, 2, or 3 and B be the event that we roll at 4, 5, or 6. Then events A and B are complementary.

Example: Suppose our experiment consists of tossing two coins. Then the event that we get a heads on the first toss is independent from the event that we get a heads on the second toss.

1.9 The Law of Total Probability

Theorem: Let A and B be events. Then,

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

Proof: We've seen the logic here before, in the proof that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

First, note that $A \cap B$ and $A \cap B^c$ are disjoint sets. It is also true that

$$A = (A \cap B) \cup (A \cap B^c)$$

Therefore, by the axioms of probability,

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

Using the definition of conditional probability, we can write:

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

which is our desired result.

The following is a more general statement of the above:

Theorem: Let the events $B_1, \cdots, B_k$ constitute a partition of the sample space S such that $P(B_i) \neq 0$ for each i. Then for any event A, we have

$$P(A) = \sum_{i=1}^{k} P(B_i \cap A) = \sum_{i=1}^{k} P(B_i)P(A|B_i)$$

Note: A partition means that the B_i are all disjoint, and

$$\bigcup_{i=1}^{k} B_i = S$$

Example: Consider an experiment which consists of flipping a coin followed by rolling a die. If we get heads, we roll a die with $P(\text{roll a six}) = 2/3$. If we get tails, we roll a fair die. Find the probability of flipping heads, given that we did not roll a 6.

Define the following events:

$$A = \{\text{Flip Heads}\} \qquad B = \{\text{Roll a 6}\}$$

Then, we wish to find $P(A|B^c)$

Firstly:

$$P(A \cap B^c) = P(A) \cdot P(B^c|A) = 1/2 \cdot 1/3 = 1/6$$

Secondly, using the law of total probability:

$$\begin{aligned} P(B^c) &= P(B^c|A)P(A) + P(B^c|A^c)P(A^c) \\ &= 1/3 * 1/2 + 5/6 * 1/2 \\ &= 1/6 + 5/12 \\ &= 7/12 \end{aligned}$$

Therefore,

$$P(A|B^c) = \frac{P(A \cap B^c)}{P(B^c)} = \frac{1/6}{7/12} = 2/7$$

1.10 Bayes Rule

Theorem: (Bayes' Rule with two events)

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Proof: Bayes rule is a double usage of the definition of conditional probability.

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(B \cap A)}{P(B)} \\ &= \frac{P(B|A) \cdot P(A)}{P(B)} \end{aligned}$$

Theorem: (Bayes' Rule Generally). Let $B_1, \cdots, B_k$ be a partition of the sample space S. For each B_r which is a member of that partition, we have that

$$\begin{aligned} P(B_r|A) &= \frac{P(A|B_r) \cdot P(B_r)}{P(A)} \\ &= \frac{P(A|B_r) \cdot P(B_r)}{\Sigma_{i=1}^{n} P(A|B_i) \cdot P(B_i)} \end{aligned}$$

Example: Recall the example from the previous section: Consider an experiment which consists of flipping a coin followed by rolling a die. If we get heads, we roll a die with $P(\text{roll a six}) = 2/3$. If we get tails, we roll a fair die. Find the probability of flipping heads, given that we did not roll a 6.

$$A = \{\text{Flip Heads}\} \qquad B = \{\text{Roll a 6}\}$$

Using Bayes' Rule:

$$P(A|B^c) = \frac{P(B^c|A)P(A)}{P(B^c|A)P(A) + P(B^c|A^c)P(A^c)}$$

where

$$\begin{aligned} P(B^c|A) &= 1/3 \\ P(A) &= 1/2 \\ P(B^c|A^c) &= 5/6 \\ P(A^c) &= 1/2 \end{aligned}$$

Plugging in, we have

$$\frac{1/3 * 1/2}{1/3 * 1/2 + 5/6 * 1/2} = 2/7$$

1.11 Random Variables

Definition: A *random variable* (abbreviated r.v.) is a function that associates a real number with each element in the sample space.

Note: We commonly denote random variables using a capital letter, usually X, Y or Z.

Example: Consider the example of flipping a coin 5 times. If we let 0 represent tails, and 1 represent heads, then our outcomes will be in $\mathbb{R}^5$. For example, $(1, 0, 0, 0, 1)$.

Suppose we are concerned only with the number of heads. Then we can define a random variable X as follows:

$$X(a, b, c, d, e) = a + b + c + d + e$$

The function X maps from the sample space $\mathbb{R}^5$ to $\mathbb{R}$. It is easy to see that the function "counts" the number of heads which were observed in the five coin tosses.

Example: Consider flipping a single coin.

If we, like above, let $S = \{0, 1\}$, then the random variable X is simply the identity mapping: $X(x) = x$ where $x = 0, 1$.

If $S = \{H, T\}$, for example, we could choose either of the following mappings, depending on whether or not we want 1 to represent heads or tails:

$$X(H) = 1, X(T) = 0 \quad \text{or} \quad X(H) = 0, X(T) = 1$$

Note: The above two examples map to discrete subsets of the real numbers $\mathbb{R}$ (e.g. X can only possibly map to the numbers $0, 1, 2, 3, 4, 5$ in the first example). We can have random variables that map to infinite subsets of $\mathbb{R}$.

Example: Consider the height of a sunflower randomly chosen from a particular field. There are uncountably many possible outcomes of this value. In reality, of course, this depends on how precisely we can record the height of the flower.

The random variable is a function, so we need an input to have an output. The output is interpreted as an "observed value" of the random variable, corresponding to a particular element of the sample space. For example, if we observe (H, H, H, T, T), this is an *element* of our sample space. The random variable "how many heads" takes the observation (an element of the sample space) as input, and outputs the real number 3. We call 3 the observed value of the random variable X. If we wanted to know the probability of X taking the value 3, we would use the notation $P(X = 3)$, as we define below.

Definition: The probability of a random variable X assuming a value x, denoted $P(X = x)$, is defined to be

$$P(X = x) = P(\{\omega : X(\omega) = x\})$$

Note: The random variable is interpreted by us as a "random" observed value, hence the name. However, like the notion of probability, this is just one interpretation of a very general mathematical idea.

The purpose of the random variable is to "translate" probability statements about the sample space to those about the real numbers.

Example: Coin flips: Flip 2 (fair) coins. What is the probability of obtaining at least one heads? Clearly, we could enumerate every element in the sample space for which this is the case:

$$\{(H,T),(T,H),(H,H)\}$$

Then we take $3/4 = 0.75$ to be the probability, since each element in the sample space has the same probability of occurring in this example. The probability function on the sample space S is defined as $P(\{\omega\}) = 1/4$ for all elements ω.

Consider instead the random variable method: define X to be the number of heads. Then the probability in question can be expressed as $P(X \geq 1)$.

Let's define everything explicitly and see how we can systematically arrive at the new probability function:

$$X((a,b)) := a + b \quad \text{for } a, b \in 0, 1$$

By definition, we have that:

$$P(X = x) = P(\{\omega : X(\omega) = x\})$$

Hence, we can calculate probabilities as follows:

$$\begin{aligned} P(X = 0) &= P(\{\omega : X(\omega) = 0\}) \\ &= P(\{(0,0)\}) = 1/4 \\ P(X = 1) &= P(\{\omega : X(\omega) = 1\}) \\ &= P(\{(0,1),(1,0)\}) = 1/2 \end{aligned}$$

and so on. This is a new probability function, which is now defined on the real numbers, rather than on the original sample space S. In its entirety, we have:

$$P(X = 0) = 1/4$$
$$P(X = 1) = 1/2$$
$$P(X = 2) = 1/4$$

So we can calculate $P(X \geq 1)$ as:

$$P(X \geq 1) = P(X = 1) + P(X = 2) = 1/2 + 1/4 = 3/4$$

Definition: A *discrete* random variable is a random variable with a countable range. That is, the "observed" values of the random variables constitute a countable set.

Note: A particular example of a countable set is a finite set, which contains a finite number of values.

Result: For a discrete random variable X, we have that:

$$P(X \in A) = \sum_{x \in A} P(X = x)$$

This follows from the third axiom of probability, since each set containing a single point x is necessarily disjoint from any other such set.

Definition: The *probability mass function* (abbreviated pmf) of a discrete random variable X is a function (denoted $f_X(x)$) defined to be

$$f_X(x) = P(X = x)$$

Note that because $f_X(x)$ is defined to be $P(X = x)$, where P is a valid probability function defined on $\mathbb{R}$, it must posses the following properties:

(1) $f_X(x) \geq 0$

(2) $\sum_x f_X(x) = 1$

Note: The notation $\sum_x$ is a sum over all possible values of x (i.e. a sum over the range of the random variable X, which is countable in the case of a discrete random variable).

Example: Equivalence of the properties of a pmf and the axioms of probability.

Let's show that the axioms of probability imply the above properties (1) and (2). First recall the definition of $P(X = x)$:

$$P(X = x) := P(\{\omega : X(\omega) = x\})$$

and the first axiom of probability:

$$0 \leq P(A) \leq 1 \text{ for any set } A$$

Therefore we can simply take $A = \{\omega : X(\omega) = x\}$, and we have that:

$$0 \leq P(X = x) \leq 1$$

and hence

$$P(X = x) = f_X(x) \geq 0$$

To prove property (2) of a pmf, let x_1 and x_2 be two distinct elements of the range of X (two distinct possible observed values). Then we have that the following sets are disjoint:

$$\{\omega : X(\omega) = x_1\} \quad \text{and} \quad \{\omega : X(\omega) = x_2\}$$

Furthermore, since X is defined on all of S, S is the union of all sets of the above form, i.e. we can write:

$$\bigcup_x \{\omega : X(\omega) = x\} = S$$

and this is a disjoint union. Therefore,

$$P(S) = P\left(\bigcup_x \{\omega : X(\omega) = x\}\right)$$
$$= \sum_x P(\{\omega : X(\omega) = x\}) = \sum_x P(X = x) = \sum_x f_X(x)$$

By the axioms of probability, $P(S) = 1$, so we have the result:

$$\sum_x f_X(x) = 1$$

■

Example: Bernoulli Random Variable with probability p: There are two possible outcomes, 0, and 1, where the probability of observing a 1 is p. (Likewise, the probability of observing a 0 is $1 - p$)

$$f_X(x) = p^x(1-p)^{1-x} \qquad x = 0, 1$$

Let's check the properties (1) and (2) from above:

1) $f_X(x) \geq 0$

We have two possibilities:

$$f_X(1) = p^1(1-p)^0 = p$$
$$f_X(0) = p^0(1-p)^1 = 1 - p$$

If $0 \leq p \leq 1$, then we have that both of the above values are ≥ 0

2) $\sum_x f(x) = 1$

$$\sum_{x \in \{0,1\}} f(x) = p + 1 - p = 1$$

Example: Binomial Random Variables: Consider the "how many heads in n coin flips" example from above:

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, 2, \cdots, n$$

Again, let's check the axioms of probability.

1) $f_X(x) \geq 0$:

$\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is always positive since the factorial is always positive. p^x and $(1-p)^{n-x}$ are also always positive, since for $0 \leq p \leq 1$, we are raising a positive number to a power so the result must be positive. Therefore the entire pmf $f_X(x)$ is positive.

2) $\sum_x f(x) = 1$:

Recall the *binomial theorem*, which states:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k$$

If we allow x in the above formula to be p, and y to be $1-p$, we arrive at the following result:

$$\sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = (p + (1-p))^n = 1^n = 1$$

Note: To the right of the probability mass function, we give the values of x for which $f_X(x)$ is nonzero. If x is not in the set specified, then we have that $f_X(x) = 0$.

In each of these cases, we're seeking to describe the randomness associated with the random samples that we observe. Most important point: In statistics, we start with *data*, which is interpreted as the observed value of some random sample (which follows some distribution), and then try and make conclusions about the associated *parameters* in its distribution function. Oftentimes, making conclusions about the distribution function (and its parameters) is what allows us to make conclusions in the real world.

Random variables are a particular method of allowing us to translate probability statements from one sample space to another. In the case

of the random variable, the "other" sample space is some subset of the real numbers. Statements about the probability of certain events are best characterized in the form of functions (as is seen above). This is why the idea of a random variable exists and is so widely studied.

Definition: A distribution *family* refers to a collection of pmfs, which can assume some set of parameter values.

Example: Recall the binomial pmf from above. It's a family of pmfs, with a unique pmf corresponding to each choice of $n = 1, 2, \cdots$ and $p \in [0, 1]$. Note how the parameter values must belong in these particularly defined sets in order for the binomial pmf to be considered a "valid" pmf, i.e. it actually corresponds to binomial probabilities.

Definition: The *cumulative distribution function* $F(x)$ of a discrete random variable X with probability mass function $f(x)$ is

$$F(x) := P(X \leq x) = \sum_{t \leq x} f(t)$$

Note: The cdf can be used for all real numbers x. Even if x is outside of the range of the random variable X, we can still evaluate the cdf at this point (in contrast to when we take $f_X(x) = 0$ for x outside of the range of X).

Example: Consider a discrete random variable X which can take values $0, 1, 2, \cdots$ (like a binomial r.v.). We evaluate the cdf simply by summing the probabilities of all the smaller values:

$$F(3) = P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3)$$

If we had a non-integer value of x, we can still evaluate the cdf:

$$F(4.05) = P(X \leq 4.05) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$$

1.12 Continuous Random Variables

Definition: A continuous random variable is a random variable X with an uncountable range.

This means that the set of potential observed values of X belong to an uncountable set, some subset the real numbers $\mathbb{R}$. Using the real numbers means we can use calculus to formalize the notion of a continuous random variable, as was done in the discrete case previously.

Definition: The function $f(x)$ is a *probability density function* (pdf) for a continuous random variable X if:

$$P(a < X < b) = \int_a^b f(x)dx$$

or more generally,

$$P(X \in A) = \int_A f(x)dx$$

As was the case for the discrete case, P being a probability function leads to the following consequences:

(1) $f(x) \geq 0$ for all $x \in \mathbb{R}$
(2) $\int_{-\infty}^{\infty} f(x)dx = 1$

Definition: The *cumulative distribution function* $F(x)$ of a continuous random variable X with density function $f(x)$ is

$$F(x) := P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

Properties:

$$P(a < X < b) = F(b) - F(a)$$

$$f(x) = \frac{dF(x)}{dx}$$

As with a discrete random variable, we can interpret the probability distribution (whether it be the pdf/pmf or cdf) as a way to completely describe how "likely" it is to observe a particular value of our random variable. For example, recall the third property of a pdf:

$$P(a < X < b) = \int_a^b f(x)dx$$

This means we can make probability statements about our random variable if we know its distribution f.

Note that the probability of observing a particular value of a continuous random variable X is 0:

$$P(X = x) = 0 \quad \forall x$$

This is because the distribution is continuous. We instead consider the probability of X belonging to some set.

1.13 The Four Most Significant Uses of a Density Function

To find the probability of $X \in A$ for some set A, one can use the density function in the following way:

$$P(X \in A) = \int_A f_X(x)dx$$

This is essentially the definition of a probability density function.

To find the *expected value* (denoted with E) of some function h of the random variable X, then we have that

$$E(h(X)) = \int h(x)f_X(x)dx$$

(the expected value will be covered in more detail later).

Knowing the probability density function also allows for random sampling via simulation. We can generate data for which we know the data generating process.

Lastly, knowing the form of the density function allows us to reduce the problem of estimating the entire density function to estimating paramters

within that density. This is introduced below and will be discussed in more detail in chapter 2.

1.14 Parameters

Consider the normal distribution, which has the following distribution function:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

μ and σ are *parameters* of the distribution. The above represents a distribution *family*, from which we choose particular values of μ and σ to arrive at a particular pdf, which is a function of the variable x. Most particular forms of distributions we consider will be expressed as an entire family of distributions, for which there are infinity many choices of parameters to choose from.

Definition: A *random sample* is an *iid* collection of random variables (iid stands for "independent and identically distributed". See definition below).

Definition: Identically Distributed: two random variables X and Y are identically distributed if they have the same distribution function ($f_X(x) = f_Y(x)$ for all x)

Example: Coin flips: Consider the random variable X which gives the number of heads in 5 coin flips. Our random sample of this random variable could be $5, 3, 2, 1, 4, 2, 1, 3$.

Example: A random sample of Bernoulli r.v.'s: $(0, 1, 0, 1, 1, 1)$

1.15 Histograms and Distributions

A histogram is a useful tool because it serves as a rough estimate for the distribution function. That is, the histogram, when plotted next to the distribution function, should generally have the same shape. Note that the histogram is obtained by using an observed sample, and the distribution function is purely theoretical.

Recall the normal distribution from above. Regardless of the choice of parameters μ and σ, the distribution will have the familiar "bell" shape,

symmetric about its mean. Because of this, a very rough method for determining if you have normal data is simply to plot the histogram and see if the histogram also has this shape. This is because of the fact that the histogram serves as an estimate for the distribution. Given infinitely many observations (and bins for our histogram), it should look increasingly similar the pdf when plotted.

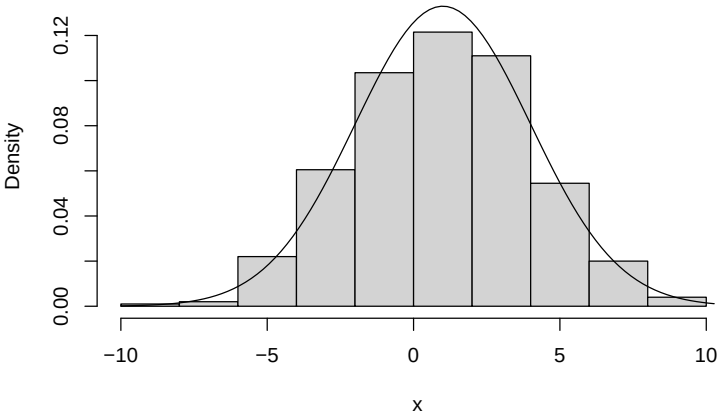


Figure 1: Histogram of a random sample of size 1000 from a normal distibution and the density function of the normal distribution.

1.16 Calculating Probabilities

The key feature of probability density functions, probability mass functions, and cumulative distribution functions is that one can use them to calculate the probability of any event related to the random variable which they describe. Let the random variable be denoted X, and recall the following general relations:

$$P(X \in A) = \sum_{x \in A} P(X = x) \qquad \text{in the discrete case}$$

$$P(X \in A) = \int_A f(x)dx \qquad \text{in the continuous case}$$

Observed values of a random variable X always assume values on some

subset of $\mathbb{R}$. For our purposes, this means that $P(X \in A)$ is expressed as X belonging to one or more *intervals*, in the continuous case:

$$A := [a, b]$$
$$P(X \in A) = \int_A f(x)dx = \int_a^b f(x)dx$$

Let the intervals $[a, b]$ and $[c, d]$ to be *disjoint* sets. Then, define

$$A := [a, b] \cup [c, d]$$
$$P(X \in A) = \int_A f(x)dx = \int_a^b f(x)dx + \int_c^d f(x)dx$$

In the discrete case, A generally has one of two forms:

$$A := \{a_1, a_2, \cdots, a_n\}, \qquad \text{where each } a_n \in \mathbb{Z}$$

in which case we have:

$$P(X \in A) = \sum_{x \in A} P(X = x) = \sum_{i=1}^{n} P(X = a_i)$$

Or we could have:

$$A := \{a_1, a_2, \cdots\}$$

where in the above, the lack of an a_n at the end of the set indicates that A has a (countably) infinite number of elements $a_i \in \mathbb{Z}$. Therefore,

$$P(X \in A) = \sum_{x \in A} P(X = x) = \sum_{i=1}^{\infty} P(X = a_i)$$

1.17 Examples

In the following examples, the functions for the pdf/pmf/cdf are all examples of well known *distribution families*, which are used due to the properties they possess for modeling certain situations. We will discuss these properties later, but for now all that matters is that they are a valid distribution function, i.e. they obey the properties given previously. Essentially, they must be non-negative and sum/integrate to 1.

Example: Consider X with the following pdf:

$$f_X(x) = \frac{1}{2}e^{-x/2} \qquad x \geq 0$$

Calculate $P(X \in (0, 1])$.

Solution:

$$P(X \in (0,1)) = \int_0^1 \frac{1}{2}e^{-x/2} = -e^{-x/2}|_0^1 = 1 - e^{-1/2} \approx 0.394$$

Example: Consider X with the following cdf:

$$F_X(x) = 1 - e^{-x/2} \qquad x \geq 0$$

Calculate $P(X \in (0, 1])$.

Solution:

$$\begin{aligned} P(X \in [0,1]) &= P(X \leq 1) - P(X \leq 0) \\ &= F_X(1) - F_X(0) \\ &= (1 - e^{-1/2}) - (1 - 1) = 1 - e^{-1/2} \end{aligned}$$

Note that The random variable X with the pdf in the first example has the cdf in the second example, so it stands to reason these probabilities should be the same, whether they were calculated using the pdf or cdf:

$$\frac{d}{dx}F_X(x) = \frac{d}{dx}(1 - e^{-x/2}) = \frac{1}{2}e^{-x/2} = f_X(x)$$

Example: Consider X with the following pdf:

$$f_X(x) = \frac{1}{2}xe^{-x}, \qquad x \geq 0$$

Calculate $P(X \in (0,1) \cup (2,3))$.

Solution:

$$\begin{aligned} P(X \in (0,1) \cup (2,3)) &= \int_0^1 \frac{1}{2}xe^{-x} + \int_2^3 \frac{1}{2}xe^{-x} \\ &= \frac{1}{2}\left(\frac{e-2}{e} + \frac{3e-4}{e^3}\right) \approx 0.236 \end{aligned}$$

(you need to integrate by parts.)

Example: Consider X with the following pdf:

$$f_X(x) = 6x(1-x) \qquad 0 \leq x \leq 1$$

Calculate $P(X < \frac{1}{2})$.

Solution: Note that $X < \frac{1}{2} \implies X \in (-\infty, \frac{1}{2})$. Therefore:

$$P(X < \frac{1}{2}) = \int_{-\infty}^{\frac{1}{2}} 6x(1-x) = \int_0^{\frac{1}{2}} 6x(1-x)$$

Notice that the second equality comes from the fact that $f_X(x)$ is only defined for x between 0 and 1, and we consider it to be zero everywhere else in $\mathbb{R}$.

$$P(X < \frac{1}{2}) = \int_0^{\frac{1}{2}} 6x(1-x) = 6(x^2/2 - x^3/3)|_0^{\frac{1}{2}} = \frac{1}{2}$$

Example: Consider X with the following pmf:

$$f_X(x) = \frac{e^{-3}3^x}{x!}$$

Calculate $P(X \in \{1, 4\})$.

Solution:

$$P(X \in \{1,4\}) = P(X = 1) + P(X = 4) = f_X(1) + f_X(4)$$
$$= 3e^{-3} + \frac{81e^{-3}}{4!} \approx 0.317$$

Example: Consider X with the following cdf:

$$F_X(x) = 1 - (0.7)^{\lfloor x \rfloor}, \qquad x = 1, 2, 3, \cdots$$

where $\lfloor \cdot \rfloor$ is the floor function.

Calculate $P(X \leq 5.7)$.

Solution:

$$P(X \leq 5) = F_X(5.7) = 1 - (0.7)^{\lfloor 5.7 \rfloor} = 1 - 0.7^5 \approx 0.832$$

Example: The *standard normal* distribution comes from the selection of parameters $\mu = 0$ and $\sigma^2 = 1$. This results in the following pdf:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Suppose we want to calculate $P(X \leq 1)$. Then we have

$$P(X \leq 1) = \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \approx 0.841$$

This is done computationally. We denote the CDF of a normal distribution as $\Phi(x)$. The integral has no closed-form solution and must be calculated using computational methods.

1.18 Mathematical Expectation

As previously mentioned, the expected value of a random variable, or functions thereof, is a quantity with which we will have great interest, so we begin exploring this concept now.

Definition: The *expected value* of a random variable X with pmf/pdf $f_X(x)$ (denoted $E(X)$) is defined to be

$$E(X) = \Sigma_x x \cdot f_X(x) \qquad \text{if } X \text{ is discrete}$$
$$E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)dx \qquad \text{if } X \text{ is continuous}$$

Note: We often denote $E(X) = \mu_x$. The expected value of X is sometimes referred to as the "mean of X".

Example: Consider a random variable X with pmf:

$$f_X(x) = \left\{ \begin{array}{ll} 0.5, & \text{if } x = 1 \\ 0.1, & \text{if } x = 2 \\ 0.4, & \text{if } x = 3 \\ 0, & \text{otherwise} \end{array} \right\}$$

To find $E(X)$, using the definition, we have:

$$E(X) = \sum_x x \cdot f_X(x) = 1 \cdot 0.5 + 2 \cdot 0.1 + 3 \cdot 0.4 = 1.9$$

Example: Consider a random variable X with pdf:

$$f_X(x) = e^{-x} \qquad 0 \leq x \leq \infty$$

Find $E(X)$ (here we use integration by parts).

$$\begin{aligned} E(X) &= \int_0^\infty xe^{-x} \\ &= -xe^{-x} + \int_0^\infty e^{-x}dx \\ &= -xe^{-x} - e^{-x}|_0^\infty \\ &= 0 - 0 - 0 + 1 = 1 \end{aligned}$$

Theorem: (Law of the Unconscious Statistician) Given a random variable, X, and a function, $h(X)$, the expected value of $h(X)$ is as follows:

$$E(h(X)) = \Sigma_x h(x) \cdot f_X(x) \qquad \text{if } X \text{ is discrete}$$
$$E(h(X)) = \int_{-\infty}^{\infty} h(x) \cdot f_X(x)dx \qquad \text{if } X \text{ is continuous}$$

if X is discrete or continuous, respectively.

Remark: Consider a random variable X. Suppose we have some function h which defines a one-to-one mapping between values of X, and values of another random variable Y, i.e. we write $Y = h(X)$. Note that the function h maps from $\mathbb{R}$ to $\mathbb{R}$. As a result, Y itself is a random variable.

Here are some more properties of the expected value:

Theorem: $E(g(X) + h(X)) = E(g(X)) + E(h(X))$

Proof: This follows immediately from the definition of an expected value, and properties of sums/integrals.

Theorem: Let a and b be constants, then

$$E(aX + b) = aE(X) + b$$

Proof:

$$E(aX + b) = \int (ax + b)f(x)dx = a\left(\int f(x)dx\right) + b = aE(x) + b$$

Theorem: For random variables $X_1, \cdots, X_n$,

$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n)$

The proof of the above theorem will be given later, using joint distributions.

Example: Consider X to be a random variable with the following pmf:

$$f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x = 0, 1, 2, \cdots$$

Define another random variable $Y = h(X) := X^2$. Then we have that the pdf of Y is:

$$f_Y(y) = \frac{e^{-\lambda}\lambda^{\sqrt{y}}}{(\sqrt{y})!} \qquad y = 0, 1, 4, 9, \cdots$$

In the case of discrete random variables, we can make the substitution like we have above, where we simply solved for X in terms of Y using the relationship between the two (this is NOT true in the continuous case, which we will go over later).

Using the the above theorem, we have that $E(Y) = E(h(X)) = E(X^2)$ is:

$$\begin{aligned} E(h(X)) &= \Sigma_x h(x) \cdot f_X(x) \\ &= \sum_{i=0}^{\infty} x^2 \frac{e^{-\lambda}\lambda^x}{x!} \\ &= \lambda(\lambda + 1) \end{aligned}$$

The details of how to show that the sum converges to $\lambda(\lambda + 1)$ have been omitted here, but are shown later. Finding the value of the sum requires some tricks, but is doable.

On the other hand, if we use the pdf of Y, we have:

$$\begin{aligned} E(h(X)) = E(Y) &= \Sigma_y y \cdot f_Y(y) \\ &= \sum_{y \text{ s.t. } y \text{ is a perfect square}} y \frac{e^{-\lambda}\lambda^{\sqrt{y}}}{\sqrt{y}!} \\ &= \lambda(\lambda+1) \end{aligned}$$

The above sum is significantly more difficult to work with, which shows the utility of the law of the unconscious statistician.

1.19 Variance and Covariance

Definition: The *variance* of a random variable X (denoted $V(X)$ or σ^2), is given by $V(X) = E((X - E(X))^2)$

Applying the definition of the expected value, we have that (in the continuous case):

$$V(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f_X(x) dx \qquad \text{if } X \text{ is continuous}$$

with a similar formula using a sum for the discrete case.

Note: The value of $E(X)$ is a constant. To emphasize this, we often write $E(X) = \mu_x$.

Definition: The square root of the variance is called the standard deviation (denoted σ).

Note that the quantity $X - \mu_x$ is itself a random variable, which represents the deviation of X from its mean. If X often assumes values far from its mean, it will have a large variance. If it stays near its mean, it will have a small variance.

Theorem: Variance Formula: $V(X) = E(X^2) - \mu_x^2$

Proof:

We will show the proof for the continuous case, which is analogous to the proof in the discrete case.

$$\begin{aligned} V(X) &= \int (x-\mu_x)^2 f_X(x)dx \\ &= \int (x^2 - 2\mu_x x + \mu_x^2) f_X(x)dx \\ &= \int x^2 f(x)dx - 2\mu_x \int x f(x)dx + \mu_x^2 \int f(x) \\ &= E(X^2) - 2\mu_x^2 + \mu_x^2 \\ &= E(X^2) - \mu_x \end{aligned}$$

Usually, in the process of calculating the variance using the definition, you will inevitably go through the above process when calculating the integral. We nearly always just use the variance formula to calculate variances.

