Data Science

thomas mcandrew

November 21, 2022

Contents

thomas mcandrew

thomas mcandrew

List of Figures

List of Tables

Part I

Probability

1

Sets, the sample space, and probability

thomas mcandrew

Lehigh University

CONTENTS

1.1 Introduction

When we decide to study a topic, we often have in mind a population, a protocol to observe members of our population, a hypothesis that posits some aspects of our observed members. We collect data, we compile results. At the close of our study we want to know how the results ("what happened") contribute to our hypotheses ("what we thought may happen") and, importantly, the chances that if we reproduced the same study that the results would lead us towards the same conclusions.

For example, suppose we decide to study the impact of percutaneous intervention (PCI) versus optimal medical therapy (OMT) to treat patients how have had a myocardial infarction, often called a "heart attack". Patients are enrolled in a study and randomly assigned PCI or OMT. We hypothesize that PCI will result in a prolonged life from when the intervention took place onward. After our study we find patients who underwent PCI live on average 100 days longer than those who underwent OMT. However, it feels unsatisfactory to have a single number describing how effective PCI was at prolong life. We may want to know the chances PCI prolongs life 80 days, 30 days, or maybe zero days, compared to OMT.

We want to understand how likely we are to see specific events. The formal method of assigning chances to a set of events is called probability theory.

1.2 Probability lingo

1.2.1 Sets

We define a **set** as a collection of items, sometimes called elements. A set is typically given a capital letter (for example A) and the elements are included inside curly braces.

Here, we define three sets

$$A = \{a, b, c\} \tag{1.1}$$

$$C = \{-1, 1, 3.14\} \tag{1.2}$$

$$Z = \{\text{orange}, 1, \text{puppy dog}\} \tag{1.3}$$

The elements inside a set are unordered and do not necessarily need to be numbers. A set can contain any object. The set A could have been defined as $\{a, b, c\}$ or $\{c, b, a\}$. We use the symbol $\in$ to express that an element "is a member of" a set. For example, $a \in A$ can be read "the element a is a member of the set A" or "the element a is in the set A". We can also communicate when an element is not a member of a set with the symbol $\notin$. For example, $c \notin C$ is read "the element c is not a member of (not in) the set C".

Generating sets: We can define a set by enclosing in brackets each individual element. $S = \{a, b, c, 10\}$. However, some sets may be easier to build with properties. A property is a function from elements to either true or false. For example we could define the property $P(x) = 1$ when x is even and 0 when x is odd. To build a set that contains elements with a specific property, we write $S = \{x | P(x)\}$ or $S = \{x | x \text{ is an even integer}\}$. This is read "the elements x such that x is an even integer". The vertical bar replaces the words "such that".

Two sets are **equal** is they contain the same elements. In other words, if for an element x, $x \in A$ implies that $x \in B$ **and** if for an element y, $y \in B$ implies that $y \in A$ then the set A and set B are equal. If a set A and set B are equal we write $A = B$. If two sets do not contain the same elements then they are **unequal** and we write $A \neq B$.

Lets look at an example.

$$A = \{a, b, c\} \tag{1.4}$$

$$B = \{b, a, c\} \tag{1.5}$$

$$C = \{b, c\} \tag{1.6}$$

Above, $A = B$ but $A \neq C$ and $B \neq C$.

A set A is a **subset** of a second set B if all of the elements in A are also elements in B. We can say that A is a subset of B if every element in x implies that element is in B, or $x \in A$ implies $x \in B$. We write $A \subset B$ to denote that A is a subset of B. To denote that a set A is **not a subset** of B, we write $A \not\subset B$. In the example above, the C is a subset of A $C \subset A$, but A is not a subset of C, or $A \not\subset C$.

We can build new sets by operating on two or more existing sets.

Set **intersection** takes as input two or more sets, A,B, and returns a set that contains all the elements that are in both A <u>and</u> B. We use the symbol "cap" to denote set intersection $A \cap B$ and often say "A intersect B".

For the above sets A and C, their intersection is

$$A \cap C = \{c, b\} \tag{1.7}$$

because the element c belongs to both set A and set C and because the element b belongs to A and C.

We can intersect more than two sets.

Let $A = \{a, b, 10, 12\}$, $Z = \{b, 10, -1\}$, $U = \{a, b, c, d, 10\}$. Then

$$A \cap Z = \{b, 10\} \tag{1.8}$$

and

$$A \cap Z \cap U = \{b, 10\}. \tag{1.9}$$

If A is a subset of B, $A \subset B$, then the only elements both sets have in common are those in A. The intersection of A and B is then

$$\begin{aligned} &\text{If } A \subset B \text{ then} \\ &A \cap B = A. \end{aligned} \tag{1.10}$$

Set **union** takes as input two or more sets and output a set that contains all the elements that belong to first set <u>or</u> the second set <u>or</u> the third set, and so on. We use the "cup" symbol to denote set union. As an example, consider the sets A, Z, and U to be the same as they are above. Then A **union** Z is

$$A \cup Z = \{a, b, 10, 12, -1\} \tag{1.11}$$

This is because each of the above elements belongs to at least one of the sets A and Z. As another example,

$$A \cup Z \cup U = \{a, b, c, 10, 12, -1\} \tag{1.12}$$

Intersection and union are the most common set operations. Before we define the set operation **complement**, we need to discuss two special sets.

The **universal set** is the set of all elements. We denote the universal set with a $\mathcal{G}$. The set which contains no elements is called the **empty set**, and we denote the empty set as $\emptyset$.

If a set A and set B have no elements in common, then $A \cap B = \emptyset$. We often say that the set A and B are **disjoint**. For example,

$$\begin{aligned} A &= \{1, 3, 5\} \\ Q &= \{2, 4, 6\} \\ A \cap Q &= \emptyset \end{aligned} \tag{1.13}$$

Because the intersection between A and Q is empty, these sets are disjoint.

Now for our final set operation—complement. The set complement takes as input a single set, A, and outputs a set with elements that are members of $\mathcal{G}$ but are not members of A. We denote set complement in one of two ways: A^c or A'.

Set operations example

Let's look at an example of how these sets and set operations work. Define the universal set, $\mathcal{G}$, to be the set of all positive integers (i.e. $1, 2, 3, 4, \cdots$). Lets define the sets $A_1 = \{1\}, A_2 = \{1,2\}, A_3 = \{1,2,3\}, \cdots, A_n = \{1,2,3,4\cdots,n\}$. We use a subscript and a number to "index" our sets. An index is an easy way to keep track of sets (and many mathematical objects).

$$\begin{aligned} A_1 \cap A_2 &= \{1\} \\ A_1 \cap A_5 &= \{1\} \\ A_3 \cap A_8 &= \{3\} \end{aligned} \tag{1.14}$$

Note that $A_3 \cap A_8$ is not the value 3, but the set {3}. The examples above suggest a pattern for any set A_i and A_j where $i \leq j$:

$$A_i \cap A_j = \{i\} \tag{1.15}$$

Lets look at set union in this example

$$\begin{aligned} A_1 \cup A_2 &= \{1,2\} \\ A_1 \cup A_5 &= \{1,2,3,4,5\} \\ A_3 \cup A_8 &= \{1,2,3,4,5,6,7,8\} \end{aligned} \tag{1.16}$$

and we see the following pattern for $i \leq j$

$$A_i \cup A_j = \{1,2,3,4\cdots,j\} = A_j \tag{1.17}$$

Because we defined a universal set, we can look at set complement

$$\begin{aligned} A_1^c &= \{2,3,4,5,\cdots\} \\ A_5^c &= \{6,7,8,9\cdots\} \\ A_8^c &= \{9,10,11,12,\cdots\} \end{aligned} \tag{1.18}$$

1.3 Applying set theory to probability

1.3.1 Foundation

The ideas about sets and set operations are the foundations for how we think about probability, about experiments, and hypotheses. We only need to recast the above set ideas to results from an experiment.

We use $\mathcal{G}$ to define the set of all possible outcomes from an experiment and call this set the **sample space**. The term "experiment" has a broad meaning. An experiment can mean everything from a randomized controlled trial to an observational study. Here an **experiment** is the process that generates outcomes.

An **outcome** is defined as an element of the sample space. An outcome is a single observation from an experiment, and we define an **event** as a set of outcomes. Most often we use o_i to denote an outcome and E_i to denote an event.

The **probability** of an event E is a mapping from the set E to a number between, or equal to, the values 0 and 1. We use $P(E)$ to denote the probability of event E. We require that the probabilities assigned to individual sets consisting of a single element in E add up to the probability of E. Lets suppose there are n outcomes in the event E. Then

$$\begin{aligned} E &= \{o_1, o_2, o_3, \cdots, o_n\} \\ P(E) &= P(\{o_1\}) + P(\{o_2\}) \cdots P(\{o_n\}) \end{aligned} \tag{1.19}$$

We further require $P(\mathcal{G}) = 1$. In other words, the probability that something happens is certain. Note that we do not need to describe how we assign probabilities to events, we only describe what values we expect those numbers to be.

Lets further detail relationships between probabilities of sets that we would expect.

1. If $A \subset B$ then $P(A) \leq P(B)$.
2. $P(A \cup B) \leq P(A) + P(B)$ (why?)
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
4. If A and B are disjoint then $P(A \cup B) = P(A) + P(B)$

The following three axioms (knowledge we assume true without proof) are called **Kolmorogov's axioms** ()

1. For any event E, $P(E) \geq 0$
2. $P(\mathcal{G}) = 1$
3. If E_1 and E_2 are disjoint then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

Example: When we want to compute the probability that an outcome fall in any one of the events, say E_1, E_2, E_3, knowing that these events have no outcomes in common, that they are disjoint, makes the computation easy. An intuitive way to think about events that are disjoint is if they cannot all occur at the same time. Suppose we wish to study the prevalence of stroke among patients who are older than sixty-five. Strokes can be categorized into transient ischemia attacks (TIA), ischemic stroke (IS), and hemorrhagic stroke (HS). Further lets assume the following probabilities for each event: $P(\{\text{TIA}\}) = 0.15$, $P(\{\text{IS}\}) = 0.75$, $P(\{\text{HS}\}) = 0.10$, and that these events cannot occur simultaneously. Then the probability of the event $E = \{TIA, HS\}$ can be broken into the union of two disjoint events $E = \{TIA\} \cup \{HS\}$ and we can use what we know about these disjoint events to compute the probability of E

$$\begin{aligned} P(E) &= P(\{\text{TIA}, \text{HS}\}) && (1.20)\\ &= P(\{\text{TIA}\} \cup \{\text{HS}\}) && (1.21)\\ &= P(\{\text{TIA}\}) + P(\{\text{HS}\}) \quad \text{disjoint} && (1.22)\\ &= 0.15 + 0.75 = 0.90 && (1.23) \end{aligned}$$

1.3.2 The principle of equally likely outcomes—one way to assign probabilities

There are many different ways to assign probabilities to events. In this text, we will only consider using the principle of equally likely outcomes to assign probabilities. The **principle of equally likely outcomes** (PELO) states that the probability we assign to every event which contains a single outcome, $E_i = \{o_i\}$, in a sample space $\mathcal{G}$ is equal.

We can use this principle to assign probabilities first to every individual outcome and then to arbitrary events. Assume the PELO is true, then

$$\begin{aligned} o_1, o_1, o_1, \cdots, o_n \in \mathcal{G} && (1.24)\\ E_i = \{o_i\} && (1.25)\\ E_1 \cup E_2 \cup E_3 \cup \cdots \cup E_n = \mathcal{G} \ \text{(why?)} && (1.26)\\ P(E_1 \cup E_2 \cup \cdots \cup E_n) = P(\mathcal{G}) = 1 && (1.27)\\ P(E_1 \cup E_2 \cup \cdots \cup E_n) = P(E_1) + P(E_2 \cup \cdots \cup E_n) \ \text{(disjoint)} && (1.28)\\ P(E_1) + P(E_2) + \cdots P(E_n) = 1 && (1.29)\\ n \cdot p = 1 \ \ (p \text{ is the constant prob. we are after}) && (1.30)\\ p = 1/n && (1.31) \end{aligned}$$

The PELO tells us that the probability of an event with a single outcome is equal to one divided by the total number of outcomes in the sample space. The PELO also tells us that for a sample space with n elements the probability of an event $E = \{o_1, o_{20}, o_4\}$ that contains three outcomes is $P(E) = \frac{3}{20}$.

An example when, and when not, the PELO applies.

PELO only works when we expect each outcome in our sample space to be equally likely.

Example 1 Suppose our experiment is a coin toss and we will observe whether or not the coin lands heads up (H) or tails up (T). If we assume the coin has not been altered in any way, then PELO applies. Our sample space is $\mathcal{G} = H, T$. Because there are two outcomes in our sample space, $P(\{H\}) = 1/2$ and $P(\{T\}) = 1/2$.

Example 2 Suppose for our experiment we observe for one year a single patient who was admitted to the hospital because of an influenza infection and we plan to observe if that patient returns to the hospital because of a second infection. Our sample space contains one outcome R if the patient is re-admitted to the hospital within one year and a second outcome if they are no re-admitted (N). Our sample space is $\mathcal{G} = R, N$. If we applied the PELO to this problem we would have to assume the probability this patient returns, or does not return, are equal. Intuitively,it feels unreasonable to assume these two events have equal probabilities.

We need an additional way to assign probabilities.

1.3.3 A frequentist approach to assigning probabilities

An empirical approach to assign probabilities to an event A with sample space $\mathcal{G}$ is to do the following:

1. Generate $1, 2, 3, 4, \cdots, N$ outcomes from $\mathcal{G}$
2. Define a variable $N(A)$ and set that variable to zero.
3. If the 1st outcome is a member of A than add one $N(A)$, otherwise move on
4. If the 2nd outcome is a member of A than add one to $N(A)$

   $\vdots$

5. If the Nth outcome is a member of A than add one to $N(A)$

The above algorithm defined a variable $N(A)$ that counts the number of outcomes that belong to A out of the N generated outcomes. We can assign to A the probability

$$P(A) = \frac{N(A)}{N} \tag{1.32}$$

that is, the number of times we observed an outcome that belonged to A divided by the number of outcomes we observed.

As an example, suppose we want to understand the probability that patients with non-medically treated type II diabetes transition to needing treatment within one year of their diagnosis. Our outcomes are $\mathcal{G}$ = {"need treatment" and "no treatment"}. To assign a probability to the event $E = \{\text{need treatment}\}$ using the frequentist approach we decide to request anonymized patient records from a set of hospitals, asking that each hospital only send those patients who had an original diagnosis of type II diabetes without medication and who have had a physician visit one year more later. After our data request, we received 5,000 patient records and find that 1,585 of these patients were asked to start medication for their type II diabetes within one year. A frequentist approach would assign

$$P(E) = 1,585/5,000 = 0.317 \tag{1.33}$$

and also assign to the event $F = \{\text{no treament}\}$

$$P(F) = 0.683 \text{ (why?)} \tag{1.34}$$

.

1.3.4 Products, Conditionals, Baye's Theorem, and Repeated Experiments

1.3.4.1 Product Sets

Suppose a novel vaccine was developed and we are asked to compare two probabilities: the probability a vaccinated patient is infected before 30 days after they receive their vaccine and the probability a patient who has not received a vaccine is infected before 30 days since they have enrolled in the experiment. To estimate these probabilities, we enroll 200 volunteer patients. We will assign 100 patients to be given the treatment and the remaining 100 patients will be observed without treatment.

Because of our experimental design, $P(\{\text{vaccinated}\}) = 1/2$ and $P(\{\text{not vaccinated}\}) = 1/2$ with a sample space $\mathcal{G}_{\text{assignment}}$, containing two outcomes $\{\text{vaccinated}, \text{not vaccinated}\}$. But our main interest is in the probability of infection within 30 days of enrolling in the study, and so our main interest is in the a related sample space $\mathcal{G}_{\text{infection}} = \{\text{infection}, \text{no infection}\}$. We need a way to combine these sample spaces together so that we can estimate the probability of a patient that is vaccinated and that is infected.

Up until this point we have learned how to assign probabilities to events in one sample space. Lets look at how we may assign events to a combination of, and later a sequence of, sample space**s**.

Products of sets: First, we'll need to talk about the product of two sets. Let $A = \{a, b, c\}$ and $B = \{1, 2, 3\}$. Then the Cartesian product $C = A \times B$ is a set where each element in A is paired with each element in B.

$$C = \{(a,1),(a,2),(a,3),(b,1),(b,2),(b,3),(c,1)(c,2)(c,3)\} \tag{1.35}$$

We use the notation $(,)$, called a "tuple", to denote pairs. A **tuple** is an outcome belonging to a sample space built from the Cartesian product of many individual sample spaces. Tuples are ordered. The tuple $(a,1) \neq (1,a)$. The above is called a Cartesian product because you could imagine creating a grid where the horizontal axis has gridlines at "a", "b", "c" and the vertical axis has grid lines at "1", "2", and "3". Where the gridlines intersect are the tuples that belong to C.

We can apply the Cartesian product to combine samples spaces. Define $\mathcal{G}_{\text{experiment}}$ as the Cartesian product between $\mathcal{G}_{\text{assignment}}$ and $\mathcal{G}_{\text{infection}}$, or

$$\mathcal{G}_{\text{experiment}} = \mathcal{G}_{\text{assignment}} \times \mathcal{G}_{\text{infection}} \tag{1.36}$$

This new sample space has the following outcomes:

$$\begin{aligned}\mathcal{G}_{\text{experiment}} = \{&(\text{vaccinated}, \text{infection}), (\text{vaccinated}, \text{not infected})\\ &, (\text{not vaccinated}, \text{infection}), (\text{not vaccinated}, \text{not infected})\}\end{aligned} \tag{1.37}$$

Our new sample space has 4 outcomes, two outcomes from the assignment sample space are paired with two outcomes from the infection sample space (2 outcomes × 2 outcomes equals 4 outcomes in the new space).

With this new space in hand, we can assign probabilities to the outcomes "vaccinated and infected" and the outcome "not vaccinated and infected". You may hear events like the above called **compound events**, or **joint events**.

Let's use the frequentist approach to estimate the probabilities of these four events. We enroll all 200 patients over a 6 months period and then follow each patient for 30 days. At 7 days, 14 days, 21 days, and at the end of there 30 day period they meet with a physician to relay information about how they feel that could indicate they had an infection. We collect the following data:

From our experimental evidence we would assign a probability of $P(\{\text{vacc}, \text{infected}\}) = 0.10$ to those who were vaccinated and infected and

Vaccinated	Infected	Frequency	Estimated Prob.
Yes	Yes	20	20/200 = 0.1
Yes	No	80	80/200 = 0.4
No	Yes	40	40/200 = 0.2
No	No	60	60/200 = 0.3

TABLE 1.1
Frequencies collected about our vaccination experiment

a probability of $P(\{\text{no vacc, infected}\}) = 0.20$ to those who were not vaccinated and were infected. But this was not what were were asked to compute. We were asked to compute the probability that someone is infected after they received (given that they already had) a vaccine, and the probability of infection given a patient was not vaccinated. Probabilities of one event, given we have observed another event are called **conditional probabilities**.

1.3.4.2 Conditional probabilities

Assume a sample space $\mathcal{G}$. We define the conditional probability of event A given event B as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1.38}$$

Example: Lets use the definition above to compute the probability of an event A given $\mathcal{G}$, the sample space (remember that the sample space is a set and so technically an event). By the definition of conditional probbility

$$P(A|\mathcal{G}) = \frac{P(A \cap \mathcal{G})}{P(\mathcal{G})} \tag{1.39}$$

We know that $P(\mathcal{G}) = 1$ (Kolmogorov axioms in **??**) and so

$$P(A|\mathcal{G}) = \frac{P(A \cap \mathcal{G})}{P(\mathcal{G})} \tag{1.40}$$

$$P(A|\mathcal{G}) = \frac{P(A \cap \mathcal{G})}{1} \tag{1.41}$$

$$= P(A \cap \mathcal{G}) \tag{1.42}$$

The set A must be a subset of $\mathcal{G}$ and so by (1.10)

$$P(A|\mathcal{G}) = P(A \cap \mathcal{G}) = P(A) \tag{1.43}$$

The above example shows that we can think of the event that we condition on as a new sample space. Lets return to our example of vaccination and the incidence of infection.

To compute the probability of infection given a patient was vaccinated we need to compute

$$P(\text{inf}|\text{vacc}) = \frac{P(\text{inf} \cap \text{vacc})}{P(\text{vacc})} \tag{1.44}$$

where the event "inf" contains those outcomes in $\mathcal{G}_{\text{experiment}}$ which have "infected" in the fist position of their tuple $(\text{infected}, \cdot)$ and the event "vacc" contains those outcomes which have "vaccinated" in the second position of their tuple $(\cdot, \text{vaccinated})$. Events which hold one or more positions in a tuple constant are often called **marginal events** and the associated probability is called a **marignal probability**. Then $P(\text{inf} \cap \text{vacc})$ are those outcomes with "infected" in the first position and "vaccinated" in the second position. We have a single outcome where this happens: {(infected, vaccinated)} and this outcome has a probability of $P(\text{inf} \cap \text{vacc}) = 0.1$.