Example: Consider a discrete r.v. X with the following pmf:

$$f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x = 0, 1, 2, \cdots$$

To find the variance using the variance formula, we must find both $E(X)$ and $E(X^2)$.

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=1}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} \\ &= \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-1)!} = \lambda \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} \\ &= \lambda \sum_{y=0}^{\infty} y\frac{e^{-\lambda}\lambda^y}{y!} = \lambda \end{aligned}$$

where we define $y = x - 1$, and the final sum is equal to 1 by nature of this being a valid pmf.

In an earlier example, we claimed that $E(X^2) = \lambda(\lambda + 1)$. We will show this by first noting the following:

$$E(X(X-1)) = E(X^2) - E(X) \implies E(X^2) = E(X(X-1)) + E(X)$$

which we get using properties of expectation. Therefore, we can find $E(X(X-1))$ and use that result to find $E(X^2)$. Using the law of the unconscious statistician:

$$\begin{aligned} E(X(X-1)) &= \sum_{x=0}^{\infty} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=2}^{\infty} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} \\ &= \sum_{x=2}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-2)!} = \lambda^2 \sum_{x=2}^{\infty} \frac{e^{-\lambda}\lambda^{x-2}}{(x-2)!} \\ &= \lambda^2 \sum_{y=0}^{\infty} y\frac{e^{-\lambda}\lambda^y}{y!} = \lambda^2 \end{aligned}$$

where we define $y = x - 2$, and the final sum is equal to 1 by nature of this being a valid pmf.

This leaves:

$$E(X^2) = E(X(X-1)) + E(X) = \lambda^2 + \lambda$$

So we caclulate the variance as follows:

$$V(X) = E(X^2) - E(X)^2 = \lambda + \lambda^2 - \lambda^2 = \lambda$$

This is a particular situation in which the expected value of the random variable X is equal to its variance.

Note: The law of the unconscious statistician applies to variances as well, in that we have

$$V(g(X)) = E(g(X)^2) - (E(g(X)))^2$$

which is true simply because the variance is defined in terms of the expected value.

1.20 Moments and Moment Generating Functions

Definition: For a random variable X, $E(X^k)$ is called the k^{th} *moment* of X.

Definition: The moment generating function $M_X(t)$ (a function of t) of a random variable X is defined to be:

$$M_X(t) = E(e^{tX})$$

if it exists.

Property:

$$\frac{d^k}{dt^k} M_X(t)|_{t=0} = E(X^k)$$

Example: This example will showcase how the above property can be useful to calculate moments.

Consider a binomial random variable X, which is a discrete random variable with the following pmf:

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, \cdots, n$$

Suppose we want to calculate the variance of X. Then:

$$V(X) = E(X^2) - E(X)^2$$

We can calculate this if we know the first moment ($E(X)$) and the second moment ($E(X^2)$). Let's calculate them using the moment generating function.

$$\begin{aligned} M_X(t) = E(e^{tX}) &= \sum_{x=0}^{n} e^{tx}\binom{n}{x}p^x(1-p)^{n-x} \\ &= \binom{n}{x}(pe^t)^x(1-p)^{n-x} \\ &= (1-p+pe^t)^n \end{aligned}$$

First moment:

$$\begin{aligned} M'_X(t) &= n(1-p+pe^t)^{n-1}(pe^t) \\ E(X) = M'_X(0) &= n(1-p+p)^{n-1}(p) = np \end{aligned}$$

Second moment:

$$\begin{aligned} M''_X(t) &= n(n-1)(1-p+pe^t)^{n-2}(pe^t)^2 + (pe^t)(n)(1-p+pe^t)^{n-1} \\ E(X^2) &= M''_X(0) \\ &= n(n-1)(1-p+p)^{n-2}(p^2) + np(1-p+p)^{n-1} \\ &= n(n-1)p^2 + np \\ &= n^2p^2 - np^2 + np \end{aligned}$$

Putting these together:

$$\begin{aligned} V(X) = E(X^2) - E(X)^2 &= n^2p^2 - np^2 + np - n^2p^2 \\ &= np - np^2 \end{aligned}$$

1.21 Covariance and Correlation

Definition: The covariance of two random variables X and Y (with $E(X) = \mu_x$ and $E(Y) = \mu_y$) is given by

$$\text{cov}(X, Y) = E((X - \mu_x)(Y - \mu_y))$$

We sometimes denote $\text{cov}(X, Y) = \sigma_{xy}$

Theorem:

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y)$$

Proof: Exercise.

Definition: The correlation of random variables X and Y is defined to be

$$\rho_{xy} = \frac{\text{cov}(X, Y)}{\sqrt{V(X)V(Y)}}$$

A consequence of this definition is that $-1 \leq \rho_{xy} \leq 1$.

1.22 Joint Distributions

Definition An n-dimensional *random vector* is a function from a sample space S to $\mathbb{R}^n$.

Note: Often, we create a random vector by assembling usual one-dimensional random variables into a vector. i.e. if X and Y are random variables, then we can define a random vector $Z = (X, Y)$, which represents the function:

$$Z(\omega) = (X(\omega), Y(\omega))$$

where ω is some element of the sample space S.

Definition: The function $f(x, y)$ is a *joint probability density function* (joint pdf) of continuous random variables X and Y if for every $A \in \mathbb{R}^2$,

$$P((X, Y) \in A) = \int \int_A f(x, y) dx dy$$

As before, P being a valid probability function implies that f has the following properties:

$$f(x,y) \geq 0 \qquad \text{for all } (x,y)$$
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$$

The following is a more general definition, extended to an n-dimensional random vector:

Definition: A function $f : \mathbb{R}^n \to \mathbb{R}$ is called a *joint probability density function* of the continuous random vector $(X_1, \cdots, X_n)$ if, for every $A \subset \mathbb{R}^n$,

$$P((X_1, \cdots, X_n) \in A) = \int \cdots \int_A f(x_1, \cdots, x_n)dx_1 \cdots dx_n$$

All of the above definitions are a natural extension of our definitions of distribution functions for random variables to higher dimensions. We are often concerned with functions of multiple random variables, say $Z = g(X, Y)$, and knowing the joint distribution function allows us to more easily work with such random variables Z. It also allows us to establish certain properties, such as independence of random variables, which is discussed later in this lecture.

1.23 Marginal and Conditional Density Functions

For the following definitions, we will consider the case of a bivariate random vector (X, Y), though these results can be generalized to an n-dimensional random vector.

Definition: The *marginal distribution* of X, given a joint $f(x, y)$, is given (in the continuous case) by

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

Likewise, the marginal distribution of Y is given by

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

Note: By integrating with respect to y, we remove all dependence on the value of y, which is why the marginal pdf is only a function of x. Also note that the marginal *is* the pdf of X, without considering values of Y, and vice versa.

Knowing the joint distribution also gives us all of the information we need to evaluate probabilities relating to both of the individual random variables X and Y.

Example: Let (X,Y) be the continuous random vector with the following joint pdf:

$$f(x,y) = x + y \qquad \text{for } x,y \in [0,1]$$

Suppose we want to evaluate $P(0.5 < X < 1)$. You can think of this as the probability that X is in the specified range, and Y is "anything". This can be done directly using the joint and evaluating the following integral:

$$P(0.5 < X < 1) = \int_0^1 \int_{0.5}^1 (x+y)dxdy = \frac{5}{8}$$

However, this is equivalent to first calculating the marginal of X:

$$f_X(x) = \int_0^1 (x+y)dy = x + \frac{1}{2}$$

and then using the marginal to evaluate the probability as usual:

$$P(0.5 < X < 1) = \int_{0.5}^1 f_X(x)dx = \int_{0.5}^1 x + \frac{1}{2}dx = \frac{5}{8}$$

Example: Let (X,Y) be the continuous random vector with the same joint pdf as above. Find $E(h(X))$.

Again, we can use either the joint or the marginal in an equivalent fashion to accomplish this. Either we evaluate the following integral using the joint:

$$E(h(X)) = \int_0^1 \int_0^1 h(x)(x+y)dxdy$$

Or we use the marginal, $f_X(x)$, and evaluate this equivalent integral:

$$E(h(X)) = \int_0^1 h(x)f_X(x)dx = \int_0^1 h(x)\left(x + \frac{1}{2}\right)dx$$

The joint gives all of the information that the marginals give, and more. Knowing the joint allows us to calculate things like covariance, which we cannot determine simply from the marginal distributions, because it requires information on how X and Y vary together. We will see more on this in the next lecture.

Definition: Let X and Y be two random variables. The *conditional distribution* of Y given that $X = x$ is

$$f(y|x) = \frac{f(x,y)}{f_X(x)}$$

supposing that $f_X(x) > 0$ for all x.

Likewise,

$$f(x|y) = \frac{f(x,y)}{f_Y(y)}$$

Notation: Here we introduce notation that we haven't seen before in order to express conditional probabilities for random variables. Let A and B be subsets of the range of random variables X and Y, respectively. Then,

$$P(X \in A|Y = y)$$

is used to express the probability of the following set of outcomes in the sample space:

$$C := \{\omega : Y(\omega) = y\}$$
$$P(X \in A | Y \in B) = P(\{\omega \in C : X(\omega) \in A\})$$

For this class, you don't need to worry about the details of the above definition. Intuitively, we can think of this as "reducing" our sample space to contain only elements ω such that $Y(\omega) = y$ (i.e. $Y = y$). Then we calculate the probability of $X \in A$, within this reduced sample space.

Example: Evaluating probabilities using a conditional distribution.

Consider the joint distribution from earlier:

$$f(x, y) = x + y \qquad \text{for } x, y \in [0, 1]$$

Suppose we want to find $P(Y < 0.5 | X = 0.1)$. We have that the conditional distribution $f(y|x)$ is:

$$f(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{x + y}{x + \frac{1}{2}}$$

Letting $x = 0.1$, we have

$$f(y|X = 0.1) = \frac{y + 0.1}{0.1 + 0.5} = \frac{y + 0.1}{0.6}$$

Now, we can evaluate the desired probability by integrating:

$$P(Y < 0.5 | X = 0.1) = \int_0^{0.5} \frac{y + 0.1}{0.6} dy = \frac{7}{24} \approx 0.292$$

Compare this to:

$$P(Y < 0.5) = \int_{0.5}^{1} \int_0^1 (x + y) dx dy = \frac{5}{8} = 0.625$$

So the probability of Y assuming a value less than 0.5 became significantly less probable when we conditioned on the fact that X was small ($X = 0.1$). This problem gives some intuition as to how random variables X and Y can vary together, and this information is contained in their joint.

Definition: The conditional expected value of a random variable X given that another random variable $Y = y$ is defined to be:

$$E(X|Y = y) = \int_{-\infty}^{\infty} xf(x|y)dx$$

i.e. we simply replace the marginal density $f_X(x)$ with the conditional density $f(x|y)$.

Example: Consider the same example from above. Evaluate $E(Y|X = 0.1)$.

$$E(Y|X = 0.1) = \int_0^1 yf(y|X = 0.1)dy = \int_0^1 \frac{y^2 + 0.1y}{0.6} = \frac{23}{36} \approx 0.639$$

Meanwhile,

$$E(Y) = \int_0^1 \int_0^1 y(x + y)dxdy = \frac{7}{12} \approx 0.583$$

1.24 Independent Random Variables

Definition: Two random variables X and Y are said to be independent if

$$f(x, y) = f_X(x)f_Y(y)$$

That is, their joint is equal to the product of their marginals.

Example: Independence is a way to describe whether or not two random variables "vary together". As was showcased throughout this lecture, X and Y with the following joint pdf:

$$f(x, y) = (x + y) \qquad \text{for } x, y \in [0, 1]$$

clearly vary together. It is also clear that the above joint distribution cannot be factored into the two marginal distributions. We have that:

$$f_X(x)f_Y(y) = \left(x + \frac{1}{2}\right)\left(y + \frac{1}{2}\right) \neq x + y$$

so X and Y are dependent random variables.

It easy to construct independent random variables. Simply take two valid pdfs for X and Y and multiply them to create the joint. For example:

$$f(x, y) = e^{-x}e^{-y} \qquad \text{for } x, y > 0$$

It is extremely common to consider the joint distribution of independent random variables and we will see this again many times.

1.25 Expectation Involving Multiple Random Variables

Definition: Consider a continuous bivariate random vector (X, Y) with joint distribution $f(x, y)$. Then we define:

$$E(h(X, Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x, y)f(x, y)dxdy$$

Note: In particular, consider $E(XY)$:

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x, y)dxdy$$

We will use this in the following example and theorem.

Example: Covariance and correlation examples. Finding E(XY) using the defn above.

Consider continuous random variables X and Y with the following joint pdf:

$$f(x, y) = \frac{2}{5}(2x + 3y) \qquad x, y \in [0, 1]$$

Find $\text{cov}(X, Y)$.

Solution:

Recall the following identity:

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y)$$

So we must calculate the three quantities used in the above expression:

$$f_X(x) = \int_0^1 \frac{2}{5}(2x + 3y)dy = \frac{4}{5}x + \frac{3}{5}$$

Using the above,

$$\begin{aligned} E(X) &= \int_0^1 x f_X(x)dx \\ &= \int_0^1 \frac{4}{5}x^2 + \frac{3}{5}xdx \\ &= \frac{17}{30} \end{aligned}$$

Similarly,

$$f_Y(y) = \int_0^1 \frac{2}{5}(2x + 3y)dx = \frac{6}{5}y + \frac{2}{5}$$

$$E(Y) = \int_0^1 y\left(\frac{6}{5}y + \frac{2}{5}\right)dy = \frac{3}{5}$$

Now for $E(XY)$:

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy \\ &= \int_0^1\int_0^1 \frac{2}{5}xy(2x+3y)dxdy \\ &= \frac{1}{3} \end{aligned}$$

Which leaves

$$\text{cov}(X,Y) = E(XY) - E(X)E(Y) = \frac{1}{3} - \frac{17}{30}\cdot\frac{3}{5} = \frac{-1}{150}$$

Theorem: If X and Y are independent random variables, then $E(XY) = E(X)E(Y)$.

Proof:

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} xf_X(x)dx \int_{-\infty}^{\infty} yf_Y(y)dy = E(X)E(Y) \end{aligned}$$

Note how this can be easily extended to n independent random variables.

Recall the following result that I stated in a previous lecture without proof. Now we can prove this.

Theorem: For random variables $X_1, \cdots, X_n$,

$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n)$

Proof: Let $f(x_1, \cdots, x_n)$ be the joint distribution of the random variables $X_1, \cdots, X_n$. Then,

$$\begin{aligned} E(X_1 + \cdots + X_n) &= \int \cdots \int (x_1 + \cdots + x_n) f(x_1, \cdots, x_n) dx_1 \cdots dx_n \\ &= \sum_{i=1}^{n} \int \cdots \int x_i f(x_1, \cdots, x_n) dx_1 \cdots dx_n \\ &= \sum_{i=1}^{n} \int x_i f_{X_i}(x_i) dx_i \end{aligned}$$

where $f_{X_i}(x_i)$ is the marginal distribution of X_i. By definition, we have that

$$\int x_i f_{X_i}(x_i) dx_i = E(X_i)$$

which leaves

$$E(X_1 + \cdots + X_n) = \sum_{i=1}^{n} E(X_i)$$

■

1.26 Common Distribution Families: Discrete Distributions

We have previously discussed the idea of a *distribution family*, which is a collection of distributions which can assume parameter values on some set. These parameter values are related to properties of the distribution, such as the mean and variance, examples of which have been previously seen. Now, we will go through some of the most important distribution families, derive these properties, and comment on the situations in which they are relavent in the context of modeling an experiment.

Notation: Each common distribution family has an abbreviation. In general, lets say a distribution family is abbreviated *fam*. Then we would write:

$$X \sim \text{fam}(\theta_1, \cdots, \theta_p)$$

to denote that the random variable X has a pdf/pmf which is a member of the distribution family "fam", and which has parameter values $\theta_1, \cdots, \theta_p$. Different distribution families will have varying numbers of parameter values, which will be represented with various letters/symbols (not necessarily θ).

1.27 Bernoulli Random Variables

Definition: A random variable X is called a Bernoulli random variable if X is discrete with the following pmf:

$$f_X(x) = p^x(1-p)^{1-x}, \qquad x \in \{0, 1\}$$

Using the notation introduced above, we write $X \sim \text{bern}(p)$. Clearly, p is the only parameter in the above pmf.

This can be thought of as a coin flip with probability of 1 equal to p, and probability of 0 equal to $1 - p$.

More generally, an experimental result in which we we observe one of two possible outcomes, can be modeled as a Bernoulli. We interpret the data from such an experiment as a random sample of independent and identically distributed Bernoulli r.v.s.

Exercise: Find expected value and variance of Bernoulli RV

1.28 Geometric Random Variables

Definition: A random variable X is called a Geometric random variable if X is discrete with the following pmf:

$$f_X(x) = (1-p)^{x-1}p \qquad x = 1, 2, 3, \cdots$$

where we write $X \sim \text{geom}(p)$

A geometric distribution describes a random variable X which is equal to the number of trials required to reach a "success". p is the probability

of success on any particular trial, and each trial is independent with the same probability of success (think X = number of coin flips required to get at least one heads).

We can derive the geometric distribution using a Bernoulli random variable:

Consider $X \sim \text{bern}(p)$, and we sample X multiple times. Consider the possible ways of obtaining a 1:

- $X = 1$ on the first sample, which has probability p.
- $X = 0$ on the first sample, with probability $1 - p$. Then on the second sample, $X = 1$, with probability p. The overall probability is $p(1 - p)$.
- Likewise, if $X = 0$ in the first two samples, and $X = 1$ on the third, then the probability is $p(1 - p)^2$.

Continuing this process yields the geometric pmf.

Exercise: Derive the expected value and variance of the geometric.

Exercise: Find the cdf of a geometric distribution.

Hint:

$$F(x) = \sum_{t=1}^{x} (1-p)^{t-1} p = \frac{p}{1-p} \sum_{t=1}^{x} (1-p)^t$$

1.29 Binomial Random Variables

Critical to understanding the binomial distribution is the binomial theorem, which we have already appealed to multiple times in previous lectures.

Theorem: (The Binomial Theorem)

$$(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x}$$

Note that if we allow $a = p$, $b = 1 - p$, where p is between 0 and 1, we have the property that:

$$(a+b)^n = (p+1-p)^n = 1^n = 1 = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x}$$

This means that if we consider a random variable X which has range $\{1, 2, \cdots, n\}$, then the pdf as defined by the quantity inside of the sum obeys the property of the pdf that it must sum to 1 over the range of X. Furthermore, each term of the sum is nonnegative, meaning that a pdf defined in that way would be a valid pdf. This is how we arrive at the binomial random variable:

Definition: A random variable X is called a Binomial random variable if X is discrete with the following pmf:

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, \cdots, n$$

where we write $X \sim \text{binom}(n, p)$.

Note how the binomial distribution has two parameters, n and p, rather than just one like the Bernoulli and the Geometric. Essentially, the binomial can be thought of as the "number of successes in n Bernoulli trials". In fact, they are often thought of as interchangeable. For example, consider flipping 5 coins, and only concerned with the number of heads in those 5 flips. Then define:

$$\begin{aligned} X &\sim \text{binom}(5, 0.5) \\ Y_1, Y_2, \cdots, Y_5 &\sim \text{bern}(0.5) \end{aligned}$$

Clearly, an observation of X is equivalent to an observation of $Y_1, \cdots, Y_5$ (still supposing of course that we are only concerned with the number of heads, not with the order in which they appear). The next example is related to this fact.

Example: Consider $Y_1, Y_2 \sim \text{bern}(0.5)$. By independence, we can write the joint distribution function $f(y_1, y_2)$ as follows:

$$\begin{aligned} f(y_1, y_2) &= p^{y_1}(1-p)^{1-y_1} p^{y_2}(1-p)^{1-y_2} \\ &= p^{y_1+y_2}(1-p)^{2-y_1-y_2} \qquad y_1, y_2 \in \{0, 1\} \end{aligned}$$

Consider the situation where we only consider how many 1's we observe. i.e. $f(1,0)$ and $f(0,1)$ refer to equivalent situations. Lets enumerate all the probabilities:

$$\begin{aligned} P(\text{two 1s}) &= f(1,1) = p^2 \\ P(\text{one 1}) &= f(1,0) + f(0,1) = p(1-p) + (1-p)p = 2p(1-p) \\ P(\text{zero 1s}) &= f(0,0) = (1-p)^2 \end{aligned}$$

This is, as expected, the pmf of a bin$(2, p)$ random variable.

Example: To make the above example more formal, we need an additional fact:

Theorem: If the moment generating functions of random variables X and Y are such that $M_X(t) = M_Y(t)$, then X and Y have the same distribution.

Lets apply this to a random sample of n Bernoulli random variables:

$$X_1, \cdots, X_n \sim \text{bern}(p)$$

Define $Z = \sum_{i=1}^n X_i$. Then,

$$M_Z(t) = E(e^{tZ}) = E(e^{t\sum X_i}) = E(e^{tX_1} \cdots e^{tX_n}) = E(e^{tX_1}) \cdots E(e^{tX_n})$$

where the last line follows from independence. Furthermore, for $X \sim$ bern(p),

$$E(e^{tX}) = \sum_{x=0}^{1} e^{tx} p^x (1-p)^{1-x} = \sum_{x=0}^{1} (pe^t)^x (1-p)^{1-x} = (1-p) + pe^t$$

which leaves

$$M_Z(t) = (1 - p + pe^t)^n$$

which is the moment generating function of a binomial random variable with parameters n and p. Therefore, by the above theorem, it is necessarily the case that:

$$Z = \sum_{i=1}^{n} X_i \sim \text{binom}(n, p)$$

i.e. a binomial random variable is the sum of n Bernoulli random variables.

Exercise: Derive expected value and variance of binomial rvs

1.30 Poisson Random Variables

A random variable X follows a *Poisson* Distribution if it has the following pdf:

$$f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, \cdots$$

Some examples of Poisson Processes:

- Telephone calls received per hour by an office.
- The number of days school is closed due to snow during the winter.
- The number of games postponed due to rain during a baseball season.
- The number of bacteria in a given culture.
- The number of typing errors per page.

That is, a Poisson random variable X will represent the number of "outcomes" which occur in a Poisson experiment.

A Poisson random variable has the property that both its expected value and variance are equal to λ. Therefore, we can think of λ as the average number of occurrences in the Poisson experiment. This was derived as an example in an earlier lecture.

1.31 Common Distribution Families: Continuous Distributions

We will now consider the continuous distributions, which are defined on some uncountable subset of $\mathbb{R}$. There are three such subsets which are of particular importance:

1) The entire real line, $\mathbb{R} = (-\infty, \infty)$
2) The positive real numbers, $\mathbb{R}^+ = [0, \infty)$
3) An interval which is bounded above and below: (a, b)

There clearly many more possible subsets of $\mathbb{R}$, which can be much less straight forward than the ones above, on which we can define a probability density function. However, we will consider the above three for the purposes of this class.

1.32 Uniform Distribution

Consider case (3) from above, where we have some interval $(a, b) \subset \mathbb{R}$, and we wish to define a continuous probability distribution function $f_X(x)$ on this interval. Naturally, we must have that

$$\int_a^b f_X(x)dx = 1$$

One possible choice for f is to define f to be constant on the entire interval. For the constant function $f_X(x) = 1$,

$$\int_a^b 1dx = x|_a^b = b - a$$

This leads to the following definition.

Definition: A continuous random variable X is said to have a *uniform* distribution if it has the following pdf:

$$f_X(x) = \frac{1}{b-a}, \qquad a \leq x \leq b$$

By the logic above, this is a valid pdf, as long as $b \geq a$, which is necessary by our definition of an interval (a, b). We write $X \sim \text{unif}(a, b)$.

Theorem: For $X \sim \text{unif}(a, b)$,

$$E(X) = \frac{a+b}{2}, \qquad V(X) = \frac{(b-a)^2}{12}$$

Proof: Exercise.

1.33 Beta Distribution

Definition: A continuous random variable X is said to have a *beta* distribution if it has the following pdf:

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 < x < 1$$

for parameters $\alpha, \beta > 0$. We write $X \sim \text{beta}(\alpha, \beta)$

Note that the value of $f_X(x)$ is bounded on the interval $(0, 1)$ (i.e. the range of the random variable X is $(0, 1)$). This is particularly useful in a modeling context when we want to consider a distribution of probabilities, which are necessarily bounded in this interval. We will see an example of this later.

Note: The *gamma function* is given by:

$$\Gamma(\alpha) := \int_0^\infty x^{\alpha-1} e^{-x} dx$$

The gamma function simply returns a constant. It has the property that:

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$$

More specifically, for n which is an integer (i.e. $n \in \mathbb{Z}$),

$$\Gamma(n) = (n-1)!$$

Theorem: For $X \sim \text{beta}(\alpha, \beta)$,

$$E(X) = \frac{\alpha}{\alpha + \beta}$$
$$V(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

These can be derived using the usual methods and the properties of the gamma function, but the details will be omitted here.

Note: A unif$(0, 1)$ distribution is a special case of the beta, where $\alpha = \beta = 1$

1.34 Gamma Distribution

Definition: A continuous random variable X is said to have a *gamma* distribution if it has the following pdf:

$$f_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \qquad x > 0$$

for $\alpha, \beta > 0$. We write $X \sim \text{gamma}(\alpha, \beta)$.

Gamma random variables are defined on $\mathbb{R}^+$.

Theorem: For $X \sim \text{gamma}(\alpha, \beta)$,

$$E(X) = \alpha\beta, \qquad V(X) = \alpha\beta^2$$

Proof: Exercise.

Definition: A continuous random variable X is said to have a *exponential* distribution if it has the following pdf:

$$f_X(x) = \frac{1}{\beta} e^{-x/\beta}, \qquad x > 0$$

for $\beta > 0$. We write $X \sim \exp(\beta)$.

The exponential distribution is a special case of the gamma, where we allow $\alpha = 1$. Its expected value and variance follow from those of the gamma.

Note: The gamma and beta distributions are of interest because they are very malleable distributions in the sense that they can assume a variety of "shapes" given their parameter values. Therefore, if we don't particularly care about what distribution a certain random variable has, there are many situations in which the beta and gamma work well. The situations in which the beta works well are when the random variable has a bounded range, and the gamma works well when it assumes values on the positive real line.

Example:

Below is a plot of beta distributions which assume a variety of parameter values of α and β:

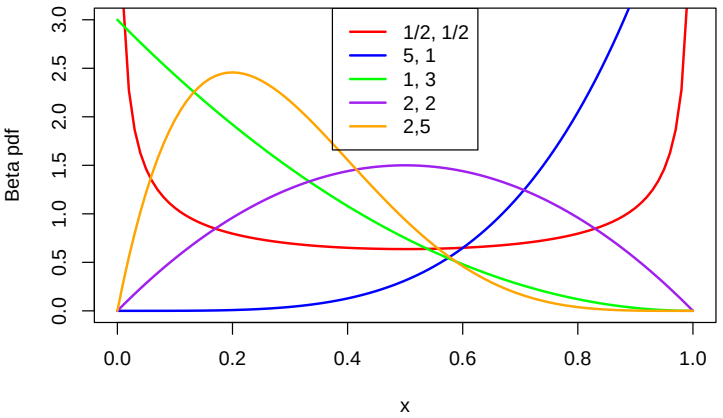


Figure 2: Beta Probability Density Function for a variety of parameter values

The beta distribution can be made to have many different general characteristics. For example, the beta$(1/2, 1/2)$ distribution gives more density to values of x close to 0 and 1, the beta$(2, 2)$ distribution is symmetric about $x = 1/2$, and the beta$(2, 5)$ distribution is skewed towards 0.

To find the pdf of a beta random variable where α and β are integer valued, we can use the property of the gamma function s.t. $\Gamma(n) = (n-1)!$. For example, consider $X \sim$ beta$(2, 3)$. Then X has pdf:

$$\begin{aligned} f_X(x) &= \frac{\Gamma(2+3)}{\Gamma(2)\Gamma(3)} x^{2-1}(1-x)^{3-1} \\ &= \frac{4!}{1!2!} x(1-x)^2 \\ &= 12x(1-x)^2 \qquad 0 < x < 1 \end{aligned}$$

Example: Consider $X \sim \text{gamma}(3, 4)$. Then X has pdf:

$$\begin{aligned} f_X(x) &= \frac{1}{4^3\Gamma(3)} x^{3-1} e^{-x/4} \\ &= \frac{1}{64 \cdot 2!} x^2 e^{-x/4} \\ &= \frac{1}{128} x^2 e^{-x/4}, \qquad x > 0 \end{aligned}$$

The constant (expressed in terms of the gamma function) which is in front of the beta or gamma distributions is called the *normalizing constant*, since it exists so that the integral of the pdf over its domain is equal to 1.

Definition: A random variable X is said to have a χ^2 (*chi-squared*) *distribution* if it has the following pdf:

$$f_X(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}, \qquad x > 0$$

for ν which is a positive integer ($\nu \in \mathbb{Z}^+$). We write $X \sim \chi^2(\nu)$. The parameter ν is referred to as the *degrees of freedom* of the chi-squared distribution, so we could also say that the r.v. X follows a chi-squared distribution with ν degrees of freedom.