Computing $P(\text{vacc})$ is only slightly more difficult. We can use what we learned about unions and disjoint events for help.

$$\begin{aligned} P(\text{vacc}) &= P(\{(\text{infected}, \text{vacc}), (\text{not infected}, \text{vacc})\}) \\ &= P(\{(\text{infected}, \text{vacc})\} \cup \{(\text{not infected}, \text{vacc})\}) \\ &= P(\{(\text{infected}, \text{vacc})\}) + P(\{(\text{not infected}, \text{vacc})\}) \\ &= 0.1 + 0.4 = 0.5 \end{aligned} \tag{1.45}$$

We arrive at our final result.

$$P(\text{inf}|\text{vacc}) = \frac{P(\text{inf} \cap \text{vacc})}{P(\text{vacc})} = \frac{0.1}{0.5} = 0.2 \tag{1.46}$$

and we can do the same for

$$P(\text{inf}|\text{no vacc}) = \frac{P(\text{inf} \cap \text{no vacc})}{P(\text{ no vacc})} = \frac{0.2}{0.5} = 0.4 \tag{1.47}$$

We report back to our experimental team that our estimated probability of infection from someone who received a vaccine is 0.2 and the probability of infection from someone who did not received a vaccine is 0.4, double the probability of a vaccinated individual.

Conditional probabilities give us, for free, an alternative way to compute the probability of the intersection of two sets. All we need to do is rearrange the definition of conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1.48}$$

$$P(A \cap B) = P(B)P(A|B) \tag{1.49}$$

Equation (1.49) is called the **multiplication rule**. Lets explore how the multiplication rule may make computing probabilities easier.

Example: We want to compute the probability that it rains and we remember to leave the house with an umbrella. We can imagine our sample space having the following four outcomes: $\mathcal{G}$ = {(rain,remember),(rain,forgot), (no rain, remember), (no rain, forgot)}. We try to remember all the times that the outcome (rain,remember) occurred in our past—but we cannot. Instead, we can certainly estimate the probability we forgot our umbrella given that it rained because it was an unpleasant experience (for some). We estimate $P(\text{forgot}|\text{rain}) = 0.95$. To compute $P(\text{forgot} \cap \text{rain})$ all we need now is the probability it rains. We take a look at the weather station reports over the past year and use the frequentest approach to estimate the probability it rains:$P(\text{rain}) = 0.25$. So then the probability it rains and we forget our umbrella is

$$\begin{aligned} P(\text{forgot} \cap \text{rain}) &= P(\text{forgot}|\text{rain})P(\text{rain}) \\ &= 0.95 \times 0.25 = 0.2375 \end{aligned} \tag{1.50}$$

The multiplication rule makes computations easier because we can think of events occurring in sequence. First it rains and second we forget our umbrella.

Conditional probabilities also allow us to express what it means for two events to be **independent**. An event A is independent from an event B if

$$P(A|B) = P(A) \tag{1.51}$$

We know the event B occurred, but this has not changed the probability that A has occurred. When two events are independent we can compute their intersection more easily.

$$\begin{aligned} &\text{Event } A \text{ and } B \text{ are independent} \\ &P(A \cap B) = P(A|B)P(B) \qquad \text{multiplication rule} \\ &P(A \cap B) = P(A)P(B) \qquad \text{independence} \end{aligned} \tag{1.52}$$

If two events are independent than the probability they occur together is the product of their probabilities. Statistical independence can be difficult to see.

Example We are recruited to track the evolution of an infectious agent for a team of public health officials (PHOs). To support future strategic planning the PHOs want to know the impact of intervention X on increases or decreases in the incidence of this infectious agent. The PHO team collected for each county in their state whether the intervention was enacted, and whether the incidence of case counts of this infectious agent increased or decreased 60 days after the intervention was in place data. When we look at this table, our first thought may be that when the intervention is enacted there is a high probability that the infection rate decreases (0.525), evidence that this is an important intervention for preventing the spread of this agent.

Intervention	Raise in infection	Probability
yes	yes	0.225
yes	no	0.525
no	yes	0.075
no	no	0.175

TABLE 1.2
Probabilities estimated from our data using a frequentest approach.

But be careful. Let's compute the probability of intervention $P(\text{intervention})$ where we define the event intervention as $\{(\text{intervention}, \text{rise}), (\text{intervention}, \text{no rise})\}$ and the probability of a raise in infection $P(\text{raise})$ where the event raise is defined as $\{(\text{intervention}, \text{rise}), (\text{no intervention}, \text{rise})\}$. To do this we should define our sample space to be clear about what outcomes we can observe from our experiment.

Our sample space is the set of four pairs

$$\begin{aligned} \mathcal{G} = \{&(\text{intervention}, \text{raise}), (\text{intervention}, \text{lower}) \\ &, (\text{no intervention}, \text{raise}), (\text{no intervention}, \text{lower})\} \end{aligned} \tag{1.53}$$

and we want to compute $P(\text{intervention}) = 0.225 + 0.525 = 0.75$ (why?) and $P(\text{raise}) = 0.225 + 0.075 = 0.30$ (why?).

One way that can test whether intervention is independent of a rise or fall in incidence, is compare the probabilities we estimated from our data (in table 1.2) to the product of individual events we computed above. Lets look at an example. The probability of an intervention and rise in incidence was estimated from our data to be $P(\text{intervention}, \text{raise}) = 0.225$. Lets compare this probability to the product $P(\text{intervention}) \cdot P(\text{raise})$.

$$P(\text{intervention}) \cdot P(\text{raise}) = 0.75 \cdot 0.30 = 0.225 \tag{1.54}$$

We have a match. Our estimated probability of an intervention and a rise occurring together is equal to the probability that an intervention occurs and a rise in incidence occurs. We can do the same procedure above for the remaining three scenario.

We see that the probabilities we collected form the data match the product of the probabilities of each individual event (Table 1.3). Intervention is independent, does not change the probability, of a rise or fall in incidence of the infectious agent we are tracking. We will need to report these results to the PHO team and recommend they try alternative interventions to curb the spread of this agent.

Intervention	Raise	Probability from data	Probability assuming indep.
yes	yes	0.225	0.75 (0.30) = 0.225
yes	no	0.525	0.75 (0.70) = 0.525
no	yes	0.075	0.25 (0.30) = 0.075
no	no	0.175	0.25 (0.70) = 0.175

TABLE 1.3
Probabilities of the cooccurance of an intervention and the rise or fall of an infectious agent, and the probability of each of these four events assuming they are independent.

1.3.4.3 The multiplication rule as a tool to compute sequential events

The multiplication rule equips us with a way to break down an event that involves many simultaneous phenomena into a sequence of, potentially easier to compute, probabilities.

Example: Suppose we wish to study the impact diabetes and smoking may have on the probability of a stroke before the age of 50. We decide to observe a population of humans in the US who are all 18 years old and host annual check-ins to record whether each person has acquired diabates, is considered obese, and has experienced a stroke.

Our sample space is $\mathcal{G} = \{(x,y,z)| \; x \in \{\text{diab}, \; \text{no diab}\}$ and $y \in \{\text{obese}, \; \text{no obese}\}$ and $z \in \{\text{stroke}, \; \text{no stroke}\}\}$. For example, one outcome in our sample is $o = (\text{diab}, \text{no obese}, \text{no stroke})$ or the outcome that a patient has acquired diabates, is not considered obese, and has not experienced a stroke.

We may find it difficult to collect data that purposely recorded these three variables together, however it may be easier to find data that has paired diabetes with obesity and diabetes with stroke. The multiplication rule can give us some intuition about the outcomes we wish to study.

Define three events:

1. $D = \{(x,y,z)|x = diab\}$
2. $O = \{(x,y,z)|y = obese\}$
3. $S = \{(x,y,z)|z = stroke\}$

We may find information to assign a probability to O. Lets suppose we found $P(O) = 0.60$. Investigating further we find a dataset that compares those who have obesity and diabetes. This data may alow us to assign the following probabilities $P(D|O) = 0.5$ and $P(D|O^c) = 0.31$. Finally, we find a dataset that measured the prevalence of stroke among patients with-/without diabetes and obesity, giving us the following probability assignments: $P(S|D \cap O) = 0.2$, $P(S|D \cap O^c) = 0.1$, $P(S|D^c \cap O) = 0.15$, and $P(S|D^c \cap O^c) = 0.01$.

If we feel that the above information applies to our study, we can compute the probability of simultaneous events by breaking them into a sequence of conditional probabilities. Suppose we wish to compute $P(D \cap O \cap S)$. This can be separated using the multiplication rule:

$$\begin{aligned} P(D \cap O \cap S) &= P(S|O \cap D)P(O \cap D) && (1.55)\\ &= P(S|O \cap D)P(D|O)P(O). && (1.56) \end{aligned}$$

We can now compute the probability of our simultaneous event with the information we collected across several studies.

$$\begin{aligned} P(D \cap O \cap S) &= P(S|O \cap D)P(D|O)P(O) && (1.57)\\ &= 0.2 \times 0.5 \times 0.60 = 0.06 && (1.58) \end{aligned}$$

If we want, we can visualize all the conditional relationships related to D, O, and S with a **tree diagram**. A tree diagram starts with a dot that represents the "root". We pick an event, say O, and create two branches: one for when O occurs and one for when O^c occurs.

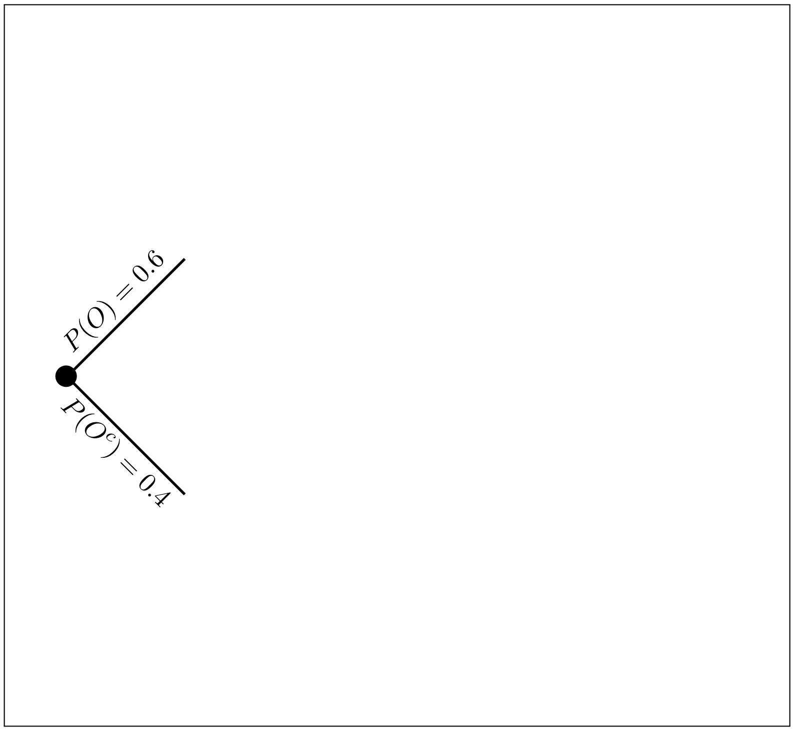


FIGURE 1.1

We pick the next event, say D, and create two branches from O and two branches from O^c for a total of four branches. Conditional on O, two branches represent the occurrence of D and occurrence of D^c. Conditional on O^c, two branches represent the occurrence of D and occurrence of D^c.

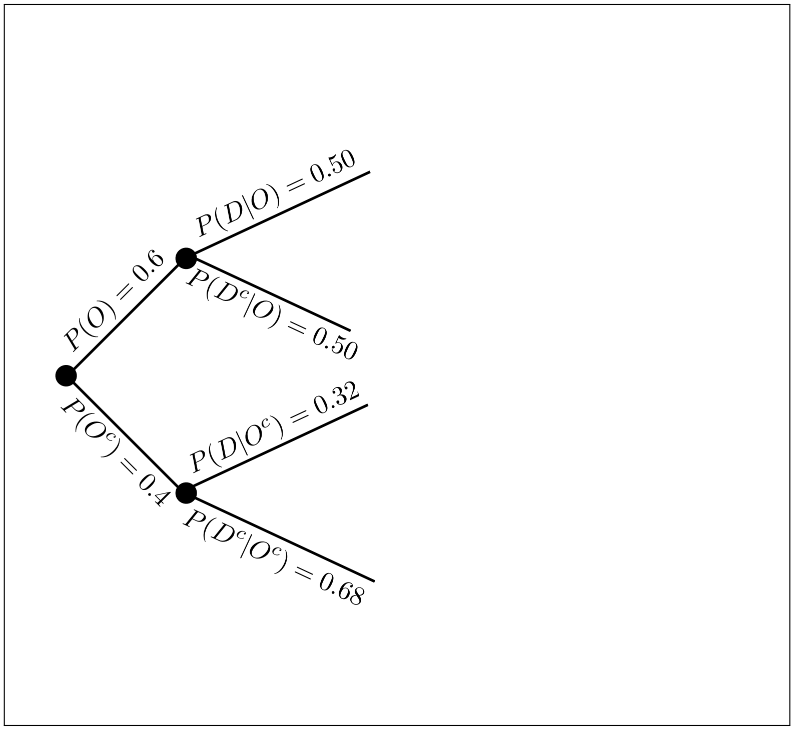


FIGURE 1.2

We can, in this example, do the same procedure for S that we did for D, but this time we need to look at the occurrence of S and S^c for the four newest branches we created.

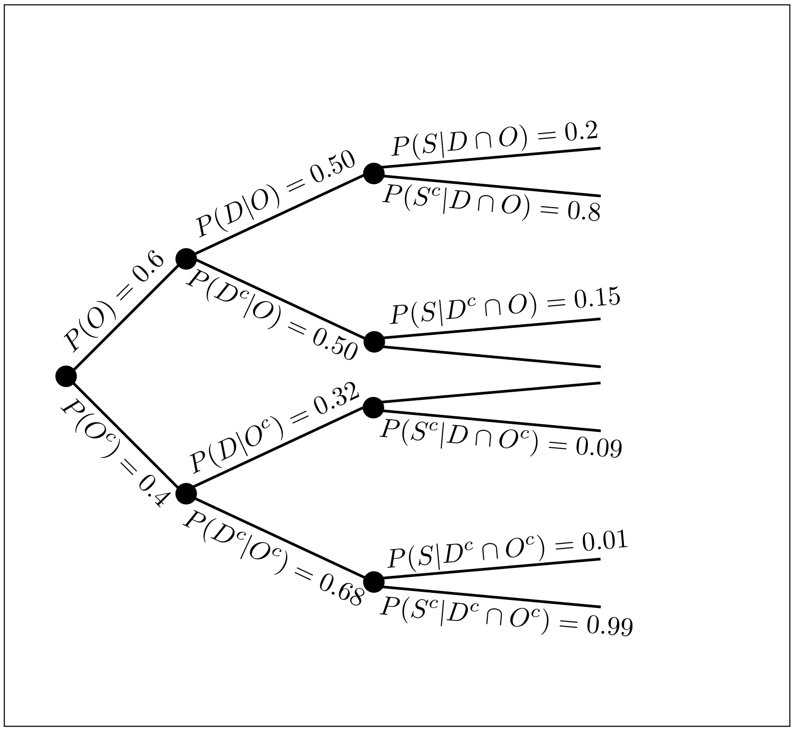

$P(O) = 0.6$
$P(O^c) = 0.4$
$P(D|O) = 0.50$
$P(D^c|O) = 0.50$
$P(D|O^c) = 0.32$
$P(D^c|O^c) = 0.68$
$P(S|D \cap O) = 0.2$
$P(S^c|D \cap O) = 0.8$
$P(S|D^c \cap O) = 0.15$
$P(S^c|D \cap O^c) = 0.09$
$P(S|D^c \cap O^c) = 0.01$
$P(S^c|D^c \cap O^c) = 0.99$

FIGURE 1.3

1.3.4.4 Partitions and the law of total probability

The multiplication rule is one way to use conditional probabilities to make computing probabilities of the intersection of many events more manageable. We can also use conditional probabilities to simplify computing single events easier.

A **partition** of the event A is a collection of sets $\mathcal{P} = \{B_1, B_2, \cdots, B_n\}$ such that $A = (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3) \cdots \cup (A \cap B_n)$ and for each i, k pair of sets $B_i \cap B_k = \emptyset$.

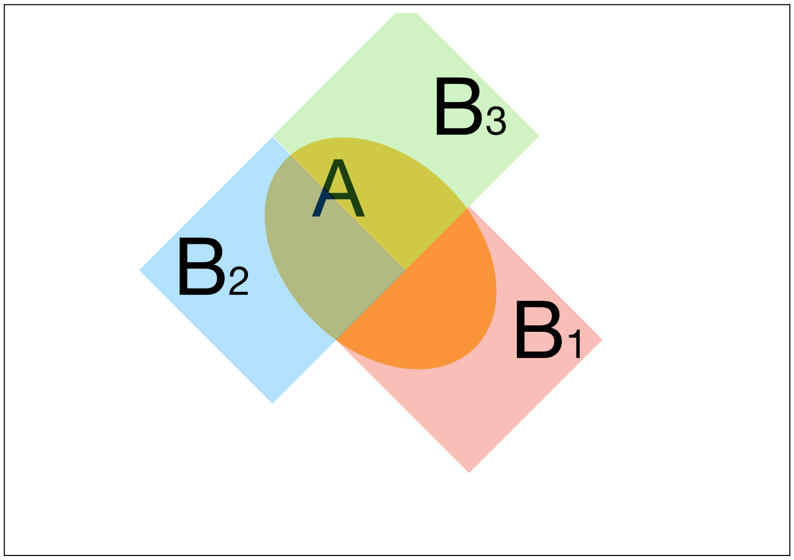


FIGURE 1.4
A partition $\mathcal{P} = \{B_1, B_2, B_2\}$ of the set A

Because B_i and B_k are disjoint for any choice of i and k, then $A \cap B_i$ and $A \cap B_k$ are also disjoint sets. We can use the disjoint property of our partition to break down the probability of A into the sum of probabilities of A intersect B_k for k from 1 to n.

$$\begin{aligned} P(A) &= P((A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3) \cdots \cup (A \cap B_n) && (1.59)\\ &= P(A \cap B_1) + P(A \cap B_2) + + \cdots + P(A \cap B_n) \quad \text{(Disjoint)} && (1.60)\end{aligned}$$

We can further manipulate the above using the multiplication rule to find

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + + \cdots + P(A \cap B_n) && (1.61)\\ &= P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \cdots + P(B_n)P(A|B_n) && (1.62)\end{aligned}$$

This is called **the law of total probability**. Intuitively the law of total probability says that the probability of an event A can be computed by first considering the probabilities of a collection of related events, and then considering the probability that A occurs given each related event.

Example: Let $\mathcal{G} = \{(a,1),(a,2),(a,3),(b,4)\}$ and further suppose we assign the following probabilities to each outcome.

Outcome	Probability
$(a,1)$	0.10
$(a,2)$	0.25
$(a,3)$	0.14
$(b,4)$	0.51

Define the events $A = \{(a,1),(b,4),(a,2)\}$. The finest (as opposed to course) partition is a a collection of sets where each outcome in A is in one, and only one, of the sets $B_1, B_2, ...$ For example a partition of A could be $\mathcal{P} = \{B_1, B_2, B_3\}$ where $B_1 = \{(a,1)\}$, $B_2 = \{(b,4)\}$, and $B_3 = \{(a,2)\}$. A courser partition is $\mathcal{P} = \{B_1, B_2\}$ where $B_1 = \{(a,1)\}$ and $B_2 = \{(a,2),(b,4)\}$.

1.3.4.5 Baye's Theorem

We have investigated how the conditional probability is related to simultaneous events, gives us a natural definition of Independence, and can allow us to compute the probability of an event if we have a partition of that event.

The final relationship we will explore is between two conditional probabilities.

Baye's Theorem relates two conditional probabilities to one another

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.63}$$

We can show that Baye's theorem is true from the definition of conditional probability.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \text{definition} \tag{1.64}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad \text{multiplication rule} \tag{1.65}$$

1.3.4.6 Repeated experiments

Many natural experiments involve repeated observations. It may be of interest to observe the whether and assign a probability to "sunny" or "cloudy" weather conditions. However, many vacationers, who plan to spend 2, 3, or more days at a remote destination may want to know the probability that the next X days are sunny.

The Cartesian product gives us a natural way to express repeated experiments. If we chose to repeat an experiment N times where each experiment produced an outcome in $\mathcal{G}$, we could imagine the set of outcomes as tuples of length N where each entry in the tuple is from our sample space. In other words, a single outcome in this repeated experiment would be $(o_1, o_2, o_3, \cdots, o_N)$ and this outcome is a member of the set $\mathcal{G} \times \mathcal{G} \times \mathcal{G} \times \cdots \times \mathcal{G}$.