The chi-squared distribution is also a special case of the gamma, where we allow $\alpha = \nu/2$ and $\beta = 2$.

Note: The reason we define the χ^2 distribution, as opposed to simply always using the gamma, is because it comes up often in statistical testing, so its worth while to give a name to this special case. The χ^2, F, and t distributions are all of particular interest in statistical testing, so we

are concerned less with the form of their pdf as we are with whether or not a certain statistic we come across in a data analysis context can be considered to be a random variable which follows one of these distributions. We will discuss these situations further in later lectures.

Example: Suppose that $X \sim \chi^2(4)$. Suppose we wish to find $P(X > 5)$. Then, naturally, we have that:

$$\begin{aligned} P(X > 5) &= \int_5^\infty \frac{1}{2^{4/2}\Gamma(4/2)} x^{4/2-1} e^{-x/2} dx \\ &= \int_5^\infty \frac{1}{4} x e^{-x/2} dx \\ &= \frac{7}{2e^{5/2}} \approx 0.287 \end{aligned}$$

Also notice that $P(X > 5) = 1 - P(X \leq 5) = 1 - F(5)$ where F is the cdf of a chi-squared distribution. In `R`, we can evaluate the cdf using the function `pchisq` (the second argument is the value of ν), so the following code will give the same result as above:

```
1 - pchisq(5, 4)
```

```
## [1] 0.2872975
```

Now consider the following situation. You are performing an experiment in order to test some hypothesis. Suppose that the experiment involves observing a single value, which we interpret as a random variable X. Should the hypothesis be true, we know that $X \sim \chi^2(4)$. If the hypothesis is false, then X may follow some other distribution, of which we do not know the form. Suppose that we observe $X = 5$. Is there sufficient evidence to suggest that our hypothesis was false?

According to the above calculation, there is about a 28.7% chance of observing $X = 5$ as the outcome of our experiment, supposing that our hypothesis is true. Now it is up to the experimenter to decide if 28.7% is "unlikely enough" of a result to conclude that our original hypothesis was false. Realistically, we would want to set a stricter criterion for making this determination.

This process is known as performing a *statistical test*. Essentially, we believe that, under our *null* hypothesis, our observation(s) should follow

a certain distribution (in this example, $\chi^2(4)$). We then accept or reject our hypothesis using the probability of observing an observation that was "at least that extreme" (in this case, it was $P(X > 5)$). We call this probability the *p-value*. This will be discussed further in later sections.

1.35 F Distribution

Definition: A continuous random variable X is said to have an F distribution if it has the following pdf:

$$f_X(x) = \frac{\Gamma(d_1/2 + d_2/2)}{x\Gamma(d_1/2)\Gamma(d_2/2)} \sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1+d_2}}}, \qquad x > 0$$

for $d_1, d_2 > 0$, usually taken to be positive integers. We write $X \sim \mathrm{F}(d_1, d_2)$. Like the χ^2 distribution, the parameters d_1 and d_2 are referred to the "degrees of freedom of the F distribution". We could say "an F distribution with d_1 and d_2 degrees of freedom.

Similarly to the χ^2, the F distribution is also used most commonly in the context of statistical tests.

1.36 Normal Distribution

Definition: A random variable X is said to have a *normal* distribution if it has the following pdf:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \qquad \text{for all } x \in \mathbb{R}$$

for μ which is any real number and σ^2 equal to any positive real number. We write $X \sim N(\mu, \sigma^2)$.

Note that now, we finally consider a distribution which has a domain of the entire real line, i.e. the range of the random variable X is the entirety of $\mathbb{R}$. Any interval $(a, b) \subset \mathbb{R}$ is such that $P(X \in (a, b)) > 0$.

Note: $Z \sim N(0, 1)$ is said to follow a *standard normal* distribution. The letter Z is usually used to denote a standard normal random variable.

The normal distribution is fundamental to the field of statistics. We will see the central limit theorem in the coming lectures, which says that the distribution of the average of a random sample approaches a normal distribution as the sample size approaches infinity. The first thing to note is that the parameter values μ and σ^2 of a normal distribution are directly equal to the expected value and variance of X:

Theorem: For $X \sim N(\mu, \sigma^2)$,

$$E(X) = \mu, \qquad V(X) = \sigma^2$$

We express the parameter as σ^2 because the standard error of a normal distribution:

$$\sqrt{V(X)} = \sigma$$

is often a quantity of interest. Furthermore, the notation σ^2 emphasizes the fact that the parameter must be positive, otherwise $f_X(x)$ would not be a valid pdf.

The following theorem will be important when we prove some results relating to a random sample later:

Theorem: Let $X \sim N(\mu, \sigma^2)$. Then the moment generating function $M_X(t)$ is given by:

$$M_X(t) := E(e^{tX}) = e^{\mu t + \sigma^2 t^2/2}$$

Some important properties of the normal distribution:

- The density function $f_X(x)$ is symmetric about the mean μ.
- The standard deviation σ determines how much the density is concentrated around the mean.
- It has non-zero density on the entire real line.
- In general, observations far from the mean have a low probability of occurring, observations near the mean have a high probability of occurring.

Example: Consider generally functions which have the form "e raised to some quadratic power", i.e.:

$$e^{ax^2+bx+c}$$

By completing the square, we can write:

$$ax^2 + bx + c = a(x + m)^2 + n$$

Note that σ^2 is a strictly positive quantity, so in order to make this general function "fit" the form of a normal distribution, we must have that the constant a above is negative. Then, we have that:

$$a = \frac{1}{-2\sigma^2} \implies \sigma^2 = \frac{1}{-2a}$$

Clearly, $\mu = -m$. Now we must consider the constant in front, which, in the normal density function, has the form $\frac{1}{\sqrt{2\pi\sigma^2}}$. Note that:

$$e^{a(x+m)^2+n} = e^n e^{a(x+m)^2}$$

e^n is a constant. Therefore this general function that we described is simply a multiplicative constant away from having the form of a normal distribution. This constant is referred to as the *normalizing* constant, which exists to enforce the property:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Our original function can therefore be made into a valid pdf by multiplying by the constant:

$$\frac{e^{-n}}{\sqrt{2\pi\sigma^2}} = \frac{e^{-(-b^2/(4a^2)+c)}}{\sqrt{2\pi(-1/2a)}}$$

As an example, consider the function

$$f(x) = e^{-2x^2+x+1}$$

In this case, the "mean" parameter $\mu = -b/(2a) = 1/4$. Functions of this form (so long as a is negative) will have the familiar "bell" shape of the normal distribution.

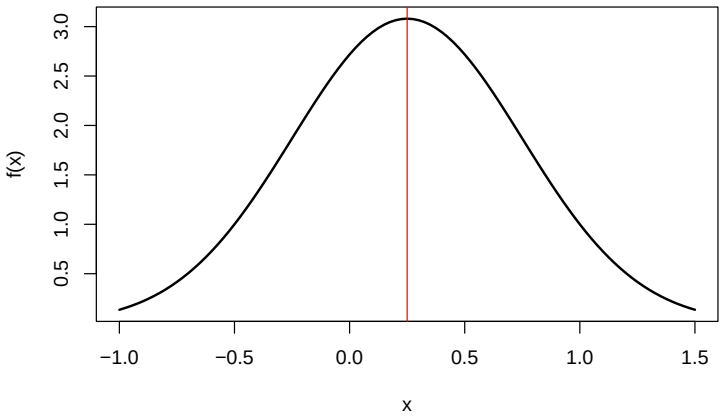


Figure 3: $f(x) = e^{-2x^2+x+1}$, where the red line is $x = 1/4$

In order to make this a valid pdf, we calculate our normalizing constant to be:

$$\frac{e^{(b^2/(4a^2)-c)}}{\sqrt{2\pi(-1/2a)}} = \frac{e^{-1/16-1}}{\sqrt{2\pi(1/4)}} = \frac{e^{-17/16}}{\sqrt{\pi/2}}$$

and define:

$$f_X(x) = \frac{e^{-17/16}}{\sqrt{\pi/2}} e^{-2x^2+x+1}$$

which is a normal distribution with mean $1/4$ and variance *something* (find this yourself as an exercise).

1.37 t Distribution

Definition: A random variable X is said to have a t distribution if it has the following pdf:

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2} \qquad \text{for all } x \in \mathbb{R}$$

for $\nu > 0$ (usually integer valued). We write $X \sim t(\nu)$. Once again, we refer to ν as the "degrees of freedom" of the t distribution.

The t distribution is also predominantly used in testing, but has uses outside of testing due to the fact that, like a normal distribution, it is symmetric and is defined on the entirety of $\mathbb{R}$. It also has the property that for low degrees of freedom, it has "heavier" tails than the normal distribution, which is to say that it gives higher density to areas further from the mean, as can be seen in the figure below.

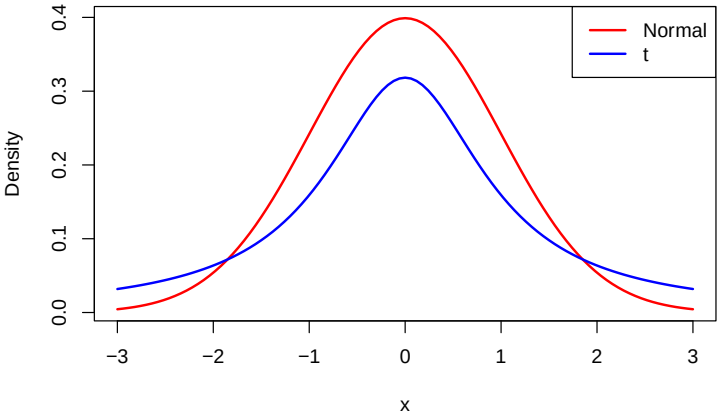


Figure 4: Standard normal distribution and a t distribution with 1 degree of freedom

1.38 Probabilities and Quantiles

Recall once again the defining property of a probability density function $f_X(x)$ of a random variable X:

$$P(X \in A) = \int_A f_X(x)dx$$

Given any of the pdfs that were given in the previous lecture, we can integrate to find probabilities. In practice, one doesn't write out these integrals and solve them by hand. We rather use a programming language like R, which has built in functions for evaluating probabilities. All that we need to know are the parameter values.

Example:

X follows a standard normal distribution. What is $P(X \leq 3)$?

$$P(X \leq 3) = \int_{-\infty}^{3} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = F(3)$$

where F is the cdf of a normal distribution. In R, we can evaluate the cdf of a normal distribution using the function `pnorm`:

```
pnorm(3, 0, 1)
```

```
## [1] 0.9986501
```

which is the probability in question. Here are some more examples:

1) $X \sim F(3, 20)$. What is $P(X \leq 0.9)$?

```
pf(0.9, 3, 20)
```

```
## [1] 0.5414603
```

2) $X \sim t(8)$. What is $P(X \geq 4)$?

```
1 - pt(4, 8)
```

```
## [1] 0.001974886
```

3) $X \sim t(8)$. What is $P(X \leq -4)$?

```
pt(-4, 8)
```

```
## [1] 0.001974886
```

This is the same as (2) because the t distribution is symmetric (draw a picture).

Related to the probabilities of a distribution are its *quantiles*, for which we give a definition now:

Definition: The *p-quantile* of a distribution, given some $0 < p < 1$, is the value q such that:

$$P(X \leq q) = p$$

Example: Let $X \sim \text{beta}(3, 1)$, which has the following pdf:

$$f_X(x) = 3x^2, \qquad 0 < x < 1$$

Suppose we wanted to find the 0.5-quantile (aka the median) of this distribution. Let q be the median, then

$$0.5 = \int_0^q 3x^2 dx = x^3|_0^q = q^3$$

and hence,

$$q = \left(\frac{1}{2}\right)^{1/3} \approx 0.794$$

Example: Let $X \sim F(3, 20)$. What is the 0.517-quantile of the distribution of X?

We'll use the result from above that $P(X \leq 0.9) = 0.517$, so this necessarily implies that $q = 0.9$ is the 0.517-quantile.

Example: Consider $X \sim N(0, 1)$. What is the 0.95-quantile of the distribution of X?

$$P(X \leq q) = 0.95 = \int_{-\infty}^q \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

In the case of the normal distribution, solving this equation is impossible in closed form, so we appeal to the function `qnorm` in `R`, which takes inputs of a probability p and outputs the p-quantile of the normal distribution:

```
qnorm(0.95, 0, 1)
```

```
## [1] 1.644854
```

Note: It used to be common to obtain these probabilities or quantiles from a table of values (e.g. a table of Z values gives probabilities or quantiles of a standard normal distribution. Chi-squared, t, and F tables give options for varying degrees of freedom, etc.). You may still be asked to do this, but probably just during an exam in a statistics class.

Note: For a discrete distribution, the p-quantile is chosen to be the smallest value of q in the range of the random variable X such that $P(X \leq q) \geq p$. For example, if we consider a binom$(100, 0.5)$ random variable:

```
qbinom(0.95, 100, 0.5)
```

```
## [1] 58
```

```
pbinom(58, 100, 0.5)
```

```
## [1] 0.955687
```

```
pbinom(57, 100, 0.5)
```

```
## [1] 0.9333947
```

The probability of being less than 58 is greater than 0.95, but it is still the 0.95 quantile because it is the smallest value q such that $P(X \leq q) \geq 0.95$

1.39 Some Relationships Between Distributions

Many of the distributions we discussed above are important because of the way that they relate to other distributions. In particular, it is common in statistics to make the claim that the data is normally distributed (often by use of the central limit theorem, which will be discussed later). This often results in certain functions of the data following one of the other common distributions, by one of the relationships below. These relationships will be stated without proof, as showing that these relationships are true is beyond the scope of this course. However, we have all of the tools to fully understand what these relationships mean.

Theorem: Let $X_1, \cdots X_n \sim N(\mu, \sigma^2)$ be *independent and identically distributed* normal random variables. Then,

$$\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Proof: Recall from the previous lecture that we have, for $X \sim N(\mu, \sigma^2)$,

$$M_X(t) = e^{\mu t + \sigma^2 t^2/2}$$

If the moment generating functions of two random variables are equal, then they have the same distribution. Now consider the mgf of our random variable $\bar{X}$:

$$\begin{aligned} M_{\bar{X}}(t) = E(e^{t/n \sum X_i}) &= E(e^{t/nX_1}) \cdots E(t/nX_n) \qquad \text{(independence)} \\ &= \left(e^{\mu t/n + \sigma^2 t^2/(2n^2)}\right)^n \qquad \text{(identically distributed)} \\ &= \left(e^{\mu t + \sigma^2/(2n) t^2}\right) \end{aligned}$$

where the final line of the above is the mgf of a $N(\mu, \sigma^2/n)$ random variable, so we have our result. ■

The normal and chi-squared distributions are related in the following way:

Theorem: Let $Z_1, \cdots, Z_k \sim N(0,1)$ be independent and identically distributed *standard* normal random variables. Then,

$$X := \sum_{i=1}^{k} Z_i^2 \sim \chi^2(k)$$

Remember again that $\sum_{i=1}^{k} Z_i^2$ is just a function of random variables, lets say $g(Z_1, \cdots, Z_k)$, which necessarily implies that the sum is itself a random variable. Considering this random variable on its own, it has all the usual properties of a random variable, including a pdf. The above statement is saying that the new random variable X is $\chi^2(k)$, which means that it has the pdf of a χ^2 random variable with k degrees of freedom, as was shown in the previous lecture.

Theorem: Let $S_1 \sim \chi^2(d_1)$ and $S_2 \sim \chi^2(d_2)$ be *independent* chi-squared random variables. Then,

$$X = \frac{S_1/d_1}{S_2/d_2} \sim F(d_1, d_2)$$

Exercise: Write down an expression for the joint distribution of S_1 and S_2. How would you use this joint to find $E(X)$, as defined above?

Theorem: Let $X_1, \cdots X_n \sim N(\mu, \sigma^2)$ be *independent and identically distributed* normal random variables. Then define:

$$\bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$T = \frac{\bar{X} - \mu}{\sqrt{\frac{1}{n} \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}} \sim t(n-1)$$

You may recognize the quantity in the bottom as the sample variance:

$$S^2 := \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

Also recall from above that $\bar{X}$, as we have defined it, itself is a normal random variable. We will revisit all of these relationships soon when we discuss statistical tests.

2 Statistical Inference

2.1 Point Estimation

Recall the definition of a random sample from earlier:

Definition: A *random sample* is an *independent and identically distributed* (iid) collection of random variables.

We have that two random variables X and Y are identically distributed if they have the same distribution function ($f_X(x) = f_Y(x)$ for all x). X and Y are independent if their joint can be expressed as the product of the marginals: $f(x, y) = f_X(x)f_Y(y)$.

Definition: Let $X_1, \cdots, X_n$ be a random sample. Then a function $T(X_1, \cdots, X_n)$ of the random sample is called a *statistic*. The distribution of the statistic is called the *sampling distribution* of the statistic.

Example: The *sample mean*, which we've come across before, is probably the most common example of a statistic:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

In the case of a *normal* random sample, we showed that the sampling distribution of $\bar{X}$ was $N(\mu, \sigma^2/n)$.

We also have the *sample variance*, which is the statistic given by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

The sample standard deviation $S = \sqrt{S^2}$ is also a statistic.

We often use the word *point estimator* to refer to a statistic (itself a random variable) which is an attempt to provide an estimate of a particular parameter θ in the probability distribution function. A *point estimate* refers to an observed value of such a random variable. Technically, we can call any statistic a "point estimator for θ", but only some of these point estimators actually serve the desired purpose. Here we define some notions of what makes a point estimator "good".

Definition: A point estimator $\hat{\theta}$ of a parameter θ is said to be *unbiased* if

$$E(\hat{\theta}) = \theta$$

2.1.1 The Method of Moments

In general, we are concerned not only with the properties of estimators but with the methods of obtaining them. One such method is the *method of moments*. Essentially, you just equate the "sample moments" (defined below) with the "population moments" ($E(X)$, $E(X^2)$, etc.) to produce a system of k equations, for a distribution which contains k parameters.

We will consider the case of a distribution with two parameters, say θ_1 and θ_2. Then, given a random sample $X_1, \cdots, X_n$, the method of moments says that we can produce estimates, $\hat{\theta_1}$ and $\hat{\theta_2}$ by solving the following system of equations:

$$E(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$E(X^2) = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

where we note that each of $E(X)$ and $E(X^2)$ are functions of one or both of the parameters θ_1, θ_2, so we have a system of two equations with two unknowns.

The first sample moment the sample mean.

Example: Suppose that we have the following random sample from a normal distribution:

$$(4.7, 5.1, 4.3, 4.5)$$

Furthermore, we know that:

$$
\begin{aligned}
E(X) &= \mu \\
E(X^2) &= V(X) + E(X)^2 = \sigma^2 + \mu^2
\end{aligned}
$$

Furthermore,

$$
\bar{X} = \frac{1}{4}(4.7 + 5.1 + 4.3 + 4.5) = 4.65
$$

and,

$$
\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{1}{4}(4.7^2 + 5.1^2 + 4.3^2 + 4.5^2) = 21.71
$$

So we have the following system of equations:

$$
\begin{aligned}
\hat{\mu} &= 4.65 \\
\hat{\sigma}^2 + \hat{\mu}^2 &= 21.71 \\
\implies \hat{\sigma}^2 &= 21.71 - 4.65^2 = 0.0875
\end{aligned}
$$

which gives us the method of moment estimators $\hat{\mu}$ and $\hat{\sigma}^2$.

Example: Note that in the above,

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2
\end{aligned}
$$

which is a *biased* (i.e. its not unbiased) estimator for σ^2 (see example above for an unbiased estimator).

To find what its expected value actually is:

$$E(\hat{\sigma}) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = = \frac{1}{n}\left(\sum E(X_i^2) - 2nE(\bar{X}^2) + nE(\bar{X}^2)\right)$$

where

$$E(\bar{X}^2) = V(\bar{X}) + E(\bar{X})^2 = \frac{\sigma^2}{n} + \mu^2$$

and hence

$$E(\hat{\sigma}) = \frac{1}{n}\left(n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right) = \frac{n-1}{n}\sigma^2$$

Something we can do when we have a biased estimator like the one above is multiply it by the multiplicative constant required to make it unbiased. In this case, we would multiply by $\frac{n}{n-1}$, which would yeild the *unbiased* estimator:

$$\hat{\sigma^2}_{\text{unbiased}} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

2.1.2 Maximum Likelihood

Consider a random sample $X_1, \cdots, X_n$ which have (the same) marginal distribution $f_X(x)$. Then, by independence, the joint distribution of this random sample can be written as:

$$f_{X_1,\cdots X_n}(x_1, \cdots x_n) = \Pi_{i=1}^{n} f_X(x_i) = f_X(x_1)f_X(x_2)\cdots f_X(x_n)$$

We often refer to this joint distribution as the *likelihood*, expressed as a function of the parameters rather than of the "data" $x_1, \cdots, x_n$. I.e suppose that $X_i \sim \text{fam}(\theta_1, \cdots, \theta_k)$, then the *likelihood* is given by:

$$L(\theta_1, \cdots, \theta_k) = f_{X_1,\cdots X_n}(x_1, \cdots x_n)$$

This is just an alternative notation. The likelihood and the joint pdf of the random sample are exactly the same function. The reason why we consider the likelihood as a function of the parameters is because we can use this function to produce estimates for our parameters in the following way.

Suppose that we have the observed value $x_1, \cdots, x_n$ of a random sample $X_1, \cdots, X_n$. Then a good way to estimate the θ's would be to find the value of the θ's which maximizes the likelihood of observing that sample. Hence we just need to maximize the likelihood function. We can do this by taking its derivative and setting it equal to zero, and solving for each θ. However, an equivalent process (which is nearly always much easier) is to take to natural logarithm of the likelihood function and then finding its maximum, which is necessarily equal to the maximum of the likelihood function. We define the *log likelihood* as follows:

$$\ell(\theta_1, \cdots, \theta_k) = log(L(\theta_1, \cdots, \theta_k))$$

Example: Consider a random sample $X_1, \cdots X_n \sim \text{pois}(\lambda)$. Then the joint pdf (i.e. the likelihood) of the random sample is given by:

$$L(\lambda) = f_{X_1, \cdots X_n}(x_1, \cdots, x_n) = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

To use *maximum likelihood* to arrive at an estimate for the parameter λ, we find the value of λ, which will be our estimate $\hat{\lambda}$, which gives the maximum of the above function.

An equivalent process to maximizing the likelihood is to maximize the *log likelihood*, defined as:

$$\ell(\lambda) = \log(L(\lambda))$$

which works because the natural logarithm is a monotonic function. Applying this process to the Poisson example, we have

$$\ell(\lambda) = -n\lambda + \sum x_i \log(\lambda) - \log\left(\prod x_i!\right)$$
$$\frac{\partial \ell}{\partial \lambda} = -n + \frac{\sum x_i}{\lambda}$$

Setting the derivative equal to zero in order to find the value of λ which maximizes the likelihood:

$$\frac{\partial \ell}{\partial \lambda} = -n + \frac{\sum x_i}{\lambda} = 0$$
$$\lambda = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

where $\bar{x}$ is the sample mean.

2.2 Bayesian Point Estimation

The *Bayesian Paradigm* is as follows: We have some random variable Y, whose probability density function $f_Y(y)$ belongs to some distribution family parameterized by θ. We allow $\theta \sim \nu$, that is the parameter θ itself is considered to be a random variable with some "prior" distribution ν (that is, $\nu(\theta)$ is a probability density function). Then we observe a random sample $Y_1, \cdots Y_n$, which are *iid* $f_Y(y)$. Our goal is to obtain the "posterior" distribution of θ, given by:

$$\nu(\theta|y_1, \cdots, y_n) := \nu(\theta|\mathbf{y}) = \frac{f(y_1, \cdots, y_n|\theta)\nu(\theta)}{f(y_1, \cdots, y_n)}$$

The form of the posterior distribution should look familiar. Recall *Bayes Rule* from when we discussed probability:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

The form of the posterior is just Bayes rule for distributions. We can show that the posterior has the form above simply by applying the definition of conditional distributions to $\nu(\theta|\mathbf{y})$.

Note that $\nu(\theta|\mathbf{y})$ is a density with respect to θ, so the marginal distribution $f(y_1, \cdots, y_n)$ in the denominator is a *constant* with regards to the posterior distribution. It is equal to:

$$f(y_1, \cdots, y_n) = \int_{-\infty}^{\infty} f(y_1, \cdots, y_n, \theta) \\ \int_{-\infty}^{\infty} f(y_1, \cdots, y_n|\theta)\nu(\theta)d\theta$$

This constant is often unknown, so the following proportionality is often referred to as *Bayes Rule*:

$$\nu(\theta|\mathbf{y}) \propto f(y_1, \cdots, y_n|\theta)\nu(\theta)$$

If we can derive the posterior distribution, then we will have a distribution for the parameter θ. If our goal is to produce a point estimate for θ, then the distribution lends itself to many possible options for point estimates. For example, the posterior mean, median, and mode can all serve as potential point estimates for θ.

Example: Consider a random sample $Y_1, \cdots Y_n \sim \text{pois}(\lambda)$. Let $\lambda \sim \exp(\beta)$, i.e. our prior distribution ν is:

$$\nu(\lambda) = \frac{1}{\beta}e^{-\lambda/\beta}$$

Using Bayes rule, our posterior distribution is proportional to:

$$\begin{aligned}
\nu(\lambda|\mathbf{y}) &\propto f(y_1, \cdots, y_n|\theta)\nu(\theta) \\
&= \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}\right) \frac{1}{\beta}e^{-\lambda/\beta} \\
&\propto e^{-n\lambda}\lambda^{\sum y_i}e^{\lambda/\beta} \\
&= e^{-\lambda(n+1/\beta)}\lambda^{\sum y_i}
\end{aligned}$$

Note that

$$\left(n + \frac{1}{\beta}\right)^{-1} = \frac{1}{n + 1/\beta} = \frac{\beta}{n\beta + 1}$$

We know that the posterior distribution must be a valid probability density function, i.e. it must integrate to 1. The above function is the kernel of a gamma density, for which we know the normalizing constant, so we can write:

$$\nu(\lambda|\mathbf{y}) = \frac{1}{\Gamma(\sum y_i + 1)\left(\frac{\beta}{n\beta+1}\right)^{\sum y_i+1}} e^{-\lambda/(\beta/(n\beta+1))} \lambda^{(\sum y_i+1)-1}$$

i.e. $\lambda|\mathbf{y} \sim \text{gamma}(\sum y_i + 1, \frac{\beta}{n\beta+1})$

To get a point estimate for λ, one option is to use the expected value of the posterior. Recall that the expected value of a gamma(α, β) distribution is $\alpha\beta$, so we can take

$$\hat{\lambda} = \left(\sum_{i=1}^{n} y_i + 1\right)\left(\frac{\beta}{n\beta + 1}\right)$$

This can be alternatively expressed as:

$$\hat{\lambda} = \left(\sum_{i=1}^{n} y_i + 1\right)\left(\frac{\beta}{n\beta + 1}\right) = \ = \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)\left(\frac{n\beta}{n\beta + 1}\right) + (\beta)\left(\frac{1}{n\beta + 1}\right)$$

As $n \to \infty$, $\hat{\lambda} \to \frac{1}{n}\sum y_i$, which was our maximum likelihood estimate for λ. Furthermore, note that $\hat{\lambda}$ is expressed above as a linear combination between the maximum likelihood estimate and the prior mean β (mean of an exp(β) distribution).

2.3 Statistical Tests

A *statistical test* is the process of using a *statistic* in order to either reject or accept one hypothesis versus another. In general, we have a null hypothesis, denoted H_0, and an alternative hypothesis, denoted H_1. The "hypothesis" is simply any statement which concerns the "population", i.e. the distribution from which our random sample came from. They are generally phrased in terms of the parameters of that distribution.

Example: Consider a random sample from a normal distribution with mean μ. We could consider testing the following hypothesis:

$$H_0 : \mu = 5$$
$$H_1 : \mu \neq 5$$

Methods of obtaining point estimators, such as maximum likelihood, the method of moments, and the Bayesian estimators that were just discussed motivate why certain *statistics* are important, in relation to certain parameters. We saw that the sample mean $\bar{X}$, arises in many situations. One such case was as the maximum likelihood estimator for μ in a $N(\mu, \sigma^2)$ distribution. In this case, $\bar{X}$ is distributed $N(\mu, \sigma^2/n)$ when each $X_i \sim N(\mu, \sigma^2)$.

Theorem: Consider a random sample $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$, then sample mean $\bar{X} \sim N(\mu, \sigma^2/n)$. Furthermore,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

i.e. Z follows a standard normal distribution. We call the process of transforming $\bar{X}$ in this way *standardization*.

2.4 One Sample Tests

Consider the situation in which we have a random sample $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$, and we want to test the following hypothesis:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

If we **assume that H_0 is TRUE**, then

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

We will call Z our test statistic, which we know follows a standard normal distribution. Assuming that we know σ, we can calculate the value of Z from our random sample, since μ_0 is just the value we choose to be in our hypothesis.