Back to our vacation example. If a vacationer wanted to understand the chances they had of enjoying 5 sunny days in a row, then we know that we could defined a sample space $\mathcal{G} = \{\text{"sunny"}, \text{"not sunny"}\}$ and we also know that the vacationer is interested in outcomes in the sample space

$$(d_1, d_2, d_3, d_4, d_5) \in \mathcal{G} \times \mathcal{G} \times \mathcal{G} \times \mathcal{G} \times \mathcal{G} \tag{1.66}$$

where d_i is the outcome on day i. We can frame the problem, and clearly layout the potential outcomes, for example on outcome is ("sunny", "sunny", "not sunny", "sunny", "not sunny").

1.3.5 The datapoint, dataset, and dataframe

The sample space, event, and outcome are all potential results from an experiment. When we conduct an experiment we will generate an outcome from our sample space and call this realized outcome a **data point**.

Example: Consider the experiment of flipping a coin and recording whether the coin lands heads or tails side up. We can define a sample space $\mathcal{G} = \{H, T\}$ where H is an outcome that represents the coin landing heads up and T represents tails up. Up until this point we have structured our experiment, but we have generated no data. We flip the coin and the coin lands tails side up. Now we have performed an experiment and generated the data point T.

Now suppose that we conduct an experiment with the same sample space $(\mathcal{G})$ a number N times and with each experiment we record a data point. A tuple of data points d is called a **data set** $\mathcal{D} = (d_1, d_2, d_3, \cdots, d_N)$ where d_i is the data point generated from the i^{th} experiment. We say that we have drawn or that we have sampled a data set $\mathcal{D}$. Further, data points (d) are often called realized outcomes because they are no longer in a set of potential possibilities but are now determined items.

A data set $\mathcal{D}$ can be unwieldy depending on the number of data points, the complexity of the sample space, or both. A **data frame** is one way to organize a data set. A data frame $\mathcal{F}$ is a table where each data point d in a dataset $\mathcal{D}$

is represented as a row in the table and if the data point is a tuple then a separate column is created for each position in the tuple.

Example: Suppose we design an experiment to collect from humans predictions, two weeks ahead from the time of our experiment, of the number of incident cases and incident deaths at the US national level of COVID-19 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7523166/). We decide to collect from each human whether they are an expert in the modeling of infectious disease, a prediction of incident cases, and a prediction of incident deaths. We draw a data set $\mathcal{D}$ of 50 human judgement predictions. We can organize this data set into a data frame

Expert	Prediction of cases	Prediction of deaths
Yes	145	52
No	215	34
Yes	524	48
Yes	265	95
No	354	35

TABLE 1.4
Example data frame $\mathcal{F}$ built from a data set $\mathcal{D}$ that contains 5 data points where each data point is a tuple of length three.

Above, the first data point is $(Yes, 145, 52)$, the second data point is $(No, 215, 34)$, and so on until the last data point $(No, 354, 35)$. A data frame can also include informative information for others such as labels for each column.

1.4 Exercises

1.

$$A = \{1,2,3,4,5,6\}; \quad B = \{1,3,6\}$$
$$C = \{7\} \quad D = \emptyset$$

(a) Please compute $A \cap B$

(b) Please compute $A \cup C$

(c) Please compute $A \cup D$

(d) Please compute $A \cap D$

(e) Please compute $(A \cap B) \cup (C \cup D)$

2. Let the sample space $\mathcal{G} = \{1, 2, 3, 4, 5, 6, 7\}$
   (a) Please compute A^c
   (b) Please compute B^c
   (c) Please compute D^c
   (d) Please compute $\mathcal{G} \cap A$
   (e) Is $A \subset \mathcal{G}$?
   (f) Is $\emptyset \subset \mathcal{G}$?

3. *Delfina Szigethy—Class of 2025:* 10 students in an Economics class and 10 students in a Biology class are asked to respond to the statement: Pick a number between 0 and 10. The data set collected for the economics class is $E = (1, 2, 4, 4, 6, 7, 7, 8, 9, 10)$ and the data set collected for the biology class is $B = (0, 2, 2, 5, 6, 6, 7, 9, 9, 10)$
   (a) Find the sample space $\mathcal{G}$
   (b) Form the **set** E from the responses from the 10 students in economics
   (c) Form the **set** B from the responses from the 10 students in Biology
   (d) Please compute $E \cup B$
   (e) Please compute $E \cap B$
   (f) Are the sets E and B disjoint?
   (g) Suppose $P(E) = 0.3$. Please compute $P(E^c)$

4. Let the sample space $\mathcal{G} = \{0, 1, 2, a, b, c\}$ and let $A = \{0, 1\}$, $B = \{x | x \text{ is a letter of the English alphabet}\}$
   (a) Please compute $A \cap B$
   (b) Please compute $A \cup B$
   (c) Please compute A^c
   (d) Is $A \cup B = \mathcal{G}$?
   (e) If we assigned probabilities to all outcomes, could $P(A \cup B) = 1$? why or why not?

5. Let $A = 0, 1, 2$ for some sample space $\mathcal{G} = \{0, 1, 2, 3, 4, 5, 6\}$. Further assume $P(A) = 0.2$.
   (a) Are the sets A and A^c disjoint? Why or why not.
   (b) Simplify $P(A \cup A^c)$ into an expression that involves $P(A)$ and $P(A^c)$.
   (c) Use Kolmogorov's axioms to show that $P(A) = 1 - P(A^c)$

6. Let $\mathcal{G} = \{x | x \text{ is a positive integer}\}$
   (a) Are the sets $\emptyset$ and $\mathcal{G}$ disjoint?
   (b) Simplify $P(\mathcal{G} \cup \emptyset)$ into an expression that involves $P(\mathcal{G})$ and $P(\emptyset)$
   (c) Use Kolmogorov's axioms to show that $P(\emptyset) = 0$
7. If $A = \{1, 2, 3\}$ and $B = \{2, 3, 4\}$ and $C = \{1, 3\}$
   (a) Can $P(A) < P(B)$? Why or why not
   (b) Can $P(A) < P(C)$? Why or why not
8. Use what you know about the intersection, about subsets, and about probability to show that $P(A \cap B) \leq P(A)$. Hint: How are $A \cap B$ and A related?
9. *Chinwe Okezie—Class of 2025:* Let $A = \{1, 2, 3, 4\}, B = \{4, 5, 6\}$ and $C = \{2, 4, 6\}$ and assume that the above outcomes follow the Principle of Equally Likely Outcomes.
   (a) Can $P(A) > P(B)$?
   (b) Can $P(B) > P(A)$?
   (c) Can $P(C) < P(B)$?
10. *Audrey Vitello—Class of 2025:* Every year at the town fair there is a pie baking contest among the four best bakers. The best tasting pie wins and the outcome associated with this outcome is the baker's number in the contest. For example, if baker one wins the outcome is the value 1, if baker two wins the outcome is the value 2 and so on.

    In the past ten years Mr. Brown, consistently contestant number 4, has won 3 times.

    (a) Compute P(4)
11. Suppose we wish to study the reemergence of cancer among patients in remition. We collect data on 1,000 patients who are in cancer remition and follow them for 5 years. At five years we are interested in the probability of a second cancer.
    (a) Define a sample space $\mathcal{G}$ we can use to assign probabilities to a second cancer and no second cancer.
    (b) After five years of followup we find that 238 patients experienced a second cancer. Use the frequentist approach to assign probabilities to a second cancer <u>and</u> no second cancer.
    (c) If you collected data on 2,000 patients do you expect the probability of a second cancer to change? How do you expect the probability to be different for 2,000 patients than with 1,000 patients?

12. A study (link = https://www.science.org/doi/10.1126/science.abj8222 found that young adults were 32 times more at risk to develop multiple sclerosis (MS) after infection with the Epstein-Barr virus compared to young adults who were not infected by the virus. The experiment enrolled 10 million young adults and observed them for a period of 20 years.
    (a) Design a sample space if we wish to study outcomes that describe the number of young adults who develop MS.
    (b) Build the event (E_1) "less than 10% of young adults develop MS" using set notation.
    (c) Build the event (E_2) "less than 5% of young adults develop MS" using set notation.
    (d) Are E_1 and E_2 disjoint? Why or why not?
    (e) Can $P(E_1) < P(E_2)$?
13. *Kareem Hargrove—Class of 2025:* Suppose we wish to study the five year incidence of asthma among patients who were infected with SARS-CoV-2. We decide to enroll 10,000 patients who were infected with SARS-CoV-2, follow them for 5 years, and count the number of patients who were diagnosed with asthma at or before five years of followup.
    (a) Define a sample space to describe whether a patient is or is not diagnosed with asthma within five years of followup.
    (b) After the observation period ends we find 4,973 patients were diagnosed with asthma. Use the Frequentist approach to assign a probability to being diagnosed with asthma.
    (c) Use the Frequentest approach to assign a probability to not being diagnosed with asthma.
14. Please compute the following
    (a) $A = \{1,2,3\}$ and $B = \{4,5,6\}$. Please compute $A \times B$ (Answer should be a set of tuples)
    (b) $A = \{1,2,3\}$. Please compute $A \times A$ (Answer should be a set of tuples)
    (c) How many elements are in $A \times A$ (looking for a number)
    (d) How many elements are in $A \times A \times A$? (looking for a number)
    (e) How many elements are in $A \times A \times A \times \cdots \times A$ where we take the Cartesian product N times? (looking for a number)
15. Define a sample space $\mathcal{G} = \{a,b,c,d,1,2,3,4,5\}$ and let $E_1 = \{1,3,5\}$, $E_2 = \{a,b,c\}$, $E_3 = \{a,d,5\}$. We further assign the following probabilities

Outcome	$P(\{\text{Outcome}\})$
a	0.10
b	0.05
c	0.15
d	0.02
1	0.14
2	0.25
3	0.08
4	0.04
5	0.17

(a) Compute $P(E_1)$ (looking for a number)

(b) Compute $P(E_2)$ (looking for a number)

(c) Compute $P(E_3)$ (looking for a number)

(d) Compute $P(E_1 \cap E_2)$ (looking for a number)

(e) Compute $P(E_1 \cap E_3)$ (looking for a number)

(f) Compute $P(E_2 \cap E_3)$ (looking for a number)

(g) Compute $P(E_1|E_2)$ (looking for a number)

(h) Compute $P(E_1|E_3)$ (looking for a number)

(i) Compute $P(E_2|E_3)$ (looking for a number)

(j) Compute $P(E_3|E_2)$ (looking for a number)

16. Define a sample space $\mathcal{G} = \{(a,1),(b,1),(c,1),(a,2),(b,2),(c,2)\}$ and let $E_1 = \{(a,1),(a,2),(c,2)\}$, $E_2 = \{(c,2),(a,1)\}$, $E_3 = \{(b,2)\}$. We further assign the following probabilities

Outcome	$P(\{\text{Outcome}\})$
(a,1)	0.05
(b,1)	0.22
(c,1)	0.15
(a,2)	0.02
(b,2)	0.13
(c,2)	0.43

(a) Compute $P(E_1)$ (looking for a number)

(b) Compute $P(E_2)$ (looking for a number)

(c) Compute $P(E_3)$ (looking for a number)

(d) Compute $P(E_1 \cap E_2)$ (looking for a number)

(e) Compute $P(E_1 \cap E_3)$ (looking for a number)

(f) Compute $P(E_2 \cap E_3)$ (looking for a number)

(g) Compute $P(E_1|E_2)$ (looking for a number)

(h) Compute $P(E_1|E_3)$ (looking for a number)

(i) Compute $P(E_2|E_3)$ (looking for a number)

(j) Compute $P(E_3|E_2)$ (looking for a number)

17. *Imani Ashman—Class of 2024*: A college student has been up studying for their midterm and decides to visit a coffee shop before the exam. The probability that they order a coffee (C) is 0.2, the probability they order a bagel (B) is 0.4, and the probability they order a bagel given that they have already ordered a coffee is 0.3. There is also the potential that they don't order a bagel or a coffee before the midterm (N). The sample space is defined as $\mathcal{G} = \{B, C, N\}$.