Now we construct a *rejection region* for this test, that is, values of the above test statistic Z for which we reject H_0 in favor of H_1. Should we observe a value of Z which is in the rejection region, we will conclude that H_0 is false and reject it. Otherwise we will fail to conclude this and accept it (or "fail to reject" it, depending on the experimental context).

Clearly, regardless of whether or not H_0 is true, it is possible to observe *any* value of $Z \in \mathbb{R}$, by nature of the normal distribution. Hence, it is always possible that our test makes an error. Therefore, we must choose the rejection region to the best of our ability, to reduce that probability. In this situation, its clear. We want to reject H_0 if our observed value of Z is far from 0, because this has a low probability of occurring when $Z \sim N(0, 1)$. Therefore, our rejection region may look like:

$$\text{Reject } H_0 \text{ if } Z < a \text{ or } Z > b$$

But what values of a and b should we choose? We can use the quantiles of the standard normal distribution to construct a region that is more exact.

Notation: Let $z_{\alpha/2}$ refer to the $(1-\alpha/2)$-quantile of a $N(0, 1)$ distribution (often referred to as a Z-value). i.e.

$$P(Z \leq z_{\alpha/2}) = 1 - \alpha/2$$

By symmetry of the normal distribution about 0, we must have that:

$$P(Z \leq -z_{\alpha/2}) = \alpha/2$$

Or, together we have that

$$\begin{aligned} P(-z_{\alpha/2} < Z \leq z_{\alpha/2}) &= P(Z \leq z_{\alpha/2}) - P(Z \leq -z_{\alpha/2}) \\ &= (1 - \alpha/2) - \alpha/2 \\ &= 1 - \alpha \end{aligned}$$

Returning to the testing example, lets construct the rejection region as follows:

$$\text{reject } H_0 \text{ if } Z < -z_{\alpha/2} \text{ or } Z > z_{\alpha/2}$$

This test has the particular property that **should H_0 be TRUE**, then the probability of (**incorrectly**) rejecting it is:

$$P(Z \leq -z_{\alpha/2}) + P(Z > z_{\alpha/2}) = \alpha/2 + \alpha/2 = \alpha$$

The probability of incorrectly rejecting a true H_0 is called the *type 1 error probability*. By choosing the $z_{\alpha/2}$ quantile to construct our rejection region, we know exactly what the type 1 error probability of our test will be. Note that choosing our rejection region to be *too* small, however, will cause the test to always accept H_0, even if its false. This is called a *type 2 error*, and much of statistical testing involves trying to strike a balance between these two types of errors.

Example: Suppose we have a normal random sample as follows:

$$(414, 402, 409, 394, 404)$$

and that we know the value of $\sigma = 10$.

Suppose we wish to test the following hypothesis:

$$H_0 : \mu = 405$$
$$H_1 : \mu \neq 405$$

Then we compute the observed value of the test statistic Z:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{404.6 - 405}{10/\sqrt{5}} = -0.089$$

Suppose we want to perform a test which has probability 0.05 of making an error. Then we test using the $z_{\alpha/2} = z_{0.025}$ quantile of a standard normal:

```
qnorm(0.025, 0, 1)
```

```
## [1] -1.959964
```

Clearly, $Z = -0.089$ is not smaller than -1.96 nor is it larger than 1.96, so we accept $H_0 : \mu = 405$ since Z lies outside of our rejection region.

This data was generated in R using `rnorm(5, 405, 10)` (and then rounded to the nearest integer). That is to say, 405 is the "true" value of the parameter μ in the distribution from which this random sample was obtained. So, our test did *not* make an error, since it correctly accepted a true null. In general, since we accept/reject based on $z_{\alpha/2}$, we should expect this test to make an error with probability α, as was shown above.

We can also consider the following probability (where Z is the random variable corresponding to our test statistic):

$$P(|Z| \geq 0.089)$$

This probability represents how likely it is to observe a value of Z which is "at least as extreme" as -0.089. "Extreme" is a bit of a vague term, but in the context of a symmetric distribution like the standard normal, "extreme" signifies "far from 0", since we expect observations to be clustered around 0. Continuing our calculation:

$$P(|Z| \geq 0.089) = P(Z \geq 0.089) + P(Z \leq -0.089) = 2P(Z \leq -0.089)$$

where the last step follows from the symmetry of the normal distribution.

We can evaluate the last probability using the cdf of a standard normal:

```
2 * pnorm(-0.089, 0, 1)
```

```
## [1] 0.9290819
```

This value is our *p-value*. Given $\alpha = 0.05$, the test we just performed is equivalent to "rejecting if the p-value is less than 0.05" (think about the relationship between quantiles and the cdf).

2.5 The Central Limit Theorem

Recall the following results:

Theorem: Consider a random sample $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$, then:

$$\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$$

and

$$Z := \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

We now state the *central limit theorem*:

Theorem: Consider a random sample $X_1, \cdots, X_n$, taken from *any* population with mean μ and variance σ^2. Then, as $n \to \infty$,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Alternatively, this is to say that as $n \to \infty$,

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

which is analogous to the result we showed earlier for a *normal* random sample.

The central limit theorem is a result which deals with the limit as $n \to \infty$. However it is also phrased as an approximation, under the assumption that n is "large enough". I.e. one might state

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{\text{approx}}{\sim} N(0,1)$$

Technically, this is always an approximation, though not always a *good* approximation. The approximation is made better with larger values of n (i.e. more elements in the random sample). It is also made better when the random sample is from a distribution that is closer to being normal, as we show in the next example.

Example: Consider a beta$(2,2)$ distribution. The plot of its density function is below:

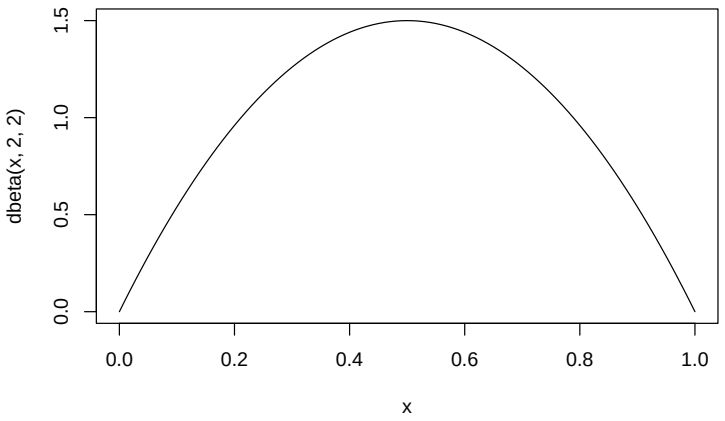


Figure 5: beta(2,2) pdf

Lets consider a sample size of $n = 5$, and see via simulation how well the CLT approximation of the distribution of $\bar{X}$ works.

This is our setup: $X_1, \cdots X_5 \sim$ beta$(2,2)$. Via the CLT, we claim that

$$\bar{X} \overset{\text{approx}}{\sim} N(\mu, \sigma^2/n)$$

Recall that μ and σ^2 are the expected value and variance of the distribution from which our random sample was drawn, respectively. Since we are working with a beta distribution, we have that these values are given by:

$$\mu = \frac{\alpha}{\alpha+\beta} = \frac{2}{2+2} = \frac{1}{2}$$
$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{4}{(4)^2(5)} = \frac{1}{20}$$

Therefore the CLT states that:

$$\bar{X} \overset{\text{approx}}{\sim} N\left(\frac{1}{2}, \frac{1}{20n}\right)$$

Using R, a random sample of size $n = 5$ is obtained 10000 times, and then $\bar{X}$ is calculated. We use this to plot an estimated density and compare that to the approximate density given by the CLT (think of the estimated density as serving the same purpose as a histogram. Its a function which estimates another function, that function being the true pdf):

```
set.seed(405)
n <- 5
nsim <- 10000
meansv <- c()
for (i in 1:nsim) {
  x <- rbeta(n, 2, 2)
  meansv <- c(meansv, mean(x))
}
```

Based on this figure, it seems that our approximation of the distribution of $\bar{X}$ is very good, even with a small sample size of 5. What if we now consider a beta distribution which looks "less normal". For example, a beta$(0.1, 0.1)$ distribution, whose density is plotted below.

Via the same equations as above, we have that the expected value is $\mu = \frac{1}{2}$, and the variance is $\sigma^2 = 5/24$.

Repeating the same simulation as above:

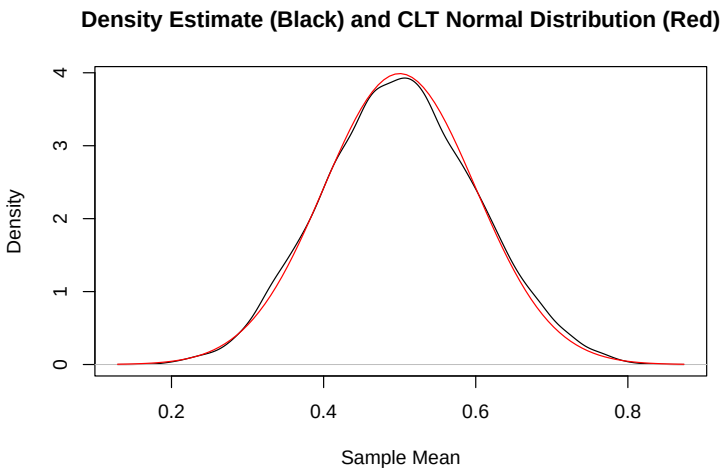


Figure 6: Density estimate created from simulated data and the true normal density

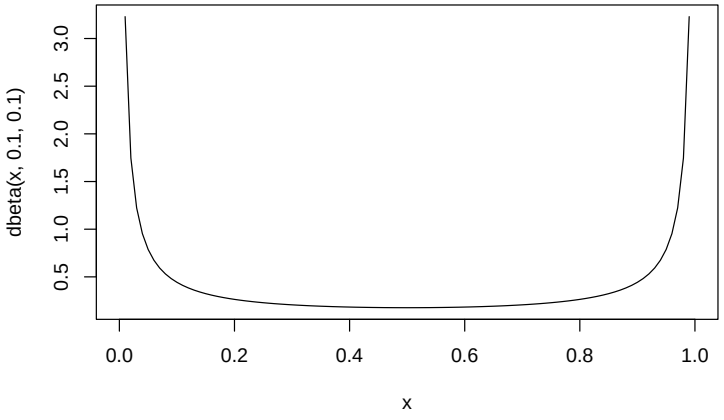


Figure 7: beta(0.1, 0.1) pdf

```r
n <- 5
nsim <- 10000
meansv <- c()
for (i in 1:nsim) {
  x <- rbeta(n, 0.1, 0.1)
  meansv <- c(meansv, mean(x))
}
```

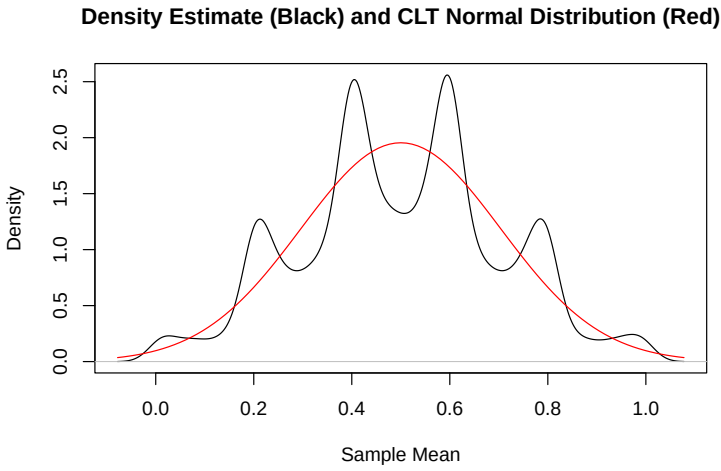


Figure 8: Density estimate from the beta(0.1, 0.1) data with random samples of size 5

Pretty weird (take a guess at what could explain the shape of the density estimate). But if we let $n = 1000$, then our approximation becomes much better (see below).

2.6 Applying the CLT to One Sample Tests

In the one sample tests which were introduced last lecture, we had a random sample $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$, and we wanted to test the following hypothesis:

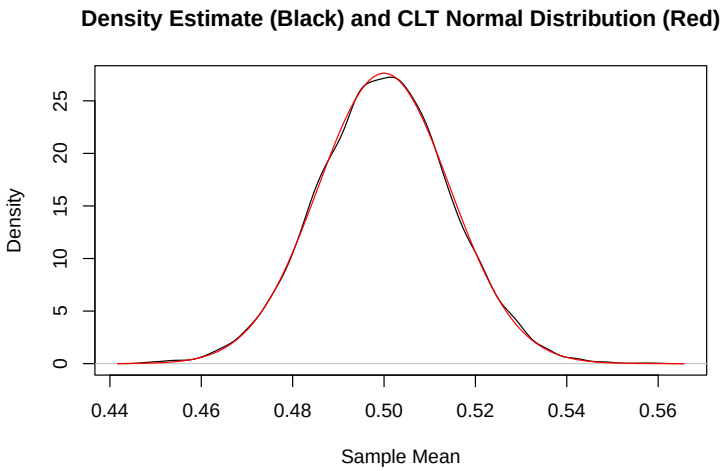


Figure 9: Density estimate from the beta(0.1, 0.1) data with random samples of size 100

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Then we constructed a test statistic $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, which is necessarily distributed standard normal, and used this information to construct our test.

Using the central limit theorem, we can extend this. That is, given a random sample $X_1, \cdots, X_n$ which follow a distribution with mean μ, variance σ^2 (*not* necessarily a normal distribution), then the test statistic:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{\text{approx}}{\sim} N(0, 1)$$

Using this approximation, we can perform one sample tests in exactly the same way. The p-values are not *exact*, as this is an approximation, so in cases of high degrees of non-normality and small sample sizes, we get

type 1 error probabilities which differ from α, when we construct the test using the appropriate quantiles. This next example illustrates this.

Example: Consider the highly non-normal beta$(0.1, 0.1)$ distribution from above. Consider the following test:

$$H_0 : \mu := \frac{\alpha}{\alpha + \beta} = \frac{1}{2}$$
$$H_1 : \mu \neq \frac{1}{2}$$

As was shown above, the true mean of the beta$(0.1, 0.1)$ distribution actually is $1/2$. Therefore, the null is true, so a rejection of the null constitutes a type 1 error. The following code generates a random sample of size 4 from the beta$(0.1, 0.1)$ distribution. Then, we estimate the type 1 error probability by doing this 100000 times, and calculating the proportion of type 1 errors which were made.

```
# proportion rejected in the simulation
reject/nsim
```

```
## [1] 0.06961
```

This is higher than expected, since theoretically we should have a type 1 error probability of 0.05. The reason for this is because of our extremely small sample size along with the high degree of non-normality of the beta$(0.1, 0.1)$ distribution. Should we increase the sample size to 5 however, we get:

```
reject/nsim
```

```
## [1] 0.05099
```

which is already close to our theoretical type 1 error probability of 0.05. So, in the case of this particular test with *known* variance, our theoretical results agree with the actual results, even for very small sample sizes like $n = 5$. We will see that this is harder when we have *unknown* variance, and we must estimate it.

2.7 t Tests

Recall the following result:

Theorem: Let $X_1, \cdots X_n \sim N(\mu, \sigma^2)$ be a random sample (i.e. they're *iid*). Then define

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Then $T \sim t(n-1)$.

Consider the situation in which we have a random sample $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$, and we want to test the following hypothesis:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

This is the same setup as the one sample tests that we have been doing in the previous lectures. However, suppose that we don't know the value of σ^2. We know that we can estimate σ^2 unbaisedly using the sample variance:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Unbiased meaning of course that $E(S^2) = \sigma^2$.

The above result regarding the t distribution gives us a way to still perform one-sample tests when we estimate σ^2 as opposed to knowing its true value. That is, our test statistic is now given by

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Under the assumption that H_0 is *true*,

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

Now we construct the *rejection region* for this test: the values of the above test statistic T for which we reject H_0 in favor of H_1. As was the case in the previous one sample tests, it will be of the form:

$$\text{reject } H_0 \text{ if } T < a \text{ or } T > b$$

This is because, like the standard normal, a t distribution is concentrated around zero. Values of T which suggest that H_0 is false are the values which fall far from the origin, i.e. values which lie in a rejection region like the one given above.

Since we are using a t distribution instead of the standard normal, we use the quantiles of a $t(n-1)$ distribution to construct this rejection region. Define $t_{n-1,\alpha/2}$ to be the $(1-\alpha/2)$ quantile of a t distribution with $n-1$ degrees of freedom, i.e.

$$P(T \leq t_{n-1,\alpha/2}) = 1 - \alpha/2$$

To construct a rejection region such that our test has type 1 error probability of α, we use the following rejection region:

$$\text{reject } H_0 \text{ if } T < -t_{n-1,\alpha/2} \text{ or } T > t_{n-1,\alpha/2}$$

Example: Consider the following random sample from a normal distribution, of which we know nothing about its parameters:

$$(2.13, 11.39, 9.50, 4.24)$$

Suppose we want to test:

$$H_0 : \mu = 12$$
$$H_1 : \mu \neq 12$$

We can estimate the variance using S^2 and compute the t statistic:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{6.815 - 12}{4.35/\sqrt{4}} = -2.38$$

The $t_{3,0.05/2}$ quantile is 3.18:

```
qt(1 - 0.05/2, 3)
```

```
## [1] 3.182446
```

So we fail to reject the null at level 0.05 since $-3.18 < -2.38 < 3.18$ (outside the rejection region).

At level 0.1:

```
qt(1 - 0.1/2, 3)
```

```
## [1] 2.353363
```

Here we reject the null (barely), since $-2.38 < -2.35$ (inside the rejection region).

In this case, our p-value is (derived in the same way as with the normal):

$$2P(T \leq -2.38)$$

```
2*pt(-2.38, 3)
```

```
## [1] 0.09761784
```

so we can see why we made a different conclusion at level 0.05 versus level 0.1.

2.8 Unknown σ^2, Large Sample

A general rule is to say that for $n \geq 30$, we have that $S^2 \approx \sigma^2$. Recall that S^2 is unbiased for σ^2:

$$E(S^2) = \sigma^2$$

It is also a *consistent* estimator of σ^2, which in this context means that

$$\lim_{n\to\infty} V(S_n^2) = 0$$

where S_n^2 is the sample variance calculated using a random sample of size n.

If we assume $S^2 \approx \sigma^2$, then

$$\frac{\bar{X} - \mu_0}{s/\sqrt{n}} \approx \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim_{H_0} N(0,1)$$

where $\sim_{H_0}$ indicates that this is the distribution under the assumption that H_0 is true.

Hence we can perform tests using the z-values as we did before.

Note: The t-test is only technically appropriate for a random sample $X_1, \cdots X_n \sim N(\mu, \sigma^2)$. The CLT does not give us a result for a more general random sample.

2.9 Power and Level

We have referred to the two types of error which can occur in a test:

- Type 1 Error: Rejecting H_0 when it is true.
- Type 2 Error: Accepting H_0 when it is false.

Now we introduce some more related terminology:

Definition: A test is *level* α if the probability of making a type 1 error is less than or equal to α.

Definition: The *power* of a test is the probability of rejecting a false null hypothesis.

You will often hear "test at a level α" to refer to "perform a test such that the test has type 1 error probability α".

Clearly, we want our test to have a low probability of comitting either kind of error. It was mentioned earlier that it is often the case that reducing the probability of type 1 error will increase the probability of type 2 error, so finding an appropriate test often involves balancing these two probabilities. The following example illustrates this.

Example: Consider a test with probability 0 of type 1 error. That is, we let $\alpha = 0$. The $1 - \alpha/2$ quantile in this case is just the 1-quantile. The 1-quantile is the number q s.t.

$$1 = \int_{-\infty}^{q} f_X(x)dx$$

so essentially its the upper bound of the range of our random variable, which could be ∞. Our rejection region would therefore be "reject if $Z > \infty$ or $Z < -\infty$", both of which are impossible, so we never reject. Never rejecting means we have probability 0 of rejecting a true null. However, since we never reject, we accept every false null. Therefore our type 2 error probability is 1, and its power is 0.

To get a test which always rejects, let $\alpha = 1$. Then we use the 0.5-quantile (aka the median) to construct our rejection region. For a distribution symmetric about 0, the median is 0, and hence our rejection region is "reject if $Z > 0$ or $Z < -0$", which is true for all values of Z. The power of our test is 1, but the level is 0.

For obvious reasons, these tests are useless, but they highlight the interplay between level and power. The tests we actually use fall somewhere between these two extremes.

2.10 P-Values

Roughly speaking, a p-value is the probability of observing a value of the test statistic that is "at least as extreme" as what was actually observed. We have examples of this in the one-sample tests, which had a rejection region of the form:

$$\text{reject } H_0 \text{ if } T < -t_{n-1,\alpha/2} \text{ or } T > t_{n-1,\alpha/2}$$

(and likewise for the test based on $Z \sim N(0, 1)$). Let T be the random variable corresponding to the test statistic, which we know follows a $t(n-1)$ distribution, and T_{obs} be the observed value of the test statistic, given an observed value of the random sample $x_1, \cdots, x_n$. Then, the p-value p is given by

$$\begin{aligned} p &= P(|T| \geq |T_{obs}|) \\ &= P(T \leq -|T_{obs}|) + P(T \geq |T_{obs}|) \\ &= 2P(T \leq -|T_{obs}|) \end{aligned}$$

where the last line follows from the symmetry of the t distribution.

Tests with rejection regions on either side of the test statistic, like the ones above, are known as *two-sided tests*. That is, the alternative hypothesis is of the form:

$$H_1 : \mu \neq \mu_0$$

Therefore, p-values are calculated in the above way, where we consider the probability of being greater than $|T_{obs}|$ AND the probability of being less than $-|T_{obs}|$ as "extreme" observations (i.e. they suggest that H_0 is false). This is not always the case, however, as will be shown now in the case of a one-sided test.

2.11 One-Sided, One-Sample Tests

Consider the following test:

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu < \mu_0 \end{aligned}$$

Under H_0, which is the same as with the previous tests, we still have:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

The above result can be exact (if the random sample is normal), approximate (if the random sample is not normal and we employ the CLT), or it could be distributed $t(n-1)$ if we use S instead of σ.

However, our rejection region changes. In this case, we consider the following rejection region:

$$\text{reject } H_0 \text{ if } Z < -z_\alpha$$

This is called a *one sided* rejection region, and it changes due to the form of our alternative. Essentially, large values of Z no longer suggest that we should reject H_0 in favor of H_1.

Similarly, if we have the test:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

our rejection region is:

$$\text{reject } H_0 \text{ if } Z > z_\alpha$$

All of these results have equivalent versions for the case of the T statistic as well.

2.12 Two-Sample Tests

A two-sample t-test is often what is being referred to when someone says "t-test". Instead of a null and alternative hypothesis which concern a single population mean, we are now interested in a hypothesis which concerns two population means. Before we do this, we need, as usual, a distributional result which will allow us to calculate a test statistic, of which we know the distribution under H_0, allowing us to perform all the usual steps for statistical tests:

Theorem: Consider two random samples of size n_1 and n_2, respectively: $X_1, \cdots X_{n_1} \sim N(\mu_1, \sigma_1^2)$ and $Y_1, \cdots Y_{n_2} \sim N(\mu_2, \sigma_2^2)$. Then,

$$Z := \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0, 1)$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of the individual random samples.

The setup for a two sample test in which we know the values of σ_1^2 and σ_2^2 is clear:

$$H_0 : \mu_1 - \mu_2 = d$$
$$H_1 : \mu_1 - \mu_2 \neq d$$

Under H_0, we have that:

$$Z := \frac{(\bar{X} - \bar{Y}) - d}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0, 1)$$

Testing is performed in exactly the same way as we have done previously, for the "known variance" tests. The most common two sample test is when we choose $d = 0$, which means we test:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

That is, whether or not the means of the two populations are equal. This has boundless applications to real-world experiments.

2.13 Unknown But Equal Variance

Theorem: Consider two random samples of size n_1 and n_2, respectively: $X_1, \cdots X_{n_1} \sim N(\mu_1, \sigma^2)$ and $Y_1, \cdots Y_{n_2} \sim N(\mu_2, \sigma^2)$ (the distributions they come from have a shared variance σ^2). Then,

$$T := \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(1/n_1 + 1/n_2\right)}} \sim t(n_1 + n_2 - 2)$$

where

$$S_p^2 := \frac{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of the individual random samples, and S_1^2 and S_2^2 are the sample variances of the individual random samples.

2.14 Paired t Tests

Consider two random samples again, both of size n: $X_1, \cdots X_n \sim N(\mu_1, \sigma_1^2)$ and $Y_1, \cdots Y_n \sim N(\mu_2, \sigma_2^2)$.

Define the random variables $D_1, \cdots, D_n$ as follows:

$$\begin{aligned} D_1 &= X_1 - Y_1 \\ D_2 &= X_2 - Y_2 \\ &\vdots \\ D_n &= X_n - Y_n \end{aligned}$$

Suppose we want to test if the $\mu_1 = \mu_2$ in a paired sense. Then we have:

$$\begin{aligned} H_0 &: \mu_D = d_0 \\ H_1 &: \mu_D \neq d_0 \end{aligned}$$

where μ_D is the mean of the $D_1, \cdots D_n$ random sample. The test statistic is given by:

$$T = \frac{\bar{D} - d_0}{S_d/\sqrt{n}} \sim_{H_0} t(n-1)$$

where S_d^2 is the sample variance of the D's.

Note: The above relies on the result that the difference of two normal random variables is still normal. The general result is as follows:

Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ be independent normal random variables. Then:

$$X - Y \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

A paired t-test is used for observations which have some sort of natural pairing. For example, X_i could be subject i's heart rate before exercise, and Y_i their heart rate after exercise. We might want to test if there is a difference between heart rate before and after exercise *in general*, not just with regards to a particular subject, in which case we could do a paired t-test.

2.15 Sampling Distribution of the Sample Variance S^2

Recall that the sample variance is given by:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Theorem: Consider a random sample $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$. Then

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

That is, the sample variance follows a χ^2 distribution with $n-1$ degrees of freedom.

The proof of the above theorem is omitted here, but follows from the fact that $(n-1)S^2/\sigma^2$ can be represented as the sum of the squares of $n-1$ standard normal random variables. Then we employ the previous result that this quantity must follow a $\chi^2(n-1)$ distribution.

This lends itself to the following test:

$$H_0 : \sigma^2 = \sigma_0^2$$
$$H_1 : \sigma^2 \neq \sigma_0^2$$

Under the assumption that H_0 is true:

$$X^2 := \frac{(n-1)S^2}{\sigma_0^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

Therefore, in order to perform this test, we calculate the test statistic X^2 and use the following rejection region:

$$\text{reject } H_0 \text{ if } X^2 < \chi^2_{n-1,1-\alpha/2} \text{ or } X^2 > \chi^2_{n-1,\alpha/2}$$

Note that here we can't use the same quantile twice and flip its sign, because the chi-squared distribution is *not symmetric*, and is only defined for the positive real numbers (it is a special case of the gamma distribution). $n-1$ is the degrees of freedom of the χ^2 distribution from which the quantile comes from.

At level $\alpha = 0.05$ with $n - 1 = 5$ degrees of freedom, for example, our quantiles would be:

```
qchisq(1-0.05/2, 5)
```

```
## [1] 12.8325
```

and

```
qchisq(0.05/2, 5)
```

```
## [1] 0.8312116
```

Example: A common use of the above test statistic is when we wish to consider the following null with a one-sided alternative:

$$H_0 : \sigma^2 = \sigma_0^2$$
$$H_1 : \sigma^2 > \sigma_0^2$$

That is, is the variance equal to a certain null value σ_0^2, or is it larger. Consider the following random sample, which we assume to be normal:

$$(16, 26, -9, 7, 22)$$

Suppose that we want to test the hypothesis:

$$H_0 : \sigma^2 = 100$$
$$H_1 : \sigma^2 > 100$$

Our test statistic X^2 is:

$$X^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{(5-1)(194.3)}{100} \approx 7.77$$

Our one-sided rejection region should look like:

$$\text{reject } H_0 \text{ if } X^2 > \chi^2_{n-1,\alpha}$$

At a level $\alpha = 0.05$, we have the quantile χ^2_α to be:

```
qchisq(1-0.05, 4)
```

```
## [1] 9.487729
```

The value of $7.77 < 9.49$, so we fail to reject H_0 and conclude that the variance is equal to 100.

2.16 Categorical Data and Contingency Tables

Categorical data is when the set of possible outcomes consists of a set of categories.

For example, the following data was collected on passengers of the Titanic. "Survived" was reported as belonging to one of two categories: "Yes" and "No":

```
## Survived
##   No  Yes
## 1490  711
```

"Class" was reported as belonging to one of four categories: "1st", "2nd", "3rd", and "Crew":

```
## Class
##  1st  2nd  3rd Crew
##  325  285  706  885
```

The numbers above represent how many passengers fell into each category. If you were sum the numbers in either table, you would get 2201, the total number of passengers on the titanic, each of whom was assigned a certain category.

We can combine the two tables above into a single table, where the rows represent whether or not someone survived, and the columns represent which class they were in:

```
##          Class
## Survived 1st 2nd 3rd Crew
##      No  122 167 528  673
##      Yes 203 118 178  212
```

This is called a *contingency table*. In particular, this is a two-way contingency table (there are two ways in which we categorize each passenger).

A natural question when analyzing this data may be: does whether or not a passenger survived depend on their class? We will construct a parametric statistical test for this hypothesis.

2.17 Distribution of Categorical Data

One way to think about a set of categories is to consider them as a discrete random variable where each category is represented by a number. For example, consider the categorical variable "Survived" for which we have two categories: "Yes" and "No". We could say that X is a random variable which obeys the following:

$$X = \begin{cases} 0 & \text{if Survived} = \text{No} \\ 1 & \text{if Survived} = \text{Yes} \end{cases}$$

Clearly, X is a Bernoulli random variable. However, we observe whether or not each of the 2201 passengers on the Titanic survived or died. This amounts to 2201 Bernoulli trials. Therefore, we let X be *binomial*, which is equal to the sum of independent Bernoulli random variables (independent meaning we must assume that whether or not a passenger survived is

independent of whether or not the other passengers survived). A binomial X would have the following relationship to the data:

$$X = \text{number of passengers who survived} \sim \text{binom}(2201, p)$$

Note: Based purely on how many people survived/died, we could estimate the parameter p using the method of moments (and our random sample of size 1):

$$E(X) = np = 771 \implies \hat{p} = \frac{771}{2201} \approx 0.35$$

Now we consider the following result, which follows from the central limit theorem, which will help us establish some distributional results relating to binomial data:

Corollary: Let $X \sim \text{binom}(n, p)$. We have seen that $E(X) = np$ and $V(X) = np(1-p)$. Define Z as:

$$Z := \frac{X - np}{\sqrt{np(1-p)}}$$

Then as $n \to \infty$, $Z \sim N(0, 1)$.

Proof: Recall that $X \sim \text{binom}(n, p)$ is the sum of n independent and identically distributed Bernoulli random variables, that is

$$X = \sum_{i=1}^{n} Y_i \qquad \text{where each } Y_i \sim \text{bern}(p)$$

If we consider $\frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$, we have that the central limit theorem gives:

$$Z := \frac{\bar{Y} - p}{\sqrt{p(1-p)/n}} \sim N(0, 1)$$

$\bar{Y} = \frac{1}{n}X$, so substituting in,

$$
\begin{aligned}
Z &= \frac{X/n - p}{\sqrt{p(1-p)/n}} \\
&= \frac{X - np}{\sqrt{np(1-p)}} \sim N(0,1)
\end{aligned}
$$

■

Note: The above approximate requires large n, but also requires that p be not too far from 0 or 1. If p is close to $1/2$, the approximation can be good, even when n is small.

The binomial distribution gives us a way to describe categorical variables with two categories as *random variables*, of which we can write the probability mass function. In the case of a categorical variable with more than two categories, we need to introduce the *multinomial distribution*:

Definition: A k dimensional random vector X is said to follow a *multinomial distribution* ($X \sim \text{multinomial}(n, p_1, \cdots, p_k)$ if X has pmf:

$$
f(x_1, \cdots, x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} \qquad \text{for } x_1, \cdots, x_k = 0, 1, 2, \cdots
$$

Now suppose Y is the categorical variable corresponding to the number of passengers in each class on the Titanic. Then we could have:

$$
Y \sim \text{multinomial}(2201, p_1, p_2, p_3, p_4)
$$

where Y is a 4 dimensional random vector, and $p_1, \cdots, p_4$ are the probabilities of falling into each of the four classes.

In general, we can say that a categorical variable X Consider the case of two categorical variables, X and Y. If X has I categories, and Y has J categories, then we can use and $I \times J$ matrix to display all possible combinations of outcomes. As was introduced above, this matrix is known as a two-way contingency table.

2.18 Chi-Squared Tests for Independence

Consider the k-dimensional random vector $X = (X_1, \cdots, X_k) \sim$ multinomial$(n, p_1, \cdots, p_k)$. Then each X_i is marginally distributed binomially. That is, for each $i = 1, \cdots, k$,

$$X_i \sim \text{binom}(n, p_i)$$

Therefore, by the result above,

$$\frac{X_i - np_i}{\sqrt{np_i(1-p_i)}} \sim N(0,1)$$

Furthermore, for large n (for which the approximation above holds), and for np_i which approaches a constant as $n \to \infty$,

$$np_i(1-p_i) = np_i\left(1 - \frac{np_i}{n}\right) \to np_i$$

Therefore

$$\frac{X_i - np_i}{\sqrt{np_i}} \sim N(0,1) \implies \left(\frac{X_i - np_i}{\sqrt{np_i}}\right)^2 \sim \chi^2(1)$$

by properties of a χ^2 distribution (this result was stated earlier in the "relationships between distributions" section. A $\chi^2(k)$ random variable is equal to the sum of k *squared* standard normal random variables).

Each $\left(\frac{X_i - np_i}{\sqrt{np_i}}\right)^2$ are *not* independent, and are correlated. We can, however, represent their sum as the sum of $n-1$ independent $\chi^2(1)$ random variables (in a similar way to S^2), which eventually leaves the following result:

$$\sum_{i=1}^{k} \left(\frac{X_i - np_i}{\sqrt{np_i}}\right)^2 \sim \chi^2(k-1)$$

where recall that k is the number of categories of our multinomial X, and n is a parameter of the multinomial distribution.

Example: Before we tackle the Titanic example, consider a simple example of a die.

An experiment which consists of 300 die rolls can be thought of as observing the value of a multinomial random variable. In particular:

$$X \sim \text{multinomial}(300, p_1, p_2, \cdots, p_6)$$

where p_i represents the probability of obtaining side i when the dice is rolled a single time.

Consider the following null and alternative hypothesis:

$$H_0 : p_1 = p_2 = \cdots = p_6 = \frac{1}{6} \text{ (i.e. the die is fair)}$$
$$H_1 : \text{For some } i \neq j \ , \ p_i \neq p_j \text{ (i.e. the die is not fair)}$$

Consider the statistic which we derived above:

$$\sum_{i=1}^{k} \left(\frac{X_i - np_i}{\sqrt{np_i}} \right)^2 \sim \chi^2(k-1)$$

Under the null hypothesis,

$$np_i = 300 * \frac{1}{6} = 50$$

Therefore, we have that, for this particular null and for 300 die rolls, *assuming that H_0 is TRUE*:

$$X^2 := \sum_{i=1}^{k} \left(\frac{X_i - 50}{\sqrt{50}} \right)^2 = \sum_{i=1}^{k} \left(\frac{X_i - np_i}{\sqrt{np_i}} \right)^2 \sim \chi^2(k-1)$$

So suppose that we perform the experiment and toss 300 die, and we observe the following:

```
##  1  2  3  4  5  6
## 34 60 81 91 17 17
```

To calculate X^2, we showed above that:

$$np_i = 300 * \frac{1}{6} = 50$$

(remember, this is given the form of the null and the sample size of this particular experiment).

Then,

$$\begin{aligned} X^2 &:= \sum_{i=1}^{k} \left(\frac{X_i - 50}{\sqrt{50}} \right)^2 \\ &= \left(\frac{34 - 50}{\sqrt{50}} \right)^2 + \left(\frac{60 - 50}{\sqrt{50}} \right)^2 + \cdots + \left(\frac{81 - 50}{\sqrt{50}} \right)^2 \\ &= 103.52 \end{aligned}$$

Using `R` to calculate it:

```
# v is the vector of die roll outcomes
sum(((v - 50)/sqrt(50))^2)
```

```
## [1] 103.52
```

This is immense, so we're going to reject H_0. Lets calculate the quantile at level $\alpha = 0.05$ to be sure:

```
qchisq(1-0.05, 6-1)
```

```
## [1] 11.0705
```

$103.52 > 11.07$ so we reject the null and conclude that the die is not fair.

Note that this is a one-sided rejection region, because large values of X^2 indicate that our observed X_i's are *far* from their expected value np_i (under the null). Therefore large values of X^2 suggest evidence against H_0, in this context.

Example: The Titanic.

Recall the contingency table with which this discussion was introduced:

```
##          Class
## Survived 1st 2nd 3rd Crew
##       No  122 167 528  673
##       Yes 203 118 178  212
```

We want to test the following hypothesis:

$$H_0 : \text{The row variable and column variable are independent}$$
$$H_1 : \text{The row variable and column variable are dependent}$$

What does independence entail in this situation? Consider each cell in the matrix to have a certain probability of falling in that cell. For simplicity, a $2x2$ matrix could have cell probabilities as follows:

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

We could rephrase the null of "rows and columns are independent" as follows:

$$H_0 : \text{Cell probabilities } \{p_{ij}\} \text{ are such that } p_{ij} = p_{i+}p_{+j}$$

Where p_{i+} is the probability of being in row i, ignoring the column, and p_{+j} is the probaility of being in column j, ignoring the row.

In the Titanic example:

```
##          Class
## Survived 1st 2nd 3rd Crew
##       No  122 167 528  673
##       Yes 203 118 178  212
```

p_{1+} would be the probability of not surviving.

p_{+1} would be the probability of being in first class.

If class and survival were independent, we would expect p_{11}, the probability of both not surviving and being in first class to be equal to the product of the probability of the two cases, i.e:

$$p_{11} = p_{1+}p_{+1}$$

This is the definition of independent events that we saw during the probability theory portion of this class.

Therefore, under H_0 that the rows and columns are independent, our contingency table should have cell probabilities as follows:

$$\begin{bmatrix} p_{1+}p_{+1} & p_{1+}p_{+2} & p_{1+}p_{+3} & p_{1+}p_{+4} \\ p_{2+}p_{+1} & p_{2+}p_{+2} & p_{2+}p_{+3} & p_{2+}p_{+4} \end{bmatrix}$$

We can obtain these "marginal" probabilities by calculating proportions. We assume that the row and column margin sums are fixed. Under this assumption, the proportions are exactly equal to the marginal probabilities, though they are really just an estimate. For example, out of the 2201 passengers, 1490 did not survive:

```
## Survived
##   No  Yes
## 1490  711
```

So we would calculate

$$p_{1+} = \frac{1490}{2201} \approx 0.677$$

and using the other marginal table:

```
## Class
##  1st  2nd  3rd Crew
##  325  285  706  885
```

we can calculate

$$p_{+1} = \frac{325}{2201} \approx 0.148$$

leaving

$$p_{11} = p_{1+}p_{+1} = 0.148 * 0.677 = 0.10$$

We can do this for *every* marginal probability.

Therefore, under H_0 (i.e. the rows and columns are independent), we can produce a table of cell probabilities:

```
tab <- marginSums(Titanic, c("Survived", "Class"))
prop_h0 <- matrix(1:8, nrow = 2)
for (i in 1:2) {
  for (j in 1:4) {
    p <- sum(tab[i, ])/2201 * sum(tab[, j])/(2201)
    prop_h0[i, j] <- p
  }
}
round(prop_h0, 2)
```

```
##      [,1] [,2] [,3] [,4]
## [1,] 0.10 0.09 0.22 0.27
## [2,] 0.05 0.04 0.10 0.13
```

and a table of expected counts (by multiplying each cell by 2201):

```
round(prop_h0*2201, 2)
```

```
##        [,1]   [,2]   [,3]   [,4]
## [1,] 220.01 192.94 477.94 599.11
## [2,] 104.99  92.06 228.06 285.89
```

All this is to say: our original H_0 of "independent rows and columns" can be rephrased in the following way. Each of the 8 entries in the matrix is considered to be one of the observations in a multinomial distribution with 8 possible outcomes.

$$X \sim \text{multinomial}(2201, p_1, p_2, \cdots, p_8)$$

Let $p_{01}, \cdots, p_{08}$ be the probabilities from the table of expected counts above. That is, $p_{01} = 0.10, p_{02} = 0.09, \cdots, p_{07} = 0.10, p_{08} = 0.13$. Then,

$$\begin{aligned} H_0 &: p_1 = p_{01}, p_2 = p_{02}, \cdots, p_8 = p_{08} \\ H_1 &: \text{The probabilities are not as above.} \end{aligned}$$

Then, under H_0,

$$X^2 := \sum_{i=1}^{k} \left(\frac{X_i - np_{0i}}{\sqrt{np_{i0}}} \right)^2 = \sum_{i=1}^{k} \left(\frac{X_i - np_i}{\sqrt{np_i}} \right)^2 \sim \chi^2((r-1)(c-1))$$

We obtain the degrees of freedom using the following equation:

$$\text{degrees of freedom} = (c-1)(r-1)$$

where c is the number of columns of the table, and r is the number of rows. This follows from the fact that we assumed that the marginal sums across the rows and columns are fixed (this comes from the way we estimated the p_{i+} and p_{+j} probabilities). Knowing $(c-1)(r-1)$ of the counts in the table will necessarily determine all the other counts, as long as we assume that we know what the marginal counts are.

np_{0i} is the "expected count" from the table above. e.g.

$$\begin{aligned} np_{01} &= 2201 * 0.10 = 220.1 \\ &\vdots \\ np_{08} &= 2201 * 0.13 = 286.13 \end{aligned}$$

Then,

$$\begin{aligned} X^2 &:= \sum_{i=1}^{k} \left(\frac{X_i - np_{0i}}{\sqrt{np_{i0}}} \right)^2 \\ &= \left(\frac{122 - 220.10}{\sqrt{220.10}} \right)^2 + \cdots + \left(\frac{212 - 286.13}{\sqrt{286.13}} \right)^2 \\ &= 190.4 \end{aligned}$$

Using `R` to calculate it:

```r
v <- as.vector(prop_h0*2201)
ex <- as.vector(tab)
sum(((v - ex)/sqrt(v))^2)
```

```
## [1] 190.4011
```

The quantile at level $\alpha = 0.05$:

```r
qchisq(1-0.05, 3)
```

```
## [1] 7.814728
```

so we reject H_0 that the row and column variables are independent, and conclude that they are dependent.

2.19 Confidence Intervals

Definition: An interval estimate of some population parameter θ is an estimate of the form:

$$\hat{\theta}_L < \theta < \hat{\theta}_U$$

That is, we find a lower bound $\hat{\theta}_L$ and an upper bound $\hat{\theta}_U$ for the population parameter θ.

Much like with errors in statistical tests, there is no way for us to construct a (finite) interval like the one above that is guaranteed to contain the true value of θ. Rather, the accuracy of our interval estimates are phrased in a probabilistic way.

The values of $\hat{\theta}_L$ and $\hat{\theta}_U$ are calculated using a random sample. That is, they are functions of our random sample $X_1, \cdots, X_n$ and are therefore *statistics*. In particular, they are themselves random variables. We can therefore write the following probability:

$$P(\hat{\theta}_L < \theta < \hat{\theta}_U)$$

This probability could be calculated, should we know the true value of the parameter θ, which would at that point be a constant.

Definition: If we have that

$$P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1 - \alpha$$

then this is called a $100 * (1 - \alpha)\%$ *confidence interval.*

The most common is a 95% confidence interval, in the case when $\alpha = 0.05$.

The process for determining confidence intervals is very similar to testing. In fact, every distributional result which yields a test also yields a confidence interval, and we will mostly talk about confidence intervals which follow from a distributional result which we have already seen.

2.20 Mean of a Single Sample

If $X_1, \cdots X_n \sim N(\mu, \sigma^2)$ is a random sample, then

$$Z := \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Therefore,

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

by definition of the $z_{\alpha/2}$ quantile.

This means:

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

So this is hence a $100(1 - \alpha)\%$ confidence interval with:

$$\hat{\theta}_L = \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$
$$\hat{\theta}_U = \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

We can only calculate this confidence interval if σ is known (or can be estimated well by s in the case when $n \geq 30$)

2.21 One-Sided Bounds

In a similar manner to a one-sided test, we can produce one-sided bounds. Here, we use the result that

$$P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_\alpha\right) = 1-\alpha$$

So we can rearrange and obtain a one-sided, $100(1-\alpha)\%$ confidence interval for μ:

$$P\left(\mu > \bar{X} - \frac{z_\alpha \sigma}{\sqrt{n}}\right) = 1-\alpha$$

and likewise for the other direction of inequality.

2.22 Non-Normal Random Sample

If $X_1, \cdots X_n$ is a random sample from a distribution with mean μ and variance σ^2, then as $n \to \infty$,

$$Z := \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

So all of the confidence intervals above hold (but are approximate).

2.23 Unknown σ^2

If $X_1, \cdots X_n \sim N(\mu, \sigma^2)$ is a random sample, then

$$T := \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

Therefore,

$$P\left(-t_{n-1,\alpha/2} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{n-1,\alpha/2}\right) = 1-\alpha$$

Which leaves:

$$P\left(\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

2.24 Standard Error

Recall that due to the central limit theorem,

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

Since $\bar{X}$ is a statistic, we call this its *sampling distribution*. Furthermore, the square root of its variance (i.e. its standard deviation) is referred to as its *standard error*:

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Note that our confidence intervals were generally of the form:

$$\bar{X} \pm z_{\alpha/2}SE(\bar{X})$$

If we must estimate σ using s, we call this the estimated standard error and denote it with a hat:

$$\hat{SE}(\bar{X}) = \frac{s}{\sqrt{n}}$$

and our confidence intervals would be:

$$\bar{X} \pm t_{\alpha/2}\hat{SE}(\bar{X})$$

Generally, if we have some kind of statistic which is asymptotically normal (like $\bar{X}$), we can construct a confidence interval using its standard error like we have above. It does not always have to be a confidence interval for the population mean μ.

2.25 Proportions

Suppose we have a random variable X which represents the number of successes in n trials (it could be binomial, in the case of independent trials). Let p be the population parameter which represents the probability of success. Then:

$$\hat{p} = \frac{X}{n}$$

is the natural way to estimate this parameter. For a binomial, this is both the method of moments estimator and the maximum likelihood estimator.

Theorem: $\hat{p} = \frac{X}{n}$ is such that it has distribution:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

asymptotically as $n \to \infty$.

In this case,

$$\hat{SE}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

So our $100(1-\alpha)\%$ confidence interval is given by:

$$\hat{p} \pm z_{\alpha/2}\hat{SE}(\hat{p})$$

where we use z values under the assumption that n is large.

2.26 Odds Ratio

Consider the following example: an experiment consists of exposing a subject to some exposure, and then seeing if that subject develops a certain disease. A potential experimental outcome could be as follows:

```
dat
```

```
##             Diseased Healthy
## Exposed            20     380
## Not Exposed         6     594
```

This is an example of a 2×2 contingency table, as was discussed before. There are two categorical variables, each with two outcomes.

We can perform a χ^2 test to test if the rows and columns are independent. Using R's `chisq.test` function,

```
chisq.test(dat)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  dat
## X-squared = 13.625, df = 1, p-value = 0.0002232
```

Based on the p-value, we would reject the null of independence at level $\alpha = 0.05$, and conclude that exposure and disease are not independent.

Another way of considering analyzing this data is via the *odds ratio*.

Definition: For a binary event with probability of success p, the odds of success are defined to be

$$\text{odds} = \frac{p}{1-p}$$

Definition: Consider a 2×2 contingency table with probability p_1 of success in row 1, and probability p_2 of success in row 2. Then the *odds ratio* is defined to be:

$$\text{OR} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

i.e. it is the ratio of the odds of the two rows.

Definition: Given a 2×2 contingency table

$$\begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}$$

then the *sample odds ratio* is given by

$$\hat{\text{OR}} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}}$$

In the example above, this would be the odds of getting the disease if exposed divided by the odds of getting the disease if not exposed:

$$\hat{\text{OR}} = \frac{20/380}{6/594} = 5.21$$

If the (sample) odds ratio is greater than 1, the odds of "success" are *higher* in row 1 than in row 2. The value of 5.21 above suggests that the odds of getting the disease given that you are exposed (row 1) are 5.21 times the odds of getting the disease given that you are not exposed (row 2). Likewise, an odds ratio smaller than 1 would suggest that the odds of "success" are *lower* in row 1 than in row 2.

Theorem: Let $\hat{\text{OR}}$ be the odds ratio obtained from a 2×2 contingency table. Then

$$\log(\hat{\text{OR}}) \overset{\text{approx}}{\sim} N\left(\log(\text{OR}), \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)$$

That is, the natrual logarithm of the odds ratio follows an approximately normal distribution.

We can therefore employ the same process for constructing a confidence interval for any quantity which has an approximate normal distribution:

$$\hat{SE}(\log(\hat{\text{OR}})) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

So our $100(1-\alpha)\%$ confidence interval is given by:

$$\log(\hat{\text{OR}}) \pm z_{\alpha/2}\hat{SE}(\log(\hat{\text{OR}}))$$

Exponentiating the endpoints of this confidence interval for $\log(\hat{\text{OR}})$ will yield a confidence interval for $\hat{\text{OR}}$, as will be shown in the following example.

Example: Consider the table given in the introduction:

```
##              Diseased Healthy
## Exposed            20     380
## Not Exposed         6     594
```

The sample odds ratio is:

$$\hat{\text{OR}} = \frac{20/380}{6/594} = 5.21$$

which means the log odds are:

$$\log(\hat{\text{OR}}) = \log(5.21) = 1.65$$

with standard error:

$$\begin{aligned} \hat{SE}(\log(\hat{\text{OR}})) &= \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \\ &= \sqrt{\frac{1}{20} + \frac{1}{380} + \frac{1}{6} + \frac{1}{594}} = 0.47 \end{aligned}$$

For a 95% confidence interval, we use $z_{0.025} = 1.96$:

$$1.65 \pm 1.96 * 0.47$$

so

$$0.73 < \log(\text{OR}) < 2.57$$

Exponentiating, we get:

$$e^{0.73} < e^{\log(\text{OR})} < e^{2.57}$$
$$2.07 < \text{OR} < 13.07$$

a 95% confidence interval for the true odds ratio OR.

The Fundamental Theorem of Epidemiology:

Consider a contingency table like the one in the example:

```
##              Diseased Healthy
## Exposed            20     380
## Not Exposed         6     594
```

Let D be the event of getting the disease, E be the event of being exposed. Then,

$$\text{OR} = \frac{P(D|E)/P(D^c|E)}{P(D|E^c)/P(D^c|E^c)} = \frac{P(E|D)/P(E^c|D)}{P(E|D^c)/P(E^c|D^c)}$$

Proof: Apply Bayes' Rule four times.

The importance of the above result is as follows. Our original interpretation of the odds ratio was "the odds of getting the disease if exposed divided by the odds of getting the disease if not exposed". In this situation, the disease is the response and the exposure is an explanatory variable.

In a case-control study, subjects are selected on the basis of being diseased or healthy. Consider the following example:

```
##              Diseased Healthy
## Exposed           172     173
## Not Exposed        90     346
```

There were 262 diseased and 519 healthy patients selected, with the experimentors aware of whether or not they were diseased or healthy. The outcome is considered to be whether or not the selected individual was ever exposed.

We can still estimate the following ratio:

$$\frac{P(E|D)/P(E^c|D)}{P(E|D^c)/P(E^c|D^c)}$$

Our estimates for the above probabilities are as follows:

$$\begin{aligned} P(E|D) &= 172/262 \approx 0.66 \\ P(E^c|D) &= 90/262 \approx 0.34 \\ P(E|D^c) &= 173/519 \approx 0.33 \\ P(E^c|D^c) &= 346/519 \approx 0.67 \end{aligned}$$

so

$$\frac{P(E|D)/P(E^c|D)}{P(E|D^c)/P(E^c|D^c)} = (0.66/0.34)/(0.33/0.67) \approx 3.82$$

Note: this could also be written as $(n_{11}/n_{12})/(n_{21}/n_{22}) = (172/173)/(90/346)$.

This is, by the above theorem, equal to the odds ratio given by "the odds of getting the disease if exposed divided by the odds of getting the disease if not exposed". That is:

$$\frac{P(D|E)/P(D^c|E)}{P(D|E^c)/P(D^c|E^c)} \approx 3.94$$

Therefore we can conclude that the odds of getting the disease if exposed is about 3.94 times larger than the odds of getting the disease if not exposed. This is true even though our data only allowed us to estimate the reverse: the odds of being exposed given that the subject is diseased.

A 95% confidence interval can be calculated in the same manner as above:

$$\log(\hat{\text{OR}}) = \log(3.94) = 1.37$$

$$\hat{SE}(\log(\hat{\text{OR}})) = \sqrt{\frac{1}{172} + \frac{1}{173} + \frac{1}{90} + \frac{1}{346}} = 0.16$$

$$1.37 \pm 1.96 * 0.16$$

so

$$1.06 < \log(\text{OR}) < 1.68$$

Exponentiating, we get:

$$e^{1.06} < e^{\log(\text{OR})} < e^{1.68}$$
$$2.88 < \text{OR} < 5.38$$

a 95% confidence interval for the true odds ratio OR.

3 Linear Regression

3.1 Linear Regression

The linear regression model is as follows: given data $Y_1, \cdots, Y_n$, and (n dimensional) covariate vectors $X_1, \cdots, X_p$, we have that

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i \qquad i = 1, \cdots, n$$

Y is a numeric *response* variable (numeric in this case meaning not categorical; it takes a real numbers as values). $X_1, ..., X_p$ are *predictor* variables. They can be numeric or categorical. We will start by considering numeric predictor variables.

ϵ is a random error term. We will start by assuming the following:

$$\epsilon_i \sim N(0, \sigma^2) \qquad i = 1, \cdots, n$$

The important thing to note is that linear regression, as formulated above, is a parametric model being specified for the sample $Y_1, \cdots, Y_n$, where we have that

$$Y_i \sim N(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}, \sigma^2) \qquad i = 1, \cdots, n$$

Knowing the distribution of each Y_i will allow us to go through the same statistical procedures as were presented in chapter 2, but in the context of this linear regression model. In particular, assigning a distribution to the Y_is will allow us to write a likelihood function of the parameters, perform maximum likelihood estimation, determine the distribution of the maximum likelihood estimates, construct Bayes' estimates, perform hypothesis tests, construct confidence intervals, and so on, largely in the same manner as chapter 2.

Our first goal is to estimate the value of each of the coefficients β_i, given the data: Y and X values. This is done using a process called least squares.

Note that:

- $\beta_1 \cdots \beta_p$ are called the "regression coefficients"

- When we calculate an estimate for a regression coefficient β_i, we, as usual, denote this with a "hat": $\hat{\beta}_i$
- The error term ϵ is always assumed to come from a distribution with mean 0 and finite variance. When it is further assumed that ϵ is normal, it allows us to perform t and F tests, in a very similar way to what we have done before.

There are two main uses for linear regression:

- Prediction. *Regression* is generally used to refer to the prediction of quantitative outputs.
- Causal Inference. Whether or not the response is related to one or more of the predictor variables (i.e. whether or not there exists a causal relationship).

3.2 Simple Linear Regression

Simple linear regression refers to the case where we have a single predictor variable (X):

$$Y = \beta_0 + \beta_1 X + \epsilon$$

In order to understand the situations in which we use linear regression, consider the following examples of some data.

Consider the following simulation in `R`:

```
x <- rnorm(100, 0, 1)
eps <- rnorm(100, 0, 1)
beta0 <- 5
beta1 <- 0.5
y <- beta0 + beta1*x + eps
```

`x` and `y` are vectors which contain 100 values each. Each value Y corresponds with a value of X:

```
head(data.frame("Y" = y, "X" = x))
```

```
##          Y          X
## 1 5.703703  0.2244980
## 2 6.397119  1.2995818
```

```
## 3 5.000033 -1.6395497
## 4 3.961134  0.4809277
## 5 5.035360  0.4408339
## 6 3.892211  0.2517918
```

Furthermore, this simulation guarantees that the assumptions of the linear regression model are satisfied. That is, by construction, we have that:

$$Y = 5 + 0.5X + \epsilon \qquad \text{where } \epsilon \sim N(0, 1)$$

The simulation allows us to know the "true" values of the following parameters:

$$\begin{aligned} \beta_0 &= 5 \\ \beta_1 &= 0.5 \\ \sigma^2 &= 1 \end{aligned}$$

The situation of interest, however, is when we have data (i.e. many observations of values of Y and X), but we don't know the parameter values. We start by assuming that the linear model holds, and then create methods for estimating these parameters under the assumption of that model.

3.3 Least Squares

Least Squares is the process of calculating estimates for β_0 and β_1 such that the squared difference between "predicted points" and "actual values" of Y are minimized.