    (a) Compute $P(B \cap C)$

    (b) Compute $P(C|B)$

    (c) Compute the probability that they order nothing?

18. Given two events A and B, show that $P(A|B) \geq P(A \cap B)$ (looking for a short explanation and mathematical argument)

19. Suppose we wish to study adverse outcomes among patients who have unprotected left main disease (`https://www.nejm.org/doi/full/10.1056/nejmoa1909406`). In this experiment (called a clinical trial) we randomize patients to receive percutaneous intervention (PCI) or a coronary artery bypass graft (CABG). We wish to study the number of patients who received a PCI and experienced a myocardial infarction (MI) between the time they had their procedure and 5 years, however we only know three pieces of information: (i) the probability a patient was randomized to PCI was 0.5, (ii) the probability a patient was randomized to CABG was 0.5, and (iii) we know that the probability of a MI was 0.2 among patients who received a PCI.

    (a) Define a sample space that will allow us to compute the probability a patient was randomized to PCI and experienced an MI (looking for a description of the sample space and a set)

    (b) Please compute the probability a patient experiences an MI and was randomized to PCI. (looking for a number)

    (c) Please compute the probability a patient does not experience an MI and was randomized to PCI. (looking for a number)

20. If two events A and B are disjoint, are they also independent? (looking for a description and mathematical argument to backup your description.)

21. Researchers studied epidemiological characteristics of a variant of SARS-CoV-2 called B.1.526 (`https://www.science.org/doi/10.1126/sciadv.abm0300`). Suppose we too wanted to study the impact of B.1.526 on a population of patients who have been admitted to the hospital for COVID-19 by collecting each patient's age, whether they have were vaccinated against COVID-19, and whether their infection was due to the B.1.526 variant.

    (a) If we define our experiment as generating the above three pieces of information from a single patient, define a sample space ($\mathcal{G}$) for these potential outcomes. (looking for a short description and the sample space)

    (b) Define a sample space if we repeated our above experiment 2 times. In other words, what would be our sample space if we collected information from 2 patients?

22. Can a data set ever have more elements than a sample space? Explain. (looking for a paragraph with some examples)

23. Suppose we collect the following data about the co-occurence of patients admitted to the hospital for influenza-like illness (ILI) and whether the patient does or does not work in a clinical setting. After data collection we estimate the following probabilities: The probability that a patient works in a clinical setting is 0.12. The probability that a patient who works in a clinical setting is admitted to the hospital for ILI is 0.2, and the probability a patient who does not work in a clinical setting was admitted to the hospital for ILI is 0.3.

    (a) Define a sample space to study outcomes related to the co-occurance of ILI/No ILI and patients who do/do not work in a clinical setting. (Looking for a set and a very brief description of an outcome or two.)

    (b) Compute the probability a patient is admitted to the hospitals for ILI and they work in a clinical setting. (Looking for an equation and number)

    (c) Compute the probability a patient is not admitted to the hospital for ILI and they work in a clinical setting. (Looking for an equation and number)

    (d) Compute the probability a patient is admitted to the hospital for ILI and they do not work in a clinical setting. (Looking for an equation and number)

    (e) Compute the probability a patient is not admitted to the hospital for ILI and they do not work in a clinical setting. (Looking for an equation and number)

24. *Matthew Jung—Class of 2025*: Pretend you are on a game show. To win a prize, you are randomly assigned either door one, door two, or door three. If you are assigned door one, there's a 9% chance to win $20,000, a 20% chance to win $10,000, and a 10% chance to win $40,000. If you are assigned door two, there's a 10% chance to win $50,000 and a 30% chance to win $5,000. If you are assigned door three, there's a 1% chance to win $100,000 and a 20% chance to win $15,000.

    (a) Define the sample space for all of the possible outcomes.

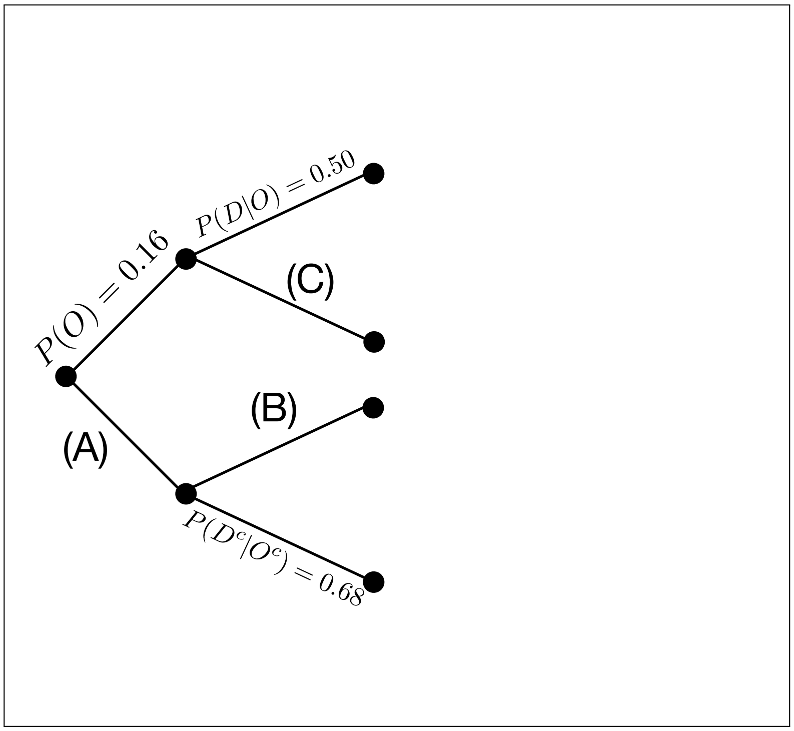
    (b) Build the table of all possible outcomes and their probabilities.

    (c) Compute the probability of winning $40,000 given you get door two.

    (d) Compute the probability of winning less than $40,000 given you get door one.

    (e) Compute the probability of winning $100,000 given you get door three.

25. Show that for two events A and B the $P(A|B)P(B) \leq P(B)$. Why is this the case intuitively? (Looking for two brief descriptions)



26. Please use the above tree diagram to answer the following questions.
    (a) Fill in the $P(O^c)$ which corresponds to (A) in the figure
    (b) Fill in the $P(D|O^c)$ which corresponds to (B) in the figure
    (c) Fill in the $P(D^c|O)$ which corresponds to (C) in the figure
    (d) Please compute $P(D)$
    (e) Please compute $P(D^c)$

27. *Walker Frisbie—Class of 2025:* Suppose we want to study the relationship between the probability you feel ill after riding the roller coaster called the "Intimidator 305" (An actual roller coaster located in VA and my #1 favorite. I have ridden 150 different coasters) and the following three events: (i) we ate two hotdogs just before the ride, (ii) you rode this same coaster 10 times in the past 30 mins, and (iii) you have not had any water and are deydrated.

    Build a probability tree diagram with a root. Next add two branches that represent the events that you ride the Intimidator and that you do not ride the Intimidator. Given you ride the Intimidator with probability 0.9, create three branches for the events that you ate two hotdogs with probability 0.3, rode the coaster 10 times with probability 0.2, and are dehydrated with probability 0.5. Given you did **not** ride the Intimidator, create three branches for the events that you ate two hotdogs with probability 0.6, rode the coaster 10 times with probability 0.2, and are dehydrated with probability 0.2. For each of the six branches you created, add two branches: one for the event you feel ill and one for the event that you do not feel ill.

(a) Assign your own probabilities to each of these ill/not ill events.

(b) Compute $P(\text{feel ill})$

28. Influenza is a contagious virus that enters the respiratory system from a human's nose, throat, or lungs. Common symptoms are cough, chills, and a fever. Suppose we collect data from a local hospital about those who were infected and not infected by the influenza virus and the above three symptoms. Let $\mathcal{G} = \{(x,y)|x \text{ is "Flu" or "not Flu" and } y \in \{\text{cough ,chills , fever}\}\}$. Suppose further that we define events F = { (x,y) | x is "Flu" }, C = { (x,y) | y is "Cough" }, H = { (x,y) | y is "Chills" } , E = { (x,y) | y is "Fever" }, and that $P(C|F) = 0.5$ and $P(C|F^c) = 0.15$, $P(H|F) = 0.25$ and $P(H|F^c) = 0.01$, $P(E|F) = 0.98$ and $P(E|F^c) = 0.35$, $P(F) = 0.02$

(a) A patient presents with the chills. What is the probability they have influenza?

(b) What symptom from the above, if present, would indicate with a high probability a patient has influenza?

1.5 Student Advice, Examples, and Insights

Jessica Berman - Class of 2025-Advice: When I see a problem pertaining to conditional probability it helps me to talk the question out and actually say to myself "what is the probability of A given B has already occured" because P(A | B) is not clear to me without saying it aloud.

Jessica Berman - Class of 2025-Example: A fun fact about conditional probability is that it is used in politics. For example, if there are 4 candidates running in a presidential election, the chances of winning for each of them is 25

Monelli Esfandiary—Class of 2025—Insight: PELO assumes you assign the same probability to every outcome in your sample space. The frequentist approach assigns a probability to each outcome that is proportional to the number of times we observe that outcome in a dataset. Though the PELO and Frequentist approach appear different, we can show that the PELO is a special case of the Frequentist approach. Suppose we create a sample space $\mathcal{G} = \{H, T\}$. PELO would assign P({H}) = C and P({T}) = C, and so (1 = 2C) = (0.5 = C). But we could arrive at the same result using the Frequentist approach if we sampled every outcome from $\mathcal{G}$ once.