Consider simple linear regression, where we have a response Y and a predictor X. Suppose that we make n observations of Y and X, and we denote these n observations by:

$$\begin{aligned} Y_1, \cdots, Y_n \\ X_1, \cdots, X_n \end{aligned}$$

The method of least squares means what we obtain estimates for β_0 and β_1 by choosing $\hat{\beta}_0$ and $\hat{\beta}_1$ such that they minimize the following quantity, which we denote using Q:

$$Q := \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

We can minimize it using calculus by differentiating it with respect to β_0 and β_1 and setting the partials equal to zero:

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0$$
$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) X_i = 0$$

which yields the following system of equations (called the *normal equations*):

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i$$
$$\sum_{i=1}^{n} X_i Y_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2$$

which we solve to yield the following estimates for β_0 and β_1:

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n \sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

3.4 Alternate Form

The equations for $\hat{\beta}_0$ and $\hat{\beta}_1$ can also be expressed as follows:

$$\hat{\beta}_1 = \frac{\text{cov}(X, Y)}{\text{var}(X)}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

where cov and var refer to the *sample* covariance and variance of the samples X and Y, given by:

$$\text{cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$
$$\text{var}(X) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

If we consider the simulated example from above, here's how we could calculate this easily in R:

```
beta1hat <- cov(x,y)/var(x)
beta0hat <- mean(y) - beta1hat*mean(x)
beta1hat
```

```
## [1] 0.5358842
```

```
beta0hat
```

```
## [1] 5.055137
```

Since we simulated our data, we can compare our estimate regression coefficients to their true values:

```
abs(beta1 - beta1hat)
```

```
## [1] 0.03588424
```

```
abs(beta0 - beta0hat)
```

```
## [1] 0.05513655
```

Example: The following data set [2] which contains `Size`, the size of the head in centimeters cubed, and `Weight`, the weight of the brain in grams. Here are the first four rows:

```
head(brainhead, 4)
```

```
##   Size Weight
## 1 4512   1530
## 2 3738   1297
## 3 4261   1335
## 4 3777   1282
```

Let the predictor X be `Size`, and the response Y be `Weight`.

Summing over these quantities, we have the following:

$$\sum_{i=1}^{237} X_i = 861256$$
$$\sum_{i=1}^{237} Y_i = 304041$$
$$\sum_{i=1}^{237} X_i Y_i = 1113176805$$
$$\sum_{i=1}^{237} X_i^2 = 3161283190$$

Using the least squares formulas for the regression coefficients, we obtain:

$$\hat{\beta}_0 = 325.57$$
$$\hat{\beta}_1 = 0.263$$

The *fitted regression equation* allows us to estimate Y for any value of X using the above estimates for the coefficients. That is, let $\hat{Y}$ denote our estimate of the value of Y, then

$$\begin{aligned}\hat{Y} &= \hat{\beta}_0 + \hat{\beta}_1 X \\ &= 325.57 + 0.263X\end{aligned}$$

If we wanted, for example, to estimate the weight of someone's brain who has a head volume of 4000 cm^3, we could use the regression equation as follows:

$$\hat{Y} = 325.57 + 0.263 * 4000 = 1377.57 \text{ grams}$$

The equation of a line defined by $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ is known as the *fitted line*. When plotted through a scatter plot of the data, you'll often hear this referred to as the *line of best fit*:

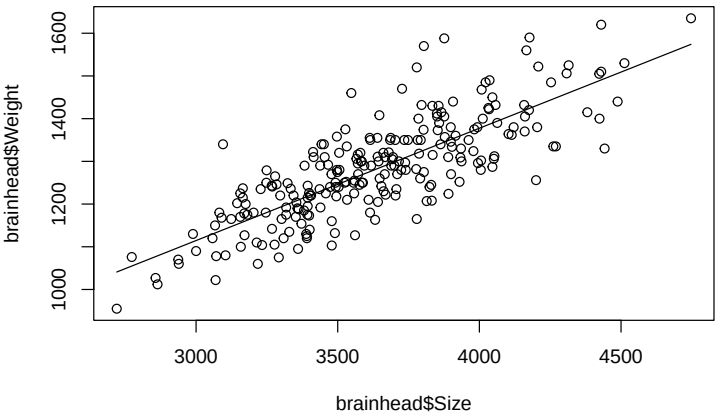


Figure 10: The fitted line.

3.5 Relationship to Maximum Likelihood

Recall the linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon$$

where we have that:

$$\epsilon \sim N(0, \sigma^2)$$

The covariates $X_{i1}, \cdots X_{ip}$ are fixed constants. It follows that:

$$Y \sim N(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}, \sigma^2)$$

Or shorthand:

$$Y \sim N(X_i^T \beta, \sigma^2)$$

where X_i and β are p-dimensional vectors.

In the simple linear regression case, we have that:

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

It follows that given a sample of Y_i's, we can write down the likelihood (i.e. the joint pdf of the data):

$$\begin{aligned} f(y_1, \cdots, y_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - (\beta_0 + \beta_1 X_i))^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 X_i))^2\right) \end{aligned}$$

To get the maximum likelihood estimates for β_0 and β_1, we maximize the above function with respect to those two parameters. Recall that we can maximize the log likelihood instead:

$$\begin{aligned} \ell(\beta_0, \beta_1) &= log(f(y_1, \cdots, y_n)) \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 X_i))^2 \end{aligned}$$

Then we take the derivative with respect to the parameters β_0 and β_1 and set them equal to zero to find the maximum likelihood estimates:

$$\frac{\partial}{\partial \beta_0} \ell(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 X_i)) = 0$$
$$\frac{\partial}{\partial \beta_1} \ell(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 X_i)) X_i = 0$$

These are exactly the normal equations from least squares. Solving for estimates for β_0 and β_1 will yield exactly the estimates that were obtained via least squares, meaning that, in this context, the two estimation methods are equivalent.

3.6 Relationship to Bayesian Statistics

Approaching regression from a Bayesian perspective would involve assigning a prior distribution to the parameters in the linear regression model. These parameters are $\beta_0, \beta_1, \cdots, \beta_p$ and σ^2. The conjugate prior on $\beta_0, \cdots, \beta_p$ is a *multivariate normal* distribution, which we will not cover here. Instead, we can consider a simpler case of regression through the origin:

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

where $\epsilon \sim N(0, 1)$.

This is simply a regression model with no intercept. It forces our regression line to pass through the origin, hence the name.

In this case, the conjugate prior on the parameter β is just a normal distribution. To phrase this in a familiar way:

$$Y_i \sim N(\beta X_i, \sigma^2)$$
$$\beta \sim N(\mu, \tau)$$

where μ and τ are our prior parameters, which we assume are known.

Assuming further that σ^2 is known, we have by Bayes rule that the posterior distribution is given by:

$$\begin{aligned}
\nu(\beta|y_1,\cdots,y_n) &\propto f(y_1,\cdots,y_n|\beta)\nu(\beta) \\
&= \left(\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(y_i-\beta X_i)^2}\right)\frac{1}{2\pi\tau^2}e^{-\frac{1}{2\tau^2}(\beta-\mu)^2} \\
&\propto \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta X_i)^2-\frac{1}{2\tau^2}(\beta-\mu)^2) \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{\sum X_i^2}{\sigma^2}+\frac{1}{\tau^2}\right)\left(\beta-\frac{\frac{\sum X_i y_i}{\sigma^2}+\frac{\mu}{\tau^2}}{\frac{\sum x_i^2}{\sigma^2}+\frac{1}{\tau^2}}\right)\right)
\end{aligned}$$

and hence:

$$\beta|y_1,\cdots,y_n \sim N\left(\frac{\frac{\sum X_i y_i}{\sigma^2}+\frac{\mu}{\tau^2}}{\frac{\sum x_i^2}{\sigma^2}+\frac{1}{\tau^2}}, \frac{1}{\frac{\sum X_i^2}{\sigma^2}+\frac{1}{\tau^2}}\right)$$

The example above is just to illustrate how Bayesian parameter estimation can be applied in the context of linear regression. Because the $Y_1,\cdots,Y_n$ are a normal random sample, these estimation procedures remain nearly the same as they were before, but will now allow us to estimate the regression parameters.

3.7 Linear Regression in R

In R, the `lm()` function is for fitting linear models. `lm()` returns an object of class "`lm`". This object has a set of functions associated with it that allow the results to be displayed, used for prediction, etc. Models are specified symbolically. Here is a simple linear regression example:

```
x <- rnorm(100, 0, 1)
eps <- rnorm(100, 0, 1)
beta0 <- 5
beta1 <- 0.5
y <- beta0 + beta1*x + eps
```

```r
beta1hat <- cov(x,y)/var(x)
beta0hat <- mean(y) - beta1hat*mean(x)
beta1hat
```

```
## [1] 0.5209224
```

```r
beta0hat
```

```
## [1] 4.778763
```

```r
mod1 <- lm(y ~ x)
summary(mod1)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7436 -0.7272  0.0668  0.6818  3.1848
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.7788     0.1060  45.096  < 2e-16 ***
## x             0.5209     0.1180   4.415 2.61e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.059 on 98 degrees of freedom
## Multiple R-squared:  0.1659, Adjusted R-squared:  0.1574
## F-statistic: 19.49 on 1 and 98 DF,  p-value: 2.606e-05
```

```r
coef(mod1)
```

```
## (Intercept)           x
##   4.7787626   0.5209224
```

We can predict values by plugging in a new value into the fitted equation:

```r
# our new value of x for which we want to predict the value of y
xnew <- 0
yhat <- beta0hat + beta1hat * xnew
```

```
yhat
```

```
## [1] 4.778763
```

```
# alternatively:
lmbetahat <- coef(mod1)
yhat <- lmbetahat[[1]] + lmbetahat[[2]] * xnew
yhat
```

```
## [1] 4.778763
```

We can plot the line of best fit as follows:

```
fittedline <- function(x) beta0hat + beta1hat*x
plot(x, y)
curve(fittedline, add = TRUE)
```

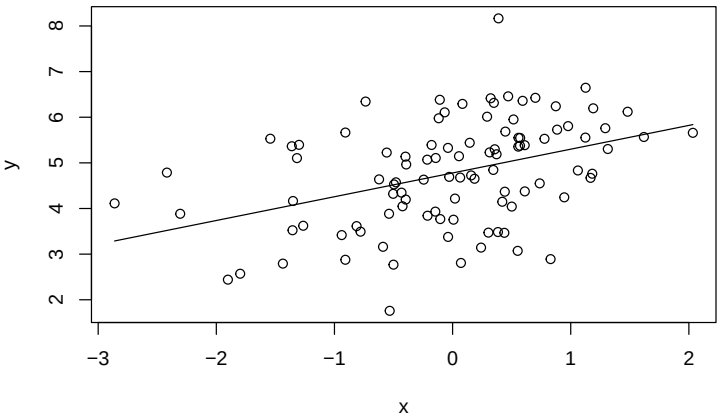


Figure 11: The fitted line.

And add our predicted value to the plot:

```
plot(x, y)
curve(fittedline, add = TRUE)
points(xnew, yhat, col = "red", pch = 19)
```

We can also use the `predict` function. It takes arguments of your linear model object (in this case `mod1`), and `newdata`. `newdata` must be a data

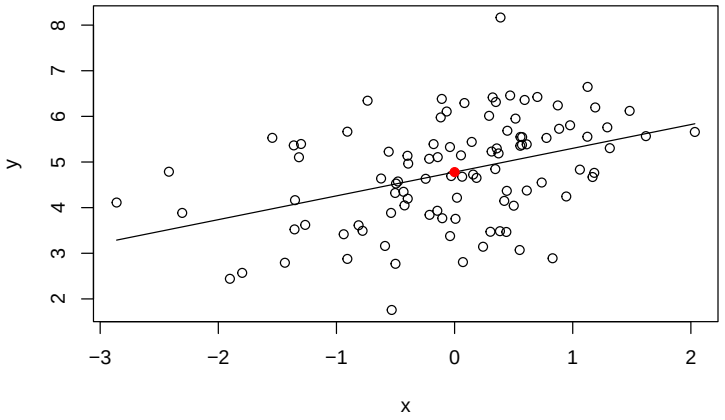


Figure 12: The fitted line.

frame with column names which are the same as the names of your predictor variables. In this case, we have a single predictor variable `x`, and the value we wish to obtain is the fitted equation evaluated at `xnew`, our new point.

```
predict(mod1, newdata = data.frame(x = xnew))
```

```
##        1
## 4.778763
```

3.8 Variance Stabalizing Transformations

Note that when we assume that the ϵ_i's be normally distributed (which leads to the Y_i's being normally distributed), we have necessarily that the variance doesn't depend on the mean. This just follows from the fact that the variance of a $N(\mu, \sigma^2)$ distribution is σ^2, and has nothing to do with whatever the value of μ may be.

In the case of $Y_1, \cdots, Y_n \sim Pois(\lambda)$, for example, the variance depends on the mean in that it is *equal* to the mean. Recall that for $Y \sim Pois(\lambda)$,

$$E(Y) = V(Y) = \lambda$$

The goal of a variance stabilizing transformation is to transform the data in such a way that the variance approximately does not depend on the value of the mean.

Consider a random sample $Y_1, \cdots, Y_n \sim Pois(\lambda)$, to which we apply some function g (the transformation). That is, we have

$$g(Y_1), \cdots, g(Y_n)$$

Note that, via a Taylor series approximation,

$$g(Y) \approx g(\lambda) + g'(\lambda)(Y - \lambda)$$

so we can therefore approximate the variance as:

$$V(g(Y)) \approx V(g(\lambda) + g'(\lambda)(Y - \lambda)) = g'(\lambda)^2 V(Y) = g'(\lambda)^2 \lambda$$

The choice of g that causes the above expression to *not* depend on λ is the choice such that $g'(\lambda)^2 = \frac{C}{\lambda}$ (for any constant C), because then:

$$V(g(Y)) \approx g'(\lambda)^2 \lambda = \frac{C\lambda}{\lambda} = C$$

This means that:

$$g'(\lambda)^2 = \frac{C}{\lambda} \implies g'(\lambda) = \frac{C^{1/2}}{\lambda^{1/2}} \implies g(\lambda) = 2C^{1/2}\lambda^{1/2}$$

Ignoring all constants, the variance stabilizing transformation is:

$$g(\lambda) = \lambda^{1/2}$$

Stabilizing the variance allows our data to become more normal in the sense that its variance will no longer depend as strongly on the mean. Based on the example above, we have that for data which we think is Poisson (i.e. "count" data), then we can apply the above transformation before performing linear regression in order to have the data better meet

the assumptions of the model (in particular, normality of the response Y).

Example:

Consider the following data frame `dashdata`, from which the first six rows are shown below:

```
head(dashdata)
```

```
##   lecture dashes_count
## 1       1           29
## 2       2           31
## 3       3           38
## 4       4            6
## 5       5           13
## 6       6           26
```

`lecture` is the lecture number (in chronological order), and `dashes_count` is the number of dashes that I made on the side of the pdf in order to let me scroll down further. Suppose we are interested in analyzing the relationship between lecture number and the number of dashes. That is, was there an increase in the number of dashes as the lectures progressed. We can fit a linear model as follows:

```
# Number of lectures
m <- 25

plot(lecture, dashes_count)
mod2 <- lm(dashes_count ~ lecture)
b <- coef(mod2)
curve(b[1] + b[2]*x, add = TRUE)
```

```
summary(mod2)
```

```
##
## Call:
## lm(formula = dashes_count ~ lecture)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.46 -10.16  -2.50   7.16  41.26
```

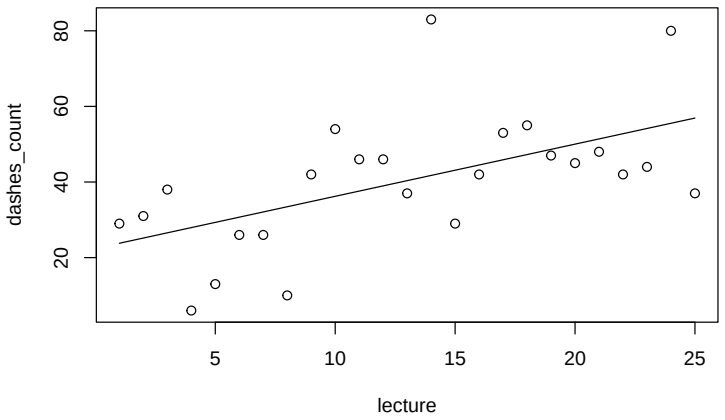


Figure 13: The fitted line.

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.4200     6.2707   3.575  0.00160 **
## lecture       1.3800     0.4218   3.272  0.00335 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.21 on 23 degrees of freedom
## Multiple R-squared:  0.3176, Adjusted R-squared:  0.2879
## F-statistic:  10.7 on 1 and 23 DF,  p-value: 0.003352
```

```
confint(mod2)
```

```
##                 2.5 %   97.5 %
## (Intercept) 9.4480100 35.39199
## lecture     0.5074095  2.25259
```

The summary of the model indicates that we should reject $H_0 : \beta_1 = 0$ based on the t-test (we'll talk about this later), indicating that there is a significant relationship between the lecture number and the number of dashes. The confidence interval for β_1 suggests that I make between 0.5 and 2.25 more dashes per lecture.

However, "dashes per lecture" is an example of "count" data. It is likely that the distribution of the dashes is closer to a Poisson than a normal, suggesting that the variance stabilizing transformation derived above would be useful in this situation. Below, a square root transformation is applied to the data and a regression model fit again:

```
sdashes <- sqrt(dashes_count)
plot(lecture, sdashes)
mod3 <- lm(sdashes ~ lecture)
b <- coef(mod3)
curve(b[1] + b[2]*x, add = TRUE)
```

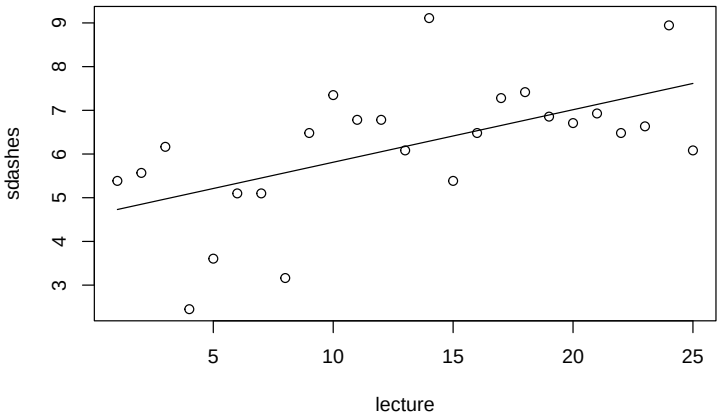


Figure 14: The fitted line.

```
summary(mod3)
```

```
##
## Call:
## lm(formula = sdashes ~ lecture)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64128 -0.74139 -0.05247  0.72993  2.81763
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   4.6100     0.5278   8.734 9.21e-09 ***
## lecture       0.1202     0.0355   3.386  0.00255 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.28 on 23 degrees of freedom
## Multiple R-squared:  0.3326, Adjusted R-squared:  0.3036
## F-statistic: 11.46 on 1 and 23 DF,  p-value: 0.002546
```

```
confint(mod3)
```

```
##                   2.5 %    97.5 %
## (Intercept) 3.51809144 5.7018068
## lecture     0.04675782 0.1936504
```

We still reject $H_0 : \beta_1 = 0$ based on the t-test, with a p-value of 0.00255.

3.9 Properties of the Least Squares Estimators

Recall the least squares estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

For convenience, define the following quantity:

$$SS_{XX} = \sum_{i=1}^n (X_i - \bar{X})^2$$

Rearranging, we can write the estimates above as follows:

$$\hat{\beta}_0 = \sum_{i=1}^n \left(\frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{SS_{XX}} \right) Y_i$$
$$\hat{\beta}_1 = \sum_{i=1}^n \frac{(X_i - \bar{X})}{SS_{XX}} Y_i$$

It is clear from these relationships that $\hat{\beta}_0$ and $\hat{\beta}_1$ are just weighted sums normal random variables. This implies that they are themselves normal.

3.10 Expected Value

To derive the expected value of $\hat{\beta}_1$:

$$\begin{aligned}
E(\hat{\beta}_1) &= E\left(\sum_{i=1}^{n} \frac{(X_i - \bar{X})}{SS_{XX}} Y_i\right) \\
&= \sum_{i=1}^{n} E\left(\frac{(X_i - \bar{X})}{SS_{XX}} Y_i\right) \\
&= \sum_{i=1}^{n} \frac{(X_i - \bar{X})}{SS_{XX}} E(Y_i) \\
&= \sum_{i=1}^{n} \frac{(X_i - \bar{X})}{SS_{XX}} (\beta_0 + \beta_1 X_i) \\
&= \frac{\beta_0}{SS_{XX}} \sum_{i=1}^{n} (X_i - \bar{X}) + \beta_1 \frac{\sum (X_i - \bar{X}) X_i}{SS_{XX}} \\
&= \frac{\beta_0}{SS_{XX}} \cdot 0 + \beta_1 \frac{SS_{XX}}{SS_{XX}} \\
&= \beta_1
\end{aligned}$$

To derive the expected value of $\hat{\beta}_0$:

$$E(\hat{\beta}_0) = E(\bar{Y} - \bar{X}\hat{\beta}_1) = E(\bar{Y}) - \bar{X}\beta_1 = \beta_0 + \beta_1 \bar{X} - \beta_1 \bar{X} = \beta_0$$

Hence both $\hat{\beta}_1$ and $\hat{\beta}_0$ are *unbiased* estimators of β_1 and β_0, respectively.

3.11 Variance

To derive the variance of $\hat{\beta}_1$, first recall the following:

Proposition: (Variance of a sum formula) Let Y_1 and Y_2 be random variables, and a and b be constants. Then,

$$V(aY_1 + bY_2) = a^2V(Y_1) + b^2V(Y_2) + 2ab\text{cov}(Y_1, Y_2)$$

For independent random variables, the covaraince is 0, which would mean that the above expression would become:

$$V(aY_1 + bY_2) = a^2V(Y_1) + b^2V(Y_2)$$

Since each Y_i for $i = 1, \cdots, n$ is independent in the case of linear regression, we can use the second formula to compute the variances.

$$\begin{aligned} V(\hat{\beta}_1) &= \sum_{i=1}^{n} V\left(\frac{(X_i - \bar{X})}{SS_{XX}} Y_i\right) \\ &= \sum_{i=1}^{n} \left(\frac{(X_i - \bar{X})^2}{SS_{XX}^2}\right) V(Y_i) \\ &= \frac{\sigma^2}{SS_{XX}} \end{aligned}$$

To derive the variance of $\hat{\beta}_0$:

$$\begin{aligned} V(\hat{\beta}_0) &= \sum_{i=1}^{n} V\left(\left(\frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{SS_{XX}}\right) Y_i\right) \\ &= \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{SS_{XX}}\right)^2 V(Y_i) \\ &= \sum_{i=1}^{n} \left(\frac{1}{n^2} - \frac{2(X_i - \bar{X})\bar{X}}{nSS_{XX}} + \frac{(X_i - \bar{X})^2\bar{X}^2}{SS_{XX}^2}\right) \sigma^2 \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{SS_{XX}}\right) \\ &= \frac{\sum X_i^2}{nSS_{XX}} \sigma^2 \end{aligned}$$

3.12 Final Distributional Results

Putting these results together, we have the following distributional results. Since $\hat{\beta}_0$ and $\hat{\beta}_1$ are statistics (i.e. functions of the data), we call the following distributions their *sampling distributions*:

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sum X_i^2}{nSS_{XX}}\sigma^2\right)$$

and

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{SS_{XX}}\right)$$

Using our typical notation, we have that the standard errors for the above estimates are given by:

$$SE(\hat{\beta}_0) = \sqrt{\frac{\sum X_i^2}{nSS_{XX}}\sigma^2}$$
$$SE(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{SS_{XX}}}$$

3.13 Inference About the Betas

Recall the distributional results from the previous lecture:

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sum X_i^2}{nSS_{XX}}\sigma^2\right)$$
$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{SS_{XX}}\right)$$

This yielded the standard errors:

$$SE(\hat{\beta}_0) = \sqrt{\frac{\sum X_i^2}{nSS_{XX}}\sigma^2}$$

$$SE(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{SS_{XX}}}$$

3.14 Error Variance

In the assumptions of linear regression, the error term ϵ was assumed to be distributed $N(0, \sigma^2)$. The maximum likelihood estimate for the parameter σ^2, the variance of the error term, is given by:

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

It is the case that:

$$E(\hat{\sigma}^2_{MLE}) = \frac{n-2}{n}\sigma^2$$

which makes the MLE a biased estimate for σ^2. The $n-2$ comes from the fact that we estimated two quantities in order to calculate $\hat{\sigma}^2_{MLE}$: β_0 and β_1.

In order to create an *unbiased* estimate from the MLE, we simply multiply by the relevant constant. In this case, we define:

$$\hat{\sigma}^2 = MSE = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

MSE is unbiased for the error variance σ^2. MSE stands for "mean squared error". We will discuss why this is further when we get to discussing ANOVA.

Similarly to S^2, it is the case that:

$$\frac{(n-2)MSE}{\sigma^2} \sim \chi^2(n-2)$$

3.15 Estimated Standard Errors

Theorem: Suppose that $Z \sim N(0,1)$ and $W \sim \chi^2(\nu)$ be independent random variables. Then,

$$\frac{Z}{\sqrt{W/\nu}} \sim t(\nu)$$

We will use the above result in the context of the regression coefficients.

With our estimate $\hat{\sigma}^2 = MSE$ for σ^2, we can estimate the standard errors of our regression coefficients as follows:

$$\hat{SE}(\hat{\beta}_0) = \sqrt{\frac{\sum X_i^2}{nSS_{XX}}\hat{\sigma}^2}$$
$$\hat{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{SS_{XX}}}$$

Recall from above that:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{SS_{XX}}\right)$$

Hence, by shifting and scaling $\hat{\beta}_1$, define Z as follows:

$$Z := \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{SS_{XX}}}} = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim N(0,1)$$

Define

$$W = \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$

such that

$$\frac{W}{n-2} = \frac{\hat{\sigma^2}}{\sigma^2}$$

Then, by the above theorem,

$$\frac{Z}{\sqrt{W/(n-2)}} = \frac{\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{SS_{XX}}}}}{\sqrt{\hat{\sigma}^2/\sigma^2}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{SS_{XX}}}} \sim t(n-2)$$

Following a similar process with β_0, we are left with the following two distributional results involving the *estimated* standard errors:

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{SE}(\hat{\beta}_0)} \sim t(n-2)$$
$$\frac{\hat{\beta}_1 - \beta_1}{\hat{SE}(\hat{\beta}_1)} \sim t(n-2)$$

3.16 Testing

We proceed using β_1, but the process is equivalent with β_0. Consider testing the following hypothesis:

$$H_0 : \beta_1 = b$$
$$H_1 : \beta_1 \neq b$$

where b is some constant. *Under the assumption that H_0 is true*, we have that

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{SE}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - b}{\hat{SE}(\hat{\beta}_1)} \sim t(n-2)$$

Therefore we calculate the test statistic:

$$T = \frac{\hat{\beta}_1 - b}{\hat{SE}(\hat{\beta}_1)} \sim_{H_0} t(n-2)$$

and have the rejection region:

$$\text{Reject } H_0 \text{ if } T > t_{n-2,\alpha/2} \text{ or } T < -t_{n-2,\alpha/2}$$

In particular, we often consider the following hypothesis:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

This is because we can interpret this test as testing whether or not the covariate X has an influence on the value of Y. This is because, supposing that $\beta_1 = 0$, the regression model reduces to:

$$Y = \beta_0 + \epsilon$$

which is to say that the value of Y is completely independent of the value of X.

3.17 Confidence Intervals

It follows from the fact that:

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{SE}(\hat{\beta}_1)} \sim t(n-2)$$

that we have that:

$$P\left(-t_{n-2,\alpha/2} < \frac{\hat{\beta}_1 - \beta_1}{\hat{SE}(\hat{\beta}_1)} < t_{n-2,\alpha/2}\right) = 1 - \alpha$$

and hence,

$$\hat{\beta}_1 - t_{n-2,\alpha/2}\hat{SE}(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + t_{n-2,\alpha/2}\hat{SE}(\hat{\beta}_1)$$

is a $100(1-\alpha)\%$ confidence interval for β_1.

Via the same process, an analogous result also applies for β_0:

$$\hat{\beta}_0 - t_{n-2,\alpha/2}\hat{SE}(\hat{\beta}_0) < \beta_0 < \hat{\beta}_0 + t_{n-2,\alpha/2}\hat{SE}(\hat{\beta}_0)$$

3.18 Prediction

Recall that given estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ in simple linear regression, we can make predictions for the value of Y at a covariate value of X_0 (call this prediction $\hat{Y}$) simply by plugging the value of X into our fitted regression equation:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

We can think of the mean squared error as the difference between the predicted value of Y at each observed covariate value X_i that is in our data, and the actual value of Y_i that was observed, i.e.

$$\begin{aligned} MSE &= \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)) \\ &= \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i) \end{aligned}$$

Definition: The quantity $Y_i - \hat{Y}_i$ is called the ith *residual*, denoted

$$e_i = Y_i - \hat{Y}_i$$

We are clearly not restricted to making predictions $\hat{Y}_i$ which correspond to X_i values present in the data. We can choose any value of X, which we will usually indicate using X_0 if it is not one of the observed X_i values.

3.19 Mean Response

The simple linear regression model states that:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$. This means that:

$$E(Y) = \beta_0 + \beta_1 X$$

Suppose that we know that the covariate X assumes a value of X_0. Then,

$$E(Y|X = X_0) = \beta_0 + \beta_1 X_0$$

We will use the quantity

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

as an estimate for $E(Y|X = X_0)$. We will denote:

$$\mu_{Y|X_0} := E(Y|X = X_0) = \beta_0 + \beta_1 X_0$$

$\hat{Y}$ is unbiased for $\mu_{Y|X_0}$ (this follows immediately from the fact that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased):

$$E(\hat{Y}) = E(\hat{\beta}_0 + \hat{\beta}_1 X_0) = \beta_0 + \beta_1 X_0 = \mu_{Y|X_0}$$

The variance of the estimator $\hat{Y}$ is given by:

$$\begin{aligned} V(\hat{Y}) &= V(\hat{\beta}_0 + \hat{\beta}_1 X_0) \\ &= V(\hat{\beta}_0) + X_0^2 V(\hat{\beta}_1) + 2X_0 \text{cov}(\hat{\beta}_0, \hat{\beta}_1) \end{aligned}$$

We have already derived $V(\hat{\beta}_0)$ and $V(\hat{\beta}_1)$. To get $\text{cov}(\hat{\beta}_0, \hat{\beta}_1)$, note that

$$V(\hat{\beta}_0 + \bar{X}\hat{\beta}_1) = V(\bar{Y})$$

$$V(\hat{\beta}_0) + \bar{X}^2 V(\hat{\beta}_1) + 2\bar{X}\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{\sigma^2}{n}$$

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{1}{2\bar{X}}\left(\frac{\sigma^2}{n} - V(\hat{\beta}_0) - \bar{X}^2 V(\hat{\beta}_1)\right)$$

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{X}\sigma^2}{SS_{XX}}$$

Therefore,

$$V(\hat{Y}) = \sigma^2\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_{XX}}\right)$$

$\hat{Y}$ is the sum of two normal random variables ($\hat{\beta}_0$ and $\hat{\beta}_1$) and is therefore itself normally distributed:

$$\hat{Y} \sim N\left(\mu_{Y|X_0}, \sigma^2\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_{XX}}\right)\right)$$

and via an analogous process to what was done with the regression coefficients:

$$T = \frac{\hat{Y} - \mu_{Y|X_0}}{\sqrt{\hat{\sigma^2}\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_{XX}}\right)}} \sim t(n-2)$$

where $\hat{\sigma}^2 = MSE$.

As usual, we denote:

$$\hat{SE}(\hat{Y}) = \sqrt{\hat{\sigma^2}\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_{XX}}\right)}$$

This leads naturally to tests and confidence intervals. In particular, a $100(1-\alpha)\%$ confidence interval for $\mu_{Y|X_0}$ is given by:

$$\hat{Y} - t_{n-2,\alpha/2}\hat{SE}(\hat{Y}) < \mu_{Y|X_0} < \hat{Y} + t_{n-2,\alpha/2}\hat{SE}(\hat{Y})$$

3.20 Single Response

Consider the random variable $Y_0 := Y|X = X_0$. That is, we fix $X = X_0$, and the regression equation becomes:

$$Y_0 = \beta_0 + \beta_1 X_0 + \epsilon$$

We wish to construct a confidence interval for observed values of Y_0. Clearly,

$$Y_0 \sim N(\beta_0 + \beta_1 X_0, \sigma^2)$$

This doesn't help us, because we don't know β_0, β_1, or σ^2. In order to construct a confidence interval out of *estimates* of these values, we can do the following trick.

Let $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$ and define the random variable $\hat{Y}_0 - Y_0$. Note that it is the sum of normal random variables, so its normal. Furthermore,

$$E(\hat{Y}_0 - Y_0) = E(\hat{\beta}_0 + \hat{\beta}_1 X_0 - (\beta_0 + \beta_1 X_0 + \epsilon)) = 0$$

$\hat{Y}_0$ isn't an estimator for Y_0, because Y_0 is itself a random variable. Rather, you could say $\hat{Y}_0 - Y_0$ is an unbiased estimator of 0.

We also have that

$$\begin{aligned} V(\hat{Y}_0 - Y_0) &= V(\hat{\beta}_0 + \hat{\beta}_1 X_0 - (\beta_0 + \beta_1 X_0 + \epsilon)) \\ &= V(\hat{\beta}_0 + \hat{\beta}_1 X_0) + V(\epsilon) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_{XX}} \right) + \sigma^2 \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_{XX}} \right) \end{aligned}$$

so defining

$$\hat{SE}(\hat{Y}_0 - Y_0) = \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_{XX}}\right)}$$

we have that

$$T := \frac{\hat{Y}_0 - Y_0}{\hat{SE}(\hat{Y}_0 - Y_0)} \sim t(n-2)$$

and hence

$$\hat{Y}_0 - t_{n-2,\alpha/2}\hat{SE}(\hat{Y} - Y_0) < Y_0 < \hat{Y}_0 + t_{n-2,\alpha/2}\hat{SE}(\hat{Y} - Y_0)$$

is a $100(1-\alpha)\%$ confidence interval for Y_0, generally referred to as a *prediction interval*, since Y_0 is a singular predicted value (as opposed to a confidence interval for the *mean* response $\mu_{Y|X_0}$).

Note: The prediction interval accounts for both the variability in the mean and the variability in the observation, and is therefore wider than the confidence interval. This is easy to see since we add 1 in the expression inside of the square root in the standard error.

Example: In this example we consider simulated data in the form of vectors `x` and `y` with $n = 90$.

Suppose we wish to create a confidence interval for the value:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 5$$

```
X0 <- 5
Xbar <- mean(x)
SSxx <- sum((x - Xbar)^2)
MSE <- 1/(n-2) * sum((y - (beta0hat + beta1hat*x))^2)

Yhat <- beta0hat + beta1hat * X0
SEYhat <- sqrt(MSE * (1/n + ((X0 - Xbar)^2)/SSxx))

Yhat + qt(c(0.025, 0.975), n-2) * SEYhat
```

```
## [1]  9.705778 10.808968
```

If we want a prediction interval for a new observation of Y when $X = 5$, we would do the following:

```
Yhat0 <- beta0hat + beta1hat * X0
SEYhatY0 <- sqrt(MSE * (1 + 1/n + ((X0 - Xbar)^2)/SSxx))

Yhat0 + qt(c(0.025, 0.975), n-2) * SEYhatY0
```

```
## [1]  6.692253 13.822493
```

3.21 ANOVA

Recall that

$$MSE = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Without the $1/(n-2)$ factor in front, we have:

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

where SSE just stands for "error sum of squares".

The *total sum of squares* SST is given by:

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

The *regression sum of squares* is given by:

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

Theorem:

$$SST = SSR + SSE$$

Proof:

$$\begin{aligned} SST &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\ &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\ &= \sum_{i=1}^{n}(Y - \hat{Y}_i)^2 + (\hat{Y}_i - \bar{Y})^2 + 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \\ &= SSE + SSR + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \end{aligned}$$

To show that the last term is equal to 0, first note that:

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = \sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)) = 0$$

because the quantity on the right is necessarily equal to 0 via the normal equations (the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ were chosen precisely such that the above equality is true).

Likewise, also as a result of the normal equations,

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)X_i = \sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))X_i = 0$$

Therefore,

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y}) \\
&= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\bar{Y} - \hat{\beta}_1\bar{X} + \hat{\beta}_1 X_i - \bar{Y}) \\
&= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)\hat{\beta}_1(X_i - \bar{X}) \\
&= -\hat{\beta}_1\bar{X}\sum_{i=1}^{n}(Y_i - \hat{Y}_i) + \hat{\beta}_1\sum_{i=1}^{n}(Y_i - \hat{Y}_i)X_i \\
&= 0 + 0 = 0
\end{aligned}$$

which implies

$$SST = SSE + SSR$$

■

The significance of these quantities is as follows. Recall that:

$$\frac{(n-2)MSE}{\sigma^2} \sim \chi^2(n-2)$$

which can be rephrased as:

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$$

Result: For SSR as defined above, we have that

$$\frac{SSR}{\sigma^2} \sim \chi^2_{NC}\left(1, \frac{\beta_1^2\sum_{i=1}^{n}(X_i - \bar{X})^2}{2\sigma^2}\right)$$

where χ^2_{NC} is a noncentral chi-squared distribution.

The noncentrality parameter, which is the second parameter in the above expression, is an additional parameter of the non central chi-squared

distribution. We will not discuss this distribution here, because we care only about the following result:

Consider testing the following hypothesis:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Then under H_0, the noncentrality parameter $= 0$, and we have that

$$\frac{SSR}{\sigma^2} \sim \chi^2\,(1)$$

i.e. our usual χ^2 distribution with 1 degree of freedom.

Recall the following result from lecture 13:

Theorem: Let $S_1 \sim \chi^2(d_1)$ and $S_2 \sim \chi^2(d_2)$ be *independent* chi-squared random variables. Then,

$$F = \frac{S_1/d_1}{S_2/d_2} \sim F(d_1, d_2)$$

Therefore, we define

$$F = \frac{SSR/\sigma^2/1}{SSE/\sigma^2/(n-2)} = \frac{SSR}{MSE} = \frac{MSR}{MSE} \sim_{H_0} F(1, n-2)$$

and we can perform tests using the quantiles of the $F(1, n-2)$ distribution.

Note that $MSR = SSR/1$ seems pointless to define here, but this equality will not be true when we generalize this to multiple linear regression.

Note that this is a one-sided test. Our rejection region is:

$$\text{Reject if } F > F_{\alpha,1,n-2}$$

The reason for the one-sided test is that only large values of F suggest rejecting H_0 in favor of H_1 in this circumstance. If MSR is large compared

to MSE, F is large and it suggests the existence of a noncentrality parameter, which would increase the value of MSR.

3.22 Coefficient of Determination

Definition: The coefficient of determination r^2 is given by:

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

r^2 represents the proportion of the variation in Y that is explained by the linear model. This value is given in the output of the `R` function `lm`.

3.23 Correlation

Definition: The *sample correlation* between X and Y is given by:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

The square of the sample correlation, r^2, is precisely the coefficient of determination given above:

$$\begin{aligned} SSR &= \sum(\hat{Y}_i - \bar{Y})^2 \\ &= \sum(\hat{\beta}_0 + \hat{\beta}_1 X_i - \hat{\beta}_0 - \hat{\beta}_1 \bar{X})^2 \\ &= \sum \hat{\beta}_1^2 (X_i - \bar{X})^2 \\ &= \hat{\beta}_1^2 SS_{XX} \\ &= \left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{SS_{XX}} \right)^2 SS_{XX} \\ &= \frac{(\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}))^2}{SS_{XX}} \end{aligned}$$

and therefore,

$$R^2 = \frac{SSR}{SST} = \frac{(\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}))^2}{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2} = r^2$$

Obviously, the notation r^2 for the coefficient of determination comes from the fact that it is equal to the square of the sample correlation.

3.24 Multiple Linear Regression

Recall the linear regression model introduced in lecture 23:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i \qquad i = 1, \cdots, n$$

where $\epsilon_i \sim N(0, \sigma^2)$.

Again, Y is a numeric response variable, $X_1, ..., X_p$ are predictor variables (either numeric or categorical). We observe n of each, which we indicate using the index i.

The method for obtaining estimates for all of the betas (there are $p+1$ of them) is exactly the same as with simple linear regression. That is, we use least squares (which is equivalent to maximum likelihood). In this case, that would mean that we want to minimize the quantity:

$$Q := \sum_{i=1}^{n}(Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}))^2$$

So we would solve the system of $p+1$ equations given by:

$$\frac{\partial Q}{\partial \beta_0} = 0$$
$$\vdots$$
$$\frac{\partial Q}{\partial \beta_p} = 0$$

For an arbitrary index $k \geq 1$,

$$\frac{\partial Q}{\partial \beta_k} = \sum_{i=1}^{n} -2(Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}))X_{ik}$$

Without the use of matrices, it is hard to express the estimates $\hat{\beta}_0 \cdots \hat{\beta}_p$ in a nice way.

3.25 Multiple Linear Regression Using Matrices

If you are unfamiliar with matrix algebra, look at the in-class notes for some review that I did.

Define the following quantities:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & & & \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The matrix X is known as the *design matrix*.

Our linear regression model can now be written as:

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

Least squares is as follows:

$$\begin{aligned} Q &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\ &= (Y - X\beta)^T (Y - X\beta) \\ &= Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta X^T X\beta \end{aligned}$$

However, $Y^T X\beta = \beta^T X^T Y$ in this case, which leaves

$$Q = Y^T Y - 2\beta^T X^T Y + \beta X^T X\beta$$

By rules for vector derivatives, we have that:

$$\begin{aligned} \frac{dQ}{d\beta} &= -2X^T Y + 2X^T X\beta = 0 \\ \hat{\beta} &= (X^T X)^{-1} X^T Y \end{aligned}$$

That is, by performing the matrix product above, we obtain the $p + 1$ dimensional vector:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$$

with entries equal to the regression coefficients that would be obtained by solving the system of equations presented earlier. Therefore, obtaining least squares estimates for the regression coefficients in the multiple linear regression case is as simple as constructing the design matrix X and then performing the above matrix product.

3.26 Simple Linear Regression Case

In the simple linear regression case, we can still use the matrix form. In this case,

$$X^T X = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$$

and

$$X^T Y = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}$$

Still,

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

Working through the algebra will yield the same least squares estimates that we had earlier.

3.27 Distributional Results, Testing, Confidence Intervals

The vector ϵ follows a multivariate normal distribution:

$$\epsilon \sim N(0, \sigma^2 I)$$

I is the identity matrix. Its entries are zero off of the diagonal, so the ϵ_i's are independent and marginally $N(0, \sigma^2)$, as we expect.

The vector Y is also multivariate normal:

$$Y \sim N(X\beta, \sigma^2 I)$$

Each $Y_i \sim N(\beta_0 + \beta_1 X_{i1} + \cdots \beta_p X_{ip}, \sigma^2)$ are independent.

The distribution of the vector $\hat{\beta}$ is also multivariate normal, where

$$\hat{\beta} \sim N(\beta, \sigma^2(X^T X)^{-1})$$

where the mean and variance follow from:

$$\begin{aligned} E(\hat{\beta}) &= E((X^T X)^{-1} X^T Y) \\ &= (X^T X)^{-1} X^T E(Y) \\ &= (X^T X)^{-1} X^T X \beta \\ &= \beta \end{aligned}$$

and

$$\begin{aligned} V(\hat{\beta}) &= V((X^T X)^{-1} X^T Y) \\ &= (X^T X)^{-1} X^T V(Y) X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

($(X^T X)^{-1}$ is symmetric since $X^T X$ is).

It is now the case that

$$\begin{aligned} \hat{\sigma}^2 = MSE &= \frac{1}{n-(p+1)} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\ &= \frac{1}{n-(p+1)} \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}))^2 \end{aligned}$$

where we estimate $p+1$ regression coefficients, so the constant that makes the above quantity unbiased changes accordingly.

For each $k = 0, \cdots, p$,

$$\hat{SE}(\hat{\beta}_k) = \sqrt{\hat{\sigma}^2 [(X^T X)^{-1}]_{k+1,k+1}}$$

where $[(X^TX)^{-1}]_{k+1,k+1}$ is the $(k+1, k+1)$ entry of the matrix $(X^TX)^{-1}$.

Also, for each $k = 0, \cdots, p$,

$$T = \frac{\hat{\beta}_k - \beta_k}{\hat{SE}(\hat{\beta}_k)} \sim t(n - (p+1))$$

and a $100(1-\alpha)\%$ confidence interval for β_k is given by:

$$\hat{\beta}_k - t_{n-(p+1),\alpha/2}\hat{SE}(\hat{\beta}_k) < \beta_k < \hat{\beta}_k + t_{n-(p+1),\alpha/2}\hat{SE}(\hat{\beta}_k)$$

3.28 Penalized Regression

Least squares is the process of minimizing the quantity Q in order to estimate β, where

$$Q = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = (Y - X\beta)^T(Y - X\beta)$$

Penalized Regression generally refers to minimizing some other quantity. Usually, there is some penalty that is added to Q which forces our regression coefficients to have favorable properties. For example, we may want to minimize:

$$Q + \lambda\beta^T S\beta$$

where S is a matrix, and the term $\beta^T S\beta$ enforces some favorable property upon the β's (we will see a specific example below). λ is a positive constant, where large values of lambda enforce the penalty more than small values do. $\lambda = 0$ returns us to the least squares case.

We can derive the form of the estimator $\hat{\beta}$ in this situation as follows. Recall that:

$$Q = Y^TY - 2\beta^TX^TY + \beta X^TX\beta$$

so, by the matrix differentiation rules,

$$\frac{d}{d\beta}(Q + \lambda\beta^T S\beta) = -2X^T Y + 2X^T X\beta + 2\lambda S\beta = 0$$

$$X^T X\beta + \lambda S\beta = X^T Y$$

$$(X^T X + \lambda S)\beta = X^T Y$$

$$\hat{\beta} = (X^T X + \lambda S)^{-1} X^T Y$$

3.29 Sparsity

In high-dimensional situations (e.g machine learning) where the dimension of β is large, penalties can be used to enforce *sparsity*. A vector is said to be sparse if many of its entries are 0 (were not using an exact definition here).

A common penalty to enforce sparseness is called the Least Absolute Selection and Shrinkage Operator (LASSO). In this case, to obtain estimates for β, we must minimize

$$\frac{1}{2}Q + \lambda \sum_{i=1}^{d} |\beta_i|$$

where d is the dimension of β. Large values of λ enforce a stricter penalty (β will be more approximately sparse for larger values of λ).

There is no closed form solution to the above equation. We can't differentiate it because the absolute value is not differentiable everywhere. It can, however, be solved very quickly. The above function is convex, but there are fast algorithms which can be used (e.g. subgradient descent).

The end result is still an estimate $\hat{\beta}$, but which is sparse, which is necessary, or just more favorable, in certain contexts.

3.30 Tests Concerning Multiple Betas

There are two kinds of F tests which we can perform in order to test if multiple regression coefficients are zero simultaneously. The first test we consider is for testing every β besides β_0.

If we have p predictors, such that we have regression coefficients $\beta_1, \cdots, \beta_p$, we can test the following hypothesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$
$$H_1 : \text{At least one } \beta_i \neq 0 \quad (\text{where } i \neq 0)$$

Note that the intercept β_0 is *not* included in either hypothesis.

Recall that in multiple linear regression,

$$\hat{\sigma}^2 = MSE = \frac{1}{n-(p+1)} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

It remains the fact that:

$$\frac{(n-(p+1))MSE}{\sigma^2} = \frac{SSE}{\sigma^2} \sim \chi^2(n-(p+1))$$

In the multiple regression case,

$$MSR := \frac{SSR}{p}$$

where p is the number of covariates.

Similarly as to the simple linear regression case, under H_0 as stated above:

$$F := \frac{MSR}{MSE} \sim F(p, n-(p+1))$$

The F test above is the test for which the F statistic is given in the summary of the output of the `lm` function in `R`.

The second kind of F test for multiple betas is for testing if some subset of the βs is equal to zero.

As an example, consider the following linear model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \qquad i = 1, \cdots, n$$

and suppose that we wish to test the following hypothesis:

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_1 : \text{At least one of } \beta_1 \text{ or } \beta_2 \neq 0$$

We can do this by fitting two models: a full model and a reduced model.

The full model contains all of the covariates. In this case, "fitting" the full model involves assuming the regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \qquad i = 1, \cdots, n$$

and then obtain an estimate of the form:

$$\hat{\beta}_F = \begin{bmatrix} \hat{\beta}_{0F} \\ \hat{\beta}_{1F} \\ \hat{\beta}_{2F} \\ \hat{\beta}_{3F} \end{bmatrix}$$

The reduced model contains only the covariates which are *not* included in the null. In this example, in order to fit the reduced model, we assume:

$$Y_i = \beta_0 + \beta_3 X_{i3} + \epsilon_i \qquad i = 1, \cdots, n$$

and then obtain an estimate of the form:

$$\hat{\beta}_R = \begin{bmatrix} \hat{\beta}_{0R} \\ \hat{\beta}_{3R} \end{bmatrix}$$

Using these predicted values, we can calculate $SSE(R)$, the sum of squares error for the reduced model, and $SSE(F)$, the sum of squares error for the full model. Then, under the assumption that H_0 as stated above is *true*,

$$F := \frac{\frac{SSE(R)-SSE(F)}{p-g}}{\frac{SSE(F)}{n-(p+1)}} \sim F(p-g, n-(p+1))$$

p is the number of predictors in the full model (in this example, 3), and g is the number of predictors in the reduced model (in this example, 1).

3.31 Categorical Predictors

We will now turn our attention to categorical response variables. As an illustrative example, consider a linear model with one numeric predictor X_1, and a binary categorical variable X_2:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \qquad i = 1, \cdots, n$$

X_2 is binary, meaning that for each i, X_{i2} is equal to either 0 or 1. In the case where $X_{i2} = 0$,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i \qquad i = 1, \cdots, n$$

In the case where $X_{i2} = 1$,

$$Y_i = (\beta_0 + \beta_2) + \beta_1 X_{i1} + \epsilon_i \qquad i = 1, \cdots, n$$

With a numeric predictor, the value of β associated with that predictor is typically interpreted to be the change in Y associated with a one unit increase in X. For a categorical predictor like X_2 above, we can interpret β_2 as the change in the intercept of the regression line as we go from the category indicated by $\beta_2 = 0$ versus the category indicated by $\beta_2 = 1$.

Fitting the regression model when using categorical predictors is exactly the same. We create the design matrix in the same way (the column associated with X_2 will contain only 0s and 1s), and estimates are obtained via:

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

3.32 Some Regression Examples

Foot Callus:

The following data set [3] contains information on the shoe use, foot sensitivity, and foot callus thickness of 76 subjects. Here is a random sample of 6 of the rows of selected columns of the data frame `kenya1`:

```
slice_sample(kenya1[-c(1, 3, 8, 9)], n = 6)
```

```
## # A tibble: 6 x 5
##   sex    height_cm weight_kg sensitivityln shoe_use
##   <fct>      <dbl>     <dbl>          <dbl> <fct>
## 1 male         170        51           0.85 Rarely
## 2 female       163        47           2.35 Rarely
## 3 male         163        69           2.01 Everyday
## 4 male         162        44           1.71 Never
## 5 male         169        64           2.67 Rarely
## 6 male         165        72           1.97 Everyday
```

The columns represent the following:

`subject` identifies each subject, with "ke" signifying that the subject was from Kenya (every subject in this data frame is from Kenya), and a unique identifying number.

`sex`, `age_yrs`, `height_cm`, and `weight_kg` are self explanatory. `sex` is categorical, whereas the rest are numeric.

`ultrasound_heel_ln` is the log of the callus thickness (in cm) at the heel.

`durometer_heel_ln` is the log of the skin hardness (Shore units, Sh) at the heel.

`sensitivityln` is a measure of the threshold at which the subject was able to sense a vibration in the metatarsal head (the "palm" of the foot).

`shoe_use` is a categorical variable with five self-reported levels of shoe use for each subject: "Never", "Rarely", "TwoDays", "ThreeDays", and "Everyday"

The variables were log-transformed to achieve a higher degree of normality.

Below, a scatter plot of log callus thickness (`ultrasound_heel_ln`) versus log skin hardness (`durometer_heel_ln`) is made, and simple linear re-

gression is fit, with a best fit line overlayed on the scatter plot (recreating figure 1d in the paper):

```
x <- kenya1$ultrasound_heel_ln
y <- kenya1$durometer_heel_ln

# Calculate the regression coefficients
beta1hat <- cov(x,y)/var(x)
beta1hat
```

```
## [1] 0.7589785
```

```
beta0hat <- mean(y) - beta1hat*mean(x)
beta0hat
```

```
## [1] 5.252366
```

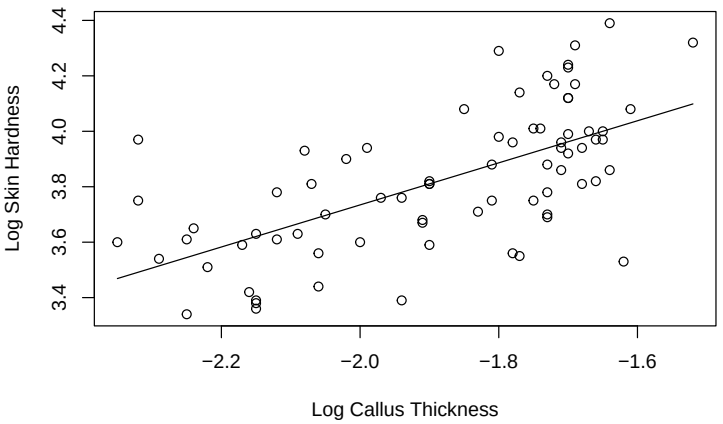


Figure 15: Log Callus Thickness vs Log Skin Hardness

Here we print a summary of the linear model using the `lm` and `summary` functions:

```
mod1 <- lm(y ~ x)
summary(mod1)
```

```
##
## Call:
## lm(formula = y ~ x)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49282 -0.13772 -0.00018  0.13259  0.47846
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.2524     0.2001  26.252  < 2e-16 ***
## x             0.7590     0.1056   7.189 4.37e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1967 on 74 degrees of freedom
## Multiple R-squared:  0.4112, Adjusted R-squared:  0.4032
## F-statistic: 51.68 on 1 and 74 DF,  p-value: 4.367e-10
```

Based on a p-value of $3.65 \cdot 10^{-10}$, we reject H_0 when testing at any reasonable level and conclude that callus thickness has an effect on skin hardness.

One of the experimental questions in the paper was to conclude whether or not people who rarely wore shoes, and hence had developed thicker calluses and more hardness on their feet, had lost sensitivity in the foot as a result. As an illustrative example (not what was done in the paper), consider the following model:

```
kenyaMod <- lm(sensitivityln ~ shoe_use + sex +
                 age_yrs + height_cm + weight_kg, data = kenya1)
summary(kenyaMod)
```

```
##
## Call:
## lm(formula = sensitivityln ~ shoe_use + sex + age_yrs + height_cm +
##     weight_kg, data = kenya1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.60380 -0.32554  0.03216  0.30978  2.05115
##
## Coefficients:
```

```
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -4.330678   2.032420  -2.131  0.03678 *
## shoe_useTwoDays     1.480342   0.687241   2.154  0.03484 *
## shoe_useThreeDays   0.100251   0.406593   0.247  0.80600
## shoe_useNever       0.373769   0.308678   1.211  0.23020
## shoe_useRarely      0.518006   0.204313   2.535  0.01357 *
## sexmale            -0.290635   0.183850  -1.581  0.11863
## age_yrs             0.018405   0.006042   3.046  0.00331 **
## height_cm           0.034372   0.013221   2.600  0.01147 *
## weight_kg          -0.002241   0.007971  -0.281  0.77949
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6326 on 67 degrees of freedom
## Multiple R-squared:  0.3013, Adjusted R-squared:  0.2179
## F-statistic: 3.612 on 8 and 67 DF,  p-value: 0.001522
```

In this model, our response is `sensitivityln`. `age_yrs`, `height_cm`, and `weight_kg` are numeric predictors and hence have one regression coefficient associated with each of them. Based on the t-tests, age and height are significant predictors for foot sensitivity.

`sex` is a categorical variable with two levels. As a result, it is sufficient to define a single variable:

$$X_5 = 0 \text{ if sex is female}, 1 \text{ if sex is male}$$

The estimate for the regression coefficient associated with this variable is $\hat{\beta}_5 = -0.29$. Since `female` is the baseline level, we can interpret this to mean that going from `female` to `male` causes a decrease in the average log foot sensitivity by 0.29 units (units of foot sensitivity in the log scale). I can also create a confidence interval for this value:

```
confint(kenyaMod)["sexmale", ]
```

```
##       2.5 %      97.5 %
## -0.65760052  0.07633044
```

The fact that the confidence interval contains 0 (and equivalently, the fact that the t-test fails to reject $H_0 : \beta_5 = 0$ at level 0.05) suggests, however, that this decrease in foot sensitivity is not significant (the decrease cannot

necessarily be said to have been "caused" by the value of the categorical variable `sex`).

`shoe_use` is a categorical variable with five categories. As we saw last lecture, this necessitates the use of four "dummy variables":

$$
\begin{aligned}
X_1 &= 1 \text{ if shoe use is two days}, 0 \text{ otherwise} \\
X_2 &= 1 \text{ if shoe use is three days}, 0 \text{ otherwise} \\
X_3 &= 1 \text{ if shoe use is never}, 0 \text{ otherwise} \\
X_4 &= 1 \text{ if shoe use is rarely}, 0 \text{ otherwise}
\end{aligned}
$$

where `shoe_use` being `Everyday` is the baseline level, which occurs when $X_1 = X_2 = X_3 = X_4 = 0$.

Note that the regression coefficient is positive for every one of the above dummy variables, which would suggest that going from everyday shoe use to any less frequent shoe use causes an increase in foot sensitivity. However, based on the t-tests, only the coefficients for `TwoDays` and `Rarely` are significantly different from 0 at level 0.05.

The natural question then arises: can we test if shoe use *in general* has an effect on sensitivity, controlling for sex, age, height, and weight? This can be accomplished using an F test, where we test if:

$$
\begin{aligned}
H_0 &: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \\
H_1 &: \text{ At least one of } \beta_1, \beta_2, \beta_3, \beta_4 \neq 0
\end{aligned}
$$

The regression coefficients in the above hypothesis are all of those which are associated with shoe use, so if we wanted to test if any deviation of shoe use from "everyday" had an effect on sensitivity, the above hypothesis could be of interest. Of course, we would want to control for other factors which have the potential to effect sensitivity, such as age, height, and sex, which is why we include them in the model as well.

Here, we fit the reduced model using the `lm` function:

```
kenyaModR <- lm(sensitivityln ~ sex + age_yrs +
                  height_cm + weight_kg, data = kenya1)
```

The `anova` function in `R` will perform the F test for the above hypothesis, when given the full and reduced model:

```
anova(kenyaModR, kenyaMod)
```

```
## Analysis of Variance Table
##
## Model 1: sensitivityln ~ sex + age_yrs + height_cm + weight_kg
## Model 2: sensitivityln ~ shoe_use + sex + age_yrs + height_cm + weight
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1     71 30.408
## 2     67 26.810  4    3.5981 2.248 0.07305 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here the p-value of 0.073 means that at level 0.05, I fail to reject H_0 that none of the coefficients associated with the different categories of shoe use have an effect on sensitivity. I can also compare the value $F = 2.248$ with the quantile of the F distribution:

```
qf(1-0.05, 4, 67)
```

```
## [1] 2.508695
```

where 4 is the difference in the number of predictors between the full and reduced model, and $76 - (8 + 1) = 67$ is the error degrees of freedom. Because $2.248 < 2.509$, I fail to reject H_0.