Anyah Kumar—Class of 2025—Statistical Independence: The topic from chapter one that I liked the most was statistical independence. Two events, A and B, are statistically independent when one probability does not affect the probability of the other. This may remind you of mutually exclusive events. However, mutually exclusive events cannot be statically independent because when one event occurs the other event **CANNOT** occur. That's why I like statistical independence, because the probability of one event doesn't change the probability of another event which makes computing the probability of the two events simpler.

Ava Barrentine—Class of 2025—Baye's thm: Baye's Theorem is the most interesting to me as it shows the probability of an event, given another has already happened. I never thought one probability could be calculated based on a second event taking place. This theory is useful for so many reasons. For example, this equation can be used to calculate the probability of there being a forest fire when a gender reveal party has happened (probably very low). Say the probability of a forest fire (event A) is $P(A) = 0.3$. The probability of a baby shower happening in/by the woods (B) is $P(B) = 0.05$. The probability of a forest fire happening due to a baby shower $P(A|B) = 0.007$. For Baye's Theorem, we can compute : (0.3 x 0.007)/0.05 = 0.042 = 4.2% chance that given that a forest fire started, a baby shower took place near the woods.

*Josephine Osroagbo—Class of 2025—PELO:*I feel that the Principle of Equally Likely Outcomes does not occur frequently when defining an experiment and assigning probabilities to events. For example, suppose we wish to study the prevalence of COVID-19 hospitalizations among those who are 60 years of age or older with a vaccine to guard against infection and without a vaccine.

We can define a sample space for a single patient as $\mathcal{G} = \{(H, V), (H, V^c), (H^c, V), (H^c, V^c)\}$ where H represents the event that a person is hospitalized and V represents that a person was vaccinated against infection by SARS-CoV-2.

If we applied PELO to this experiment we would have to assume the above four events occur with equal probability. Intuitively, this is likely not the case.

Neha Paras—Class of 2025—A connection between PELO and Uniform: While looking back at my chapter 1 notes, I grazed through the section about PELO. PELO states that the probability of each event in a sample space with a single outcome is equal. This reminded me of the class lesson where we were introduced to the uniform discrete distribution. I thought PELO and the uniform distribution were similar because they both assign to an event with a single outcome a probability of one divided by the number of outcomes in the sample space. I thought the connection between these two concepts was interesting, and it made me think more about how distributions fall outside of uniform. I am excited for future learning about the uniform distribution, for random events, as well as for the normal distribution.

1.6 Glossary

Set:

Subset:

Set equality:

Set intersection:

Set union:

Universal set:

Empty set:

Sample space:

Experiment:

Outcome:

Event:

Probability:

Kolmorogov's Axioms:

Principle of Equally Likely Outcomes:

Product Set:

Compound event:

Conditional probability:

Multiplication Rule:

Independence:

2

Random variables

thomas mcandrew

Lehigh University

CONTENTS

2.1 Introduction

The foundations of probability are built on sets, yet data is more naturally stored and more easily computed on if it is represented numerically.

Random variables match each outcome in our sample space to a value on the number line.

In addition to computational advantages, random variables help us extract from our data the most important characteristics, and they serve as building blocks which we can use to create powerful models. Random variables are also a language we can use to communicate our modeling efforts to other mathematicians, statisticians, and data scientists.

Suppose we hypothesize that the frequency of social media posts on some popular outlet are related to influenza-like illness (ILI)—a syndromic diagnosis suggesting a patient may have influenza. A patient is diagnosed with ILI if their temperature is measured to be at or above 38C and symptoms resembling the flu. Because influenza is most active in winter and spring, we collect a random sample, each day, of social media posts from September to May and in addition we collect the proportion of patients who are admitted to the hospital and are diagnosed with influenza-like illness at the US national level.

The above hypothesis, data collection, and future inference has numerous details. However, we will see shortly that we can simplify our hypothesis by using random variables.

2.2 Maps from the sample space to the number line

Given a sample space $\mathcal{G}$, a **random variable**, (e.g. X), is a function from each element in $\mathcal{G}$—from each outcome—to a value on the real number line. The real number line contains all numbers: integer and decimal, from negative to positive infinity.

Example: Suppose our sample space contains two elements $\mathcal{G} = \{a, b\}$. We may decide to define a random variable X that maps the outcome a to the value -1 and the outcome b to the value 1. In otherwords, $X(a) = -1$ and $X(b) = 1$. We could as well define a random variable Y on the same sample space such that $Y(a) = 0$ and $Y(b) = 1$.

Example: Suppose our sample space contains all integers from 0 to 1000 $\mathcal{G} = \{0, 1, 2, 3 \cdots, 1000\}$. We may be most interested in when an integer is even or odd, and so we can define a random variable $Y(y) = 0$ when y, our outcome, is an odd integer and $Y(y) = 1$ when y is even. This is an example of how a random variable can distill down a sample space with many outcomes into a random variable with two.

Example: Suppose we decide to study the relationship between the cumulative total number of cigarettes smoked by a person form the date that they started smoking and the presence of lung cancer. We define our sample space to be $\mathcal{G} = \{(x,y)|x \in \mathbb{Z}, y \in \{0,1\}\}$. We define two random variables, a random variable X that maps the outcome (x,y) to the value in the first position x, and a random variable Y that maps the $outcome(x,y)$ to the value in the second position y. Though our outcomes are linked, we can use random variables to think about two separate outcomes—cigarettes smoked and lung cancer—and how they interact.

2.3 A new sample space

When we build a random variable (X) that maps outcomes to values on the number line we create a new sample space which we will call the support of X or $supp(X)$. Define a sample space $\mathcal{G}$ without outcomes o_i. Then the **support of X** is

$$supp(X) = \{x|X(o) = x \text{ for some outcome } o \text{ in } \mathcal{G}\} \tag{2.1}$$

Our new sample space is the set of all the potential values that our random variable X can produce. This is a sample space linked to $\mathcal{G}$, but in practice after we develop a random variable we often no longer reference $\mathcal{G}$.

Example: In our above example where $\mathcal{G} = \{a,b\}$, the random variable X has support $supp(X) = \{-1,1\}$ and $supp(Y) = \{0,1\}$. Lets look at another example, when above $\mathcal{G}$ is the set of all integers from 0 to 1000. Even though the sample space is quite large, the random variable that maps the integers to 0 when they are odd and 1 when even has a small support $(supp(Y) = \{0,1\})$.

2.4 How to assign probabilities to a random variable

Random variable themselves do not require that we include the probability of each of their values. Random variables are a function from outcomes to the real numbers—nothing more. That said, in practice we build random variables expecting that the probabilities we assign to outcomes in our sample space will correspond to probabilities assigned to values of our random variable.

We assign a probability to the value x, which belongs in the support of random variable X, the sum of the probabilities of all the outcomes that X maps to x.

$$P(X = x) = P(o_1) + P(o_2) + \cdots + P(o_n) \tag{2.2}$$

where each outcome $o_1, o_2, \cdots, o_n$ is mapped by X to the value x. In other words, $X(o) = x$ for each of $o_1, o_2, \cdots, o_n$.

Example: Define a $\mathcal{G} = \{a, b, c, d, e\}$ and a random variable X that maps the outcomes to the following values

Outcome	P(outcome)	X(outcome)
a	0.1	0
b	0.25	1
c	0.15	1
d	0.3	2
e	0.2	0

We assign the probability that $X = 0$ as the sum of the probabilities assigned to outcome a and outcome e, or

$$\begin{aligned} P(X = 0) &= P(\{a\}) + P(\{e\}) && (2.3) \\ &= 0.1 + 0.2 = 0.3 && (2.4) \end{aligned}$$

We can run the same procedure for all the elements in the support of X,

$$\begin{aligned} P(X = 1) &= P(\{b\}) + P(\{c\}) && (2.5) \\ &= 0.25 + 0.15 = 0.40 && (2.6) \\ P(X = 2) &= P(\{d\}) = 0.3 = 0.30, && (2.7) \end{aligned}$$

and organize our work in a table

X	P(X=x)
0	0.30
1	0.40
2	0.30

A **probability distribution** for a random variable X is a set of tuples where the first position in each tuple is a value in the support of X and the second position in the tuple is the corresponding probability assigned to that value.

Example: A probability distribution for the random variable X above is $\{(0, 0.30), (1, 0.40), (2, 0.30)\}$.

Example: Imagine we run an experiment that collects data on marathon runners. We decide to collect the number of elapsed minutes until they finish the race. Our sample space is defined as all positive integers $\mathcal{G} = \{1, 2, 3, \cdots, \}$. We may decide to build a random variable X that maps outcomes less than 60 to the value 1, outcomes from 61 to 120 to the value 2, and outcomes greater than 120 to the value 3. One potential probability distribution for X is $\{(1, 0.10), (2, 0.50), (3, 0.40)\}$. For this probability distribution, $P(X = 1) = 0.10$, $P(X = 2) = 0.50$, and $P(X = 3) = 0.40$.

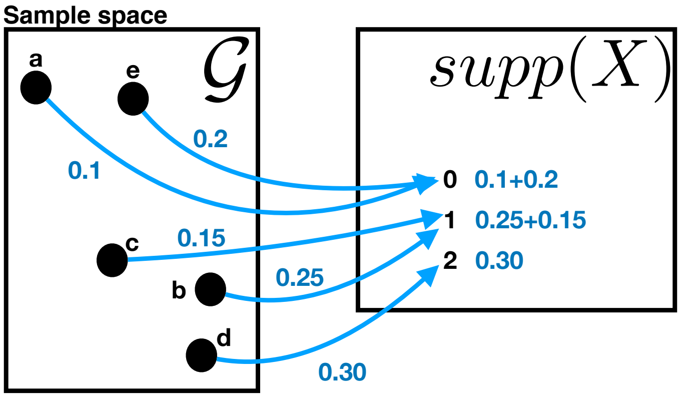


FIGURE 2.1
A sample space $\mathcal{G}$ with elements $\{a, b, c, d, e\}$ and a random variable X that maps each element in $\mathcal{G}$ to one the values: 0, 1, or 2. Probabilities corresponding to outcomes in the sample space and how they map to probabilities for each value of X are shown in blue.

2.5 Probability mass function

There are several supportive tools that we can use to help us better understand random variable we create. The first is the probability mass function, or p.m.f. The **probability mass function** is a function that maps values in the support of a random variable X to their corresponding probabilities. Inputs are values of X, outputs are probabilities.