In the context of the experiment, this suggests that foot sensitivity is not effected by shoe use.

Squatting:

The data set `squat` contains data on 177 workouts which involved some variation of a barbell squat.

```
slice_sample(squat[-c(2, 6)], n = 6)
```

```
## # A tibble: 6 x 7
##   Count Weight  Reps  Sets Program  Days orm_lombardi
```

```
##   <dbl>  <dbl> <dbl> <dbl> <fct>   <dbl>        <dbl>
## 1   157    295     5     7 NG2       592         347.
## 2    79    260     5     7 NG        274         305.
## 3   175    295     8     5 SF2       647         363.
## 4    17    275     6     3 OG         37         329.
## 5    52    265     4     7 NG        203         304.
## 6    55    285     3     7 NG        210         318.
```

The columns which matter for this analysis:

`Date`: The date on which the workout was performed.

`Days`: Days since January 31st 2022, the date of the first workout.

`Weight`: The weight which was used.

`Reps`: The total number of repetitions performed per set.

`Sets`: The number of sets.

`Program`: A unique identifier for the training program which was being followed.

`orm_lombardi`: An estimated one rep max weight calculated using the following formula: Let w be the weight performed for r reps. Then the Lombardi one-rep-max estimate is given by:

$$ORM = wr^{0.1}$$

Here a simple linear model is fit:

```
mod1 <- lm(orm_lombardi ~ Days, data = squat)
summary(mod1)
```

```
##
## Call:
## lm(formula = orm_lombardi ~ Days, data = squat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -112.752  -16.421    2.915   27.918   59.540
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 319.76072    5.33672  59.917  < 2e-16 ***
## Days          0.05917    0.01419   4.169 4.81e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.55 on 175 degrees of freedom
## Multiple R-squared:  0.09034,    Adjusted R-squared:  0.08514
## F-statistic: 17.38 on 1 and 175 DF,  p-value: 4.805e-05
```

```
plot(squat$Days, squat$orm_lombardi)
lines(squat$Days, predict(mod1))
```

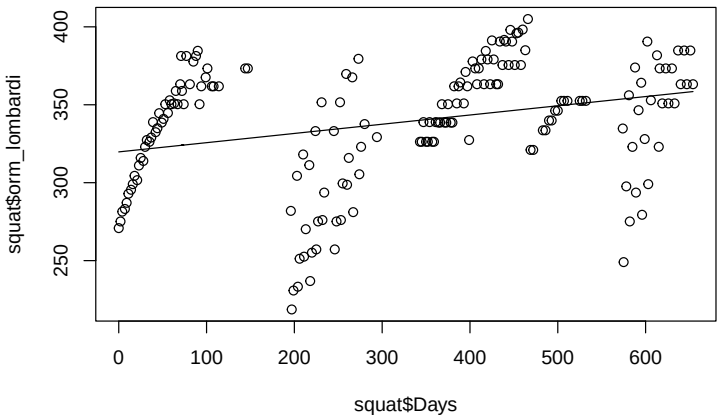


Figure 16: Simple linear model for the squatting data

The coefficient value for `Days` suggests that the change in mean ORM weight (as given by `orm_lombardi`) corresponding to the passage of 1 day (a 1 unit increase in `Days`) is an increase of 0.059 pounds. The t-test suggests that this value is significantly different from 0. We can construct a confidence interval for the value:

```
coef(mod1)["Days"]
```

```
##       Days
## 0.05917419
```

```
confint(mod1)["Days", ]
```

```
##      2.5 %     97.5 %
## 0.03116068 0.08718770
```

Suppose that we want to give a point estimate for the a new observation of `orm_lombardi` on day 466. `R`'s `predict` function can be used to generate a prediction. The `interval` argument being set to `"prediction"` will give a prediction interval along with the estimate:

```
predict(mod1, newdata = data.frame(Days = 466), interval = "prediction")
```

```
##        fit      lwr     upr
## 1 347.3359 272.9058 421.766
```

Recall that a prediction interval is a confidence interval for a new observation Y_0, in this case at a covariate value of $X = 466$:

$$Y_0 = \beta_0 + \beta_1 \cdot 466 + \epsilon$$
$$\hat{Y}_0 - t_{n-2,\alpha/2}\hat{SE}(\hat{Y} - Y_0) < Y_0 < \hat{Y}_0 + t_{n-2,\alpha/2}\hat{SE}(\hat{Y} - Y_0)$$

Now suppose that we wish to test if the *rate* of progression was different for each program. This means that we need to add `Program` as a categorical predictor. There are 6 different programs, so this will require 5 dummy variables in the model. However, we want to test if the *rate* of progression changes, so we should include interaction terms between `Program` and `Days`, which will allow the slope of the regression line to vary with the category of `Program`.

We set `SF` to be the baseline level for `Program`. Then, for example, consider the case when `Program` is `OG`, then the regression model reduces to:

$$Y = (\beta_0 + \beta_{OG0}) + (\beta_1 + \beta_{OG1})X + \epsilon$$

where X is `Days` and Y is the estimated one rep max. β_{OG0} is the coefficient in front of the dummy variable for program `OG`, and β_{OG1} is the coefficient in front of the interaction term between that dummy variable and X.

To test if the rate of progression varies with program in general, we'll perform an F test to see if the coefficients associated with every interaction term are equal to 0:

$$H_0 : \beta_{OG1} = \beta_{NG1} = \beta_{SUM1} = \beta_{NG21} = \beta_{SF21} = 0$$
$$H_1 : \text{At least one of the above coefficients} \neq 0$$

If we fail to reject the null we can conclude that program had no effect on the rate of progression.

Here we fit the full and reduced models required to test the above hypothesis, and perform the test using the `anova` function:

```
mod2 <- update(mod1, . ~ . + Program)
mod3 <- update(mod1, . ~ . + Program*Days)
anova(mod2, mod3)
```

```
## Analysis of Variance Table
##
## Model 1: orm_lombardi ~ Days + Program
## Model 2: orm_lombardi ~ Days + Program + Days:Program
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    170 102155
## 2    165  95745  5    6410.6 2.2095 0.05571 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

At level $\alpha = 0.05$, we (barely) fail to reject the null that the rate of progression does *not* change with program.

If we were to plot the regression line for each model, it is clear that including interaction terms allows the slope to vary with each program.

```
squat %>%
        bind_cols(fitted = predict(mod3)) %>%
        ggplot(aes(x = Days, y = orm_lombardi, color = Program)) +
        geom_jitter() +
        geom_line(aes(y = fitted), linewidth = 0.75) +
        labs(x = "Days", y = "Estimated ORM (lbs)")
```

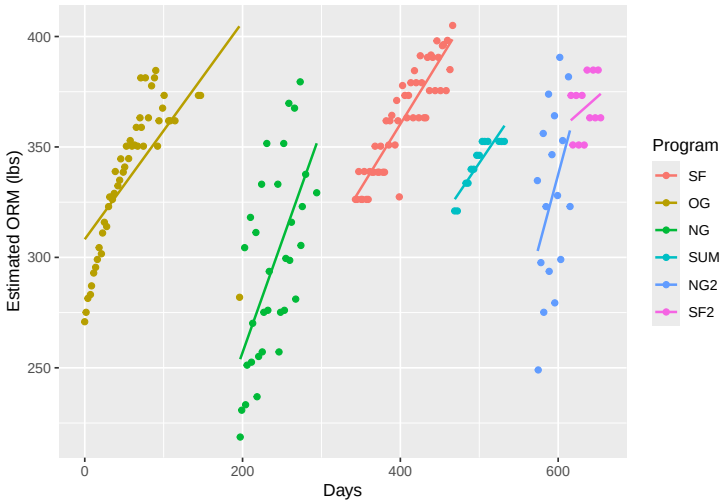


Figure 17: Model with interaction terms

In contrast, if we use the model without interaction terms, the slope must stay constant, but the intercept is allowed to change:

```
squat %>%
        bind_cols(fitted = predict(mod2)) %>%
        ggplot(aes(x = Days, y = orm_lombardi, color = Program)) +
        geom_jitter() +
        geom_line(aes(y = fitted), linewidth = 0.75) +
        labs(x = "Days", y = "Estimated ORM (lbs)")
```

3.33 Transformations of Covariates

Consider the following model:

$$Y_i = \beta_0 + \beta_1 f(X_i) + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \cdots, n$, and where f is any function. It is clear that:

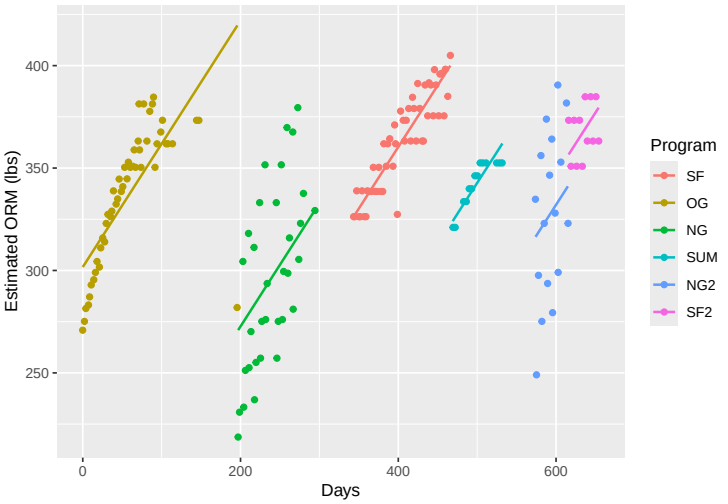


Figure 18: Model without interaction terms.

$$Y_i = \beta_0 + \beta_1 X_i^* + \epsilon_i$$

where $X_i^* = f(X_i)$ for each i.

The second model is of the form of a simple linear regression model, which can be fit using least squares, allowing us to obtain estimates for β_0 and β_1. Generalizing this, if we have

$$Y_i = \beta_0 + \beta_1 f_1(X_{i1}) + \cdots + \beta_p f_p(X_{ip}) + \epsilon_i$$

where $f_1, \cdots, f_p$ are p different functions, we can simply transform every covariate:

$$\begin{aligned} X_{i1}^* &= f_1(X_{i1}) \\ &\vdots \\ X_{ip}^* &= f_p(X_{ip}) \end{aligned}$$

and then fit multiple linear regression using these transformed covariates.

3.34 Transformations of the Response

Consider the model:

$$f(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i$$

Then we could transform Y_i and obtain estimates for the betas using least squares. Define $f(Y_i) = Y_i^*$.

Predictions are for $f(Y_i)$, not Y_i, so to predict values of Y_i, we would invert the function:

$$\hat{Y}_i = f^{-1}(\hat{Y}_i^*) = f^{-1}(\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

All of our usual interpretations, confidence intervals, prediction intervals, etc. now apply to $\hat{Y}_i^*$, not $\hat{Y}_i$. We could back-transform the endpoints of a confidence interval or prediction interval to obtain an interval for $\hat{Y}_i$.

We could also have a slightly more bizzare model that could be transformed into a linear model. Take for instance:

$$Y_i = \beta_0 X_i^{\beta_1} \epsilon_i$$

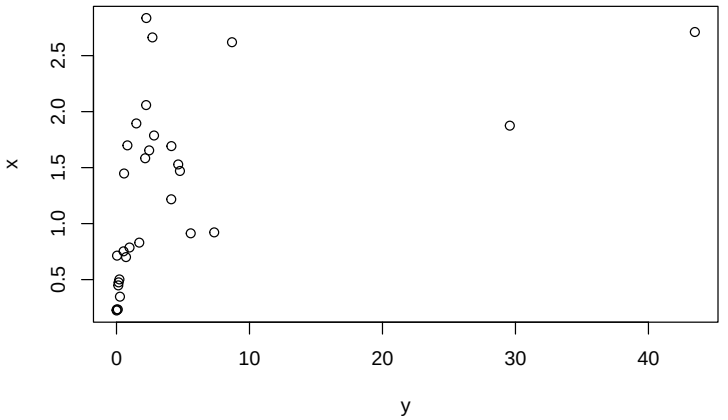
where ϵ_i is such that $E(\log(\epsilon_i)) = 0$. Then,

$$\begin{aligned} \log(Y_i) &= \log(\beta_0) + \beta_1 \log(X_i) + \log(\epsilon_i) \\ &= \beta_0^* + \beta_1 X_i^* + \epsilon_i^* \end{aligned}$$

We then perform linear regression as usual. β_1 which is obtained by fitting the second model is exactly the same as the one which appeared in the original model. The interpretation of the regression coefficients may have significance or no significance depending on the reasoning behind the construction of the original model.

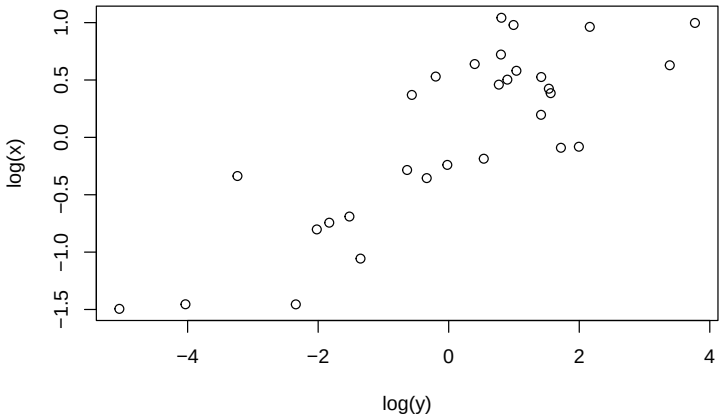
Example: In many cases a plot can elucidate the need for a transformation. Consider the following data `x` and `y`:

```
plot(y, x)
```

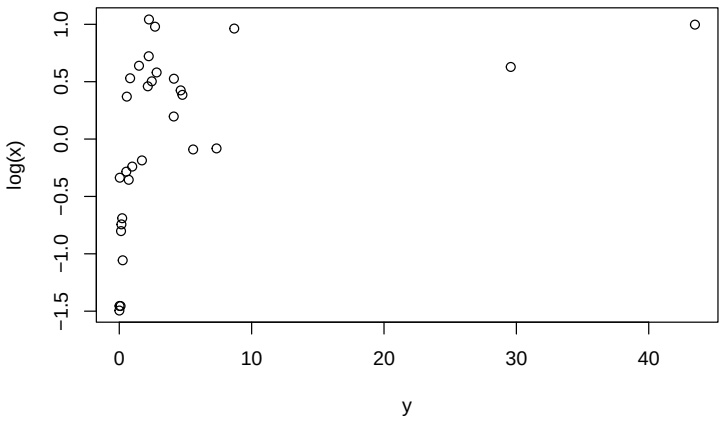


A linear model does not look like it would be a good fit. We can transform using `log` to see when the relationship becomes linear:

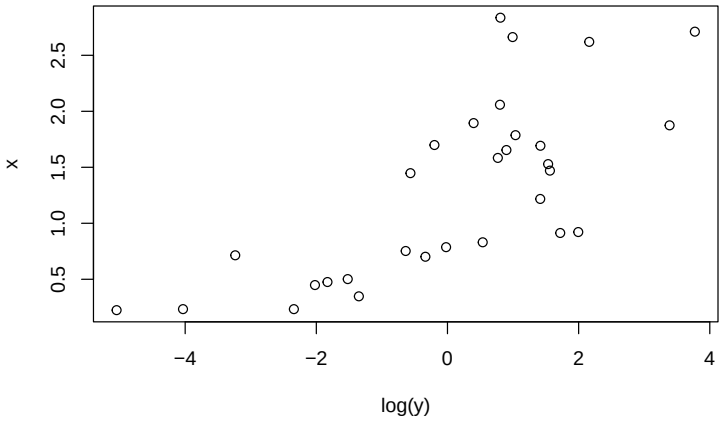
```
plot(log(y), log(x))
```




```
plot(y, log(x))
```




```
plot(log(y), x)
```



Transforming both `x` and `y` with the log produces the most visually linear relationship, so we would use that transformation and then perform least squares as usual.

The log transformation (on the response Y) also has theoretical significance in that it is the variance stabilizing transformation for data which comes from a distribution whose variance is proportional to the square of its expected value (as was the case for the square root transformation with Poisson data).

3.35 Additive Models

The simple additive model is as follows:

$$Y_i = f(X_i) + \epsilon_i$$

Our goal, instead of estimating coefficients (β), is to estimate the function f which relates the response Y to the covariate X.

One way to do this is to assume that f belongs to some function space with a particular basis. We denote the basis as:

$$\{b_1(x), \cdots, b_k(x)\}$$

where $b_1, \cdots, b_k$ are *functions*. Then, we claim that f can be represented as a linear combination of the b functions:

$$f(x) = \sum_{j=1}^{k} \beta_j b_j(x)$$

Estimating f with least squares then becomes easy. We simply create the design matrix:

$$X = \begin{bmatrix} b_1(X_1) & b_2(X_1) & \cdots & b_k(X_1) \\ b_1(X_2) & b_2(X_2) & \cdots & b_k(X_2) \\ \vdots & & & \\ b_1(X_n) & b_2(X_n) & \cdots & b_k(X_n) \end{bmatrix}$$

and then estimate the betas using least squares:

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

to produce the following estimate for the function f:

$$f(x) = \sum_{j=1}^{k} \hat{\beta}_j b_j(x)$$

Example: The simplest example of an additive model is a polynomial basis. This is when we define:

$$b_j(x) = x^{j-1}$$

This means that we claim that f is of the form:

$$f(x) = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k$$

Suppose that we had the following data:

$$Y = (1, 2, 3)$$
$$X = (2.2, 3.9, 8.4)$$

and we want to fit a polynomial of degree 2 to the data (a quadratic). The design matrix would be as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & 2.2 & 2.2^2 \\ 1 & 3.9 & 3.9^2 \\ 1 & 8.4 & 8.4^2 \end{bmatrix}$$

Least squares estimates for β would be obtained via the following matrix product:

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

and we can estimate the function f using:

$$\hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$$

and we estimate Y using:

$$\hat{Y} = \hat{f}(X)$$

Example: The following example from [4] considers the dataset `spdfuel`, which contains ship speed X (in knots) and fuel consumption Y (in tons/day), plotted below.

The two "bends" in the scatter plot suggest the use of a cubic regression model. Using our above notation, let

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2 + \hat{\beta}_3 X^3$$

Fitting the model using the `lm` function in `R` and plotting:

```
mod1 <- lm(fuel_leg3 ~ speed_leg3 + I(speed_leg3^2) + I(speed_leg3^3))
```

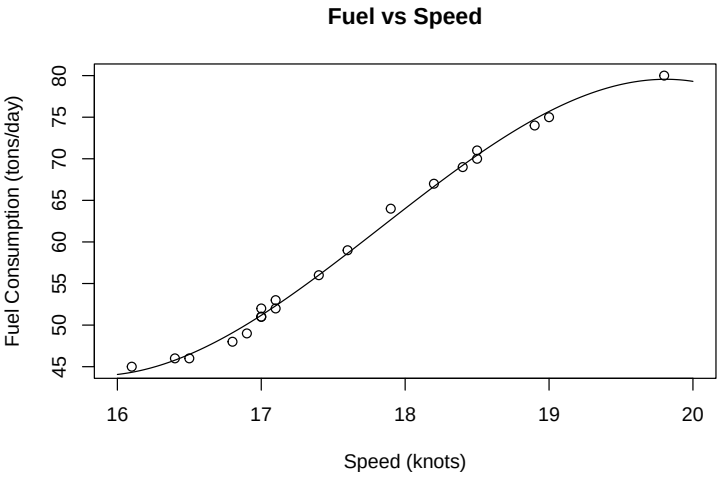


Figure 19: Plot of the data with a third degree polynomial fit

Suppose that we wanted to test if there was a nonlinear relationship between Y and X. Then the test:

$$H_0 : \beta_2 = \beta_3 = 0$$
$$H_1 : \text{One of } \beta_2 \text{ or } \beta_3 \neq 0$$

would be reasonable.

To perform this test, we can fit a reduced model with no quadratic or cubic term and use the F test for a subset of the betas:

```
mod2 <- lm(fuel_leg3 ~ speed_leg3)
anova(mod2, mod1)
```

```
## Analysis of Variance Table
##
## Model 1: fuel_leg3 ~ speed_leg3
## Model 2: fuel_leg3 ~ speed_leg3 + I(speed_leg3^2) + I(speed_leg3^3)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     18 46.369
## 2     16  8.228  2    38.141 37.083 9.831e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a p-value of $9.8 \cdot 10^{-7}$ we reject H_0 and conclude that there exists a nonlinear relationship between X and Y.

Example: Polynomials are not the most flexible choice of basis when trying to estimate a general function f. Another potential option is the *piecewise linear basis*.

Define knots $x_1^*, \cdots, x_k^*$ to be *fixed* values where our piecewise function will be non-differentiable (see the plot below). The piecewise linear basis is then given by:

$$b_1(x) = \begin{cases} (x_2^* - x)/(x_2^* - x_1^*) & \text{for } x < x_2^* \\ 0 & \text{elsewhere} \end{cases}$$

$$b_j(x) = \begin{cases} (x - x_{j-1}^*)/(x_j^* - x_{j-1}^*) & \text{for } x_{j-1}^* < x \leq x_j^* \\ (x_{j+1}^* - x)/(x_{j+1}^* - x_j^*) & \text{for } x_j^* < x \leq x_{j+1}^* \\ 0 & \text{elsewhere} \end{cases} \qquad j = 2, \cdots, k-1$$

$$b_k(x) = \begin{cases} (x - x_{k-1}^*)/(x_k^* - x_{k-1}^*) & \text{for } x_{k-1}^* < x \leq x_k^* \\ 0 & \text{elsewhere} \end{cases}$$

The jth basis function is given as `R` code below:

```
b_j<-function(x, xj, j){
  k<-length(xj)
  if(j>0 & j<k){
    bj<-(x-xj[j-1])*I(xj[j-1]<x)*
      I(x<=xj[j])/(xj[j]-xj[j-1])+(xj[j+1]-x)*
      I(xj[j]<x)*I(x<xj[j+1])/(xj[j+1]-xj[j])
  }
  if(j==1){
    bj<-(xj[2]-x)*I(x<xj[2])/(xj[2]-xj[1])
  }
  if(j==k){
    bj<-(x-xj[k-1])*I(x>xj[k-1])/(xj[k]-xj[k-1])
  }
  return(bj)
}
```

Now we simulate the data according to the following relationship:

$$Y_i = a + \log(b + cX_i) + \epsilon_i$$

This could be fit via nonlinear regression (which allows for the specification of a model with a nonlinear function and an additive error term, like the one above). However, it cannot be fit via transformation alone. Exponentiating both sides, for example, would yield:

$$e^{Y_i} = e^a \cdot (b + cX_i) \cdot e^{\epsilon_i}$$

so our error term is no longer additive and linear regression is not a good model.

```
# Set parameters and generate the data
a<-10
b<-4.05
c<-22.5
n<-100
X<-1:n
eps<-rnorm(n,0,.25)
Y<-a+log(b+c*X)+eps
```

```r
# Design matrix
Xf<-function(x, xj){
  n<-length(x)
  k<-length(xj)
  X<-matrix(rep(0,n*k), ncol=k)
  for(i in 1:n){
    for(j in 1:k){
      X[i,j]<-b_j(x[i],xj,j)
    }
  }
  return(X)
}

# Plot and fit the model
plot(X, Y)
xj<-seq(1,100, length.out=20)                    # 20 uniformly spaced knots
X_m<-Xf(X, xj)
beta<-solve(t(X_m)%*%X_m)%*%t(X_m)%*%Y
Y_hat<-X_m%*%beta
points(X, Y_hat, col="red", type="l")
```

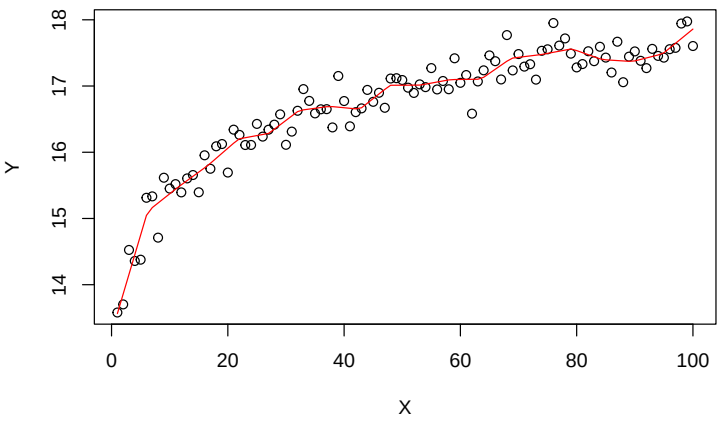


Figure 20: Fit using the piecewise linear basis.

We can penalize "wiggliness" in the piecewise linear basis approximation

by minimizing the following function instead:

$$Q + \lambda \beta^T S \beta$$

where

$$S = D^T D$$

and

$$D = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 & \cdots \\ 0 & 1 & -2 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & -2 & 1 & 0 & \cdots \\ \vdots & & & & & & \end{bmatrix}$$

is a $(k-2) \times k$ matrix.

We still have a closed form solution for $\hat{\beta}$, which was derived in lecture 29:

$$\hat{\beta} = (X^T X + \lambda S)^{-1} X^T Y$$

3.36 Bootstrap

Suppose that we have a random sample $X_1, \cdots, X_n \sim P$, and $T = g(X_1, \cdots, X_n)$ is some statistic (a function of our random sample). Suppose we wish to estimate the standard error of T in order to construct a confidence interval.

If we know the distribution of T, we usually have a good way to do this. In linear regression with normal error terms, for example, we have estimated standard errors for the statistic $\hat{\beta}$ which come directly from estimating the error variance using $\hat{\sigma}^2 = MSE$.

It is often the case that we do not know the distribution of T, as could be the case if we did not assume normal error terms or performed another procedure other than least squares in order to arrive at our estimator $\hat{\beta}$ (such as penalized regression). Bootstrap allows for a general procedure that can be used in these kinds of situations.

In order to perform bootstrap, sample, with replacement $X_1^*, \cdots, X_n^*$ from the random sample $X_1, \cdots, X_n$. Then compute $T_1 = g(X_1^*, \cdots, X_n^*)$. Repeat this until you obtain a bootstrap sample of your desired size, denoted B:

$$T_1, \cdots .T_B$$

Then we estimate the quantity $\text{Var}(T) = \tau^2$ using:

$$\hat{\tau}^2 = \frac{1}{B}\sum_{j=1}^{B}(T_j - \bar{T})^2$$

Suppose that we have an estimator $\hat{\theta}$ for a parameter θ. If $\hat{\theta}$ is *asymptotically* normal (e.g. we can appeal to the CLT in some way), then

$$(\hat{\theta} - z_{\alpha/2}\hat{\tau}, \hat{\theta} + z_{\alpha/2}\hat{\tau})$$

is a $1 - \alpha$ confidence interval for θ.

If we have no asymptotic normality, we can just use the quantiles of the boostrap values of $\hat{\theta}$: $\hat{\theta}_1^*, \cdots, \hat{\theta}_B^*$:

$$(\hat{\theta}^*_{\alpha/2}, \hat{\theta}^*_{1-\alpha/2})$$

where $\hat{\theta}^*_{\alpha/2}$ and $\hat{\theta}^*_{1-\alpha/2}$ are the $\alpha/2$ and $1 - \alpha/2$ quantiles of the sample $\hat{\theta}_1^*, \cdots, \hat{\theta}_B^*$.

3.37 Residual Boostrap

Suppose that

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Our residuals are given by:

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

Sample with replacement from our n residuals $e_1, \cdots, e_n$ to obtain $e_1^*, \cdots, e_n^*$. Then create:

$$Y_i^* = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i^*$$

for each i. Then create estimates $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ using least squares and the above Y_i^* values.

Repeat in order to get a bootstrap sample of $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ values and create confidence intervals using the above techniques.

4 References

[1] 2023. Statistical model. (2023). Retrieved from https://en.wikipedia.org/w/index.php?title=Statistical_model&oldid=1185851310

[2] R. J. Gladstone. 1905. A study of the relations of the brain to to the size of the head. *Biometrika* 4, (1905), 105–123.

[3] Wynands Holowka N.B. 2019. Foot callus thickness does not trade off protection for tactile sensitivity during walking. *Nature* 571, (2019), 261–264.

[4] Larry Winner. 2021. *Statistical regression analysis*. University of Florida.