The probability mass function is a convenient way to organize a probability distribution and it allows us to transfer all the information we know about functions to random variables.

Example: Define a random variable Y with support $\{-1, 0, 1\}$, probability distribution {(-1,0.2),(0,0.5),(1,0.3)}, and probability mass function

$$f(y) = \begin{cases} 0.2 & \text{when } y = -1 \\ 0.5 & \text{when } y = 0 \\ 0.3 & \text{when } y = 1 \end{cases} \tag{2.8}$$

The function—our probability mass function—is a type of function called a **piecewise** function.

We can ask our pm.f. to return the probability for a given value

$$f(1) = 0.3 \tag{2.9}$$

and we can visualize our probability mass function using, for example, a barplot.

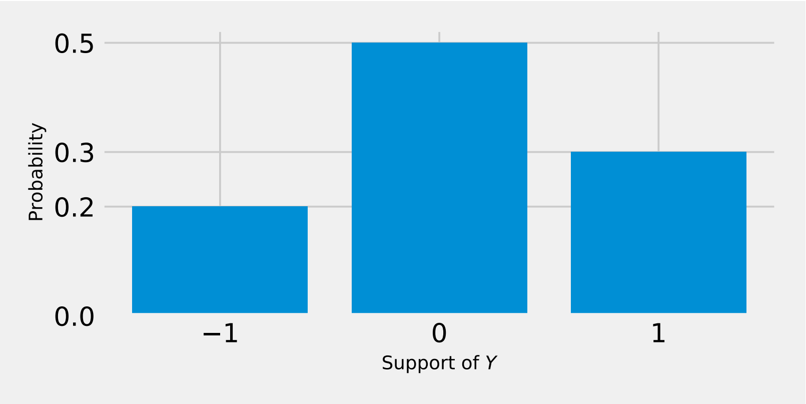


FIGURE 2.2
A barplot for visualizing the probability mass function of our random variable Y. The support of Y is plotted on the horizontal axis and height of each bar corresponds to the probability assigned to that value in the support.

Distributed as f: Because we can use the probability mass function to describe the probability distribution of a random variable, we will often write

$$Y \sim f \tag{2.10}$$

The above formula is read "the random variable Y is distributed as f", and what we mean when a random variable is distributed as f is that the support of Y is the same as the domain of the function f and that the probability of a value y is equal to $f(y)$, or

$$supp(Y) = dom(f) \tag{2.11}$$
$$P(Y = y) = f(y) \tag{2.12}$$

The probability mass function is a convenient method for assigning probabilities to random variables and visualizing the distribution of a random variable.

2.6 Cumulative mass function

The **cumulative mass function** is a function that maps values in the support of a random variable X to the probability that the random variable is less than or equal to this value, or $P(X \leq x)$.

We use a capital F to denote a cumulative mass function (c.m.f). The c.m.f. corresponding to random variable X has a domain equal to the support of X and produces values between 0 and 1 (the values a function can produce is called the function's **image**).

$$supp(X) = dom(F) \tag{2.13}$$
$$image(F) = [0,1] \tag{2.14}$$
$$P(X \leq x) = F(x) \tag{2.15}$$

The c.m.f. too can be visualized and we could also use the c.m.f. to describe the probability distribution of a random variable. This is because we can use the c.m.f. to derive the p.m.f.

Example: For a random variable

$$X \sim f \tag{2.16}$$
$$supp(X) = \{0,1,2,3\} \tag{2.17}$$
$$f(x) = \begin{cases} 0 & 0.1 \\ 1 & 0.3 \\ 2 & 0.2 \\ 3 & 0.4 \end{cases} \tag{2.18}$$

The c.m.f is then

$$F(x) = \begin{cases} 0 & 0.1 \\ 1 & 0.3 + 0.1 = 0.4 \\ 2 & 0.2 + 0.3 + 0.1 = 0.6 \\ 3 & 0.4 + 0.2 + 0.3 + 0.1 = 1 \end{cases} \tag{2.19}$$

We can use the c.m.f $(F(x))$ to compute any p.m.f $(f(x))$ too by noticing that

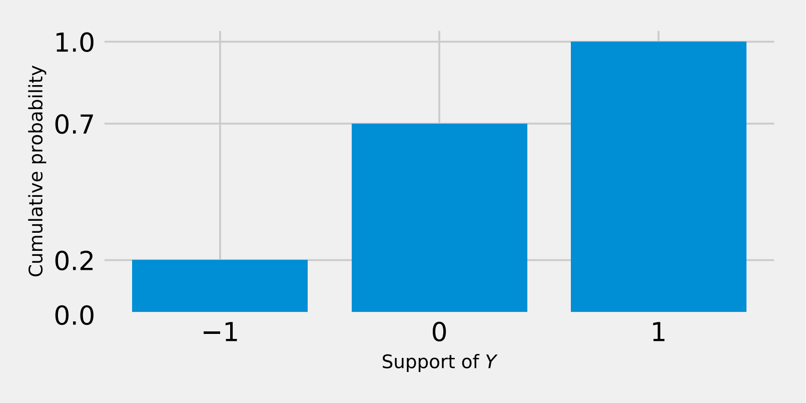


FIGURE 2.3
A barplot for visualizing the cumulative mass function of our random variable Y. The support of Y is plotted on the horizontal axis and height of each bar corresponds to the probability assigned to values (v) less than or equal to v in the support.

for a support $\{x_0, x_1, x_2, x_3, \cdots, x_{n-1} \cdots, x_n\}$ where the values in this set are ordered form smallest to largest

$$f(x_i) = [f(x_i) + f(x_{i-1}) + \cdots f(x_0)] - [f(x_{i-1}) + \cdots f(x_0)] \tag{2.20}$$

$$= F(x_i) - F(x_{i-1}). \tag{2.21}$$

Because the p.d.f. and c.m.f equivalently describe the probability distribution of a random variable we can write $X \sim F$ or $X \sim f$.

2.7 Two or more random variables

A single random variable is a useful way to describe a sample space and the probabilities assigned to values on the number line corresponding to outcomes.

But more complicated scientific questions often require several random variables and rules for how they interact. We will explore how to generate multiple random variables on the same sample space. Then we will develop an approach to define probabilities for combinations of random variables, a joint probability mass function and joint cumulative mass function.

Suppose we design a sample space $\mathcal{G}$ and build two random variables on this space X and Y. Because X and Y are random variables we can think of the support of X and support of Y as two new sample spaces. Lets assume that the outcomes in $supp(X) = \{x_1, x_2, x_3, \cdots, x_N\}$ and that the outcomes in $supp(Y) = \{y_1, y_2, \cdots, y_M\}$. Then we can define a new set of outcomes in the space $supp(X) \times supp(Y) = \{x_i, y_j\}$ for all combinations of i from 1 to N and j from 1 to M.

This new space above maps outcomes in $\mathcal{G}$ to a tuple (x_i, y_j) where the outcomes were mapped by X to the value x_i and by Y to the value y_j. If we assign the probability of (x_i, y_j) to be the the sum of the probabilities of all outcomes in $\mathcal{G}$ that X maps to x_i and Y maps to y_i then we call this a **joint probability distribution**.

Example: A pharmaceutical company launches a clinical trial to study adverse events from a novel medicine. The trial enrolls patients and randomizes them to receive the novel drug or optimal medical therapy. The trial expects three potential adverse events which we call ae_1, ae_2, and ae_3. We may choose to define a sample space $\mathcal{G} = \{(\text{nov}, \text{ae}_1), (\text{nov}, \text{ae}_2), (\text{nov}, \text{ae}_3), (\text{omt}, \text{ae}_1), (\text{omt}, \text{ae}_2), (\text{omt}, \text{ae}_3)\}$ and further build a random variable X that maps outcomes to values 0 when an adverse event was experienced by a control patient and 1 when an adverse event was experienced by a novel drug patient, and a random variable Y that maps adverse event 1 to the value 1, adverse event 2 to the value 2, and adverse event 3 to the value 3.

$$\mathcal{G} = \{(\text{nov}, \text{ae}_1), (\text{nov}, \text{ae}_2), (\text{nov}, \text{ae}_3), (\text{omt}, \text{ae}_1), (\text{omt}, \text{ae}_2), (\text{omt}, \text{ae}_3)\}$$

Event	X
$(\text{nov}, \text{ae}_1)$	1
$(\text{nov}, \text{ae}_2)$	1
$(\text{nov}, \text{ae}_3)$	1
$(\text{omt}, \text{ae}_1)$	0
$(\text{omt}, \text{ae}_2)$	0
$(\text{omt}, \text{ae}_3)$	0

Event	Y
$(\text{nov}, \text{ae}_1)$	1
$(\text{nov}, \text{ae}_2)$	2
$(\text{nov}, \text{ae}_3)$	3
$(\text{omt}, \text{ae}_1)$	1
$(\text{omt}, \text{ae}_2)$	2
$(\text{omt}, \text{ae}_3)$	3

A joint probability distribution would assign probabilities to all possible pairs of values for X and Y.

We write $P(X = x, Y = y)$ to denote the probability assigned to the joint

X	Y	prob
1	1	0.1
1	2	0.05
1	3	0.23
0	1	0.05
0	2	0.30
0	3	0.27

probability that the random variable X equals the value x and the random variable Y equals the value y. With the above example in mind, $P(X = 0, Y = 2) = 0.30$ and this probability could be described as the probability an OMT patient experiences adverse event 2.

Joint probability distributions are often visualized as a table with one random variable's values representing the rows and the second random variable's values representing the columns

	Y = 1	Y = 2	Y = 3
X=0	0.05	0.30	0.27
X=1	0.10	0.05	0.23

A joint distribution can be used to compute probabilities that the random variable X equals the value x for any value of Y—$P(X = x)$ and also the probability that the random variable Y equals the value y for any value of x—$P(Y = y)$. These probabilities are called **marginal probabilities**.

The marginal probabilities that X equals 0, or $P(X = 0)$, is computed by summing the joint probabilities where $X = 0$. We can use the same procedure to compute the marignal probability that $X = 1$.

$$\begin{aligned}
P(X = 0) &= P(X = 0, Y = 1) + P(X = 0, Y = 2) + P(X = 0, Y = 3) \\
&= 0.05 + 0.30 + 0.27 = 0.62 \\
P(X = 1) &= P(X = 1, Y = 1) + P(X = 1, Y = 2) + P(X = 1, Y = 3) \\
&= 0.10 + 0.05 + 0.23 = 0.38
\end{aligned}$$

The same procedure can be done for the random variable Y to find marginal probabilities:

$$\begin{aligned}P(Y=1) &= P(X=0,Y=1)+P(X=1,Y=1)\\ &= 0.1+0.05=0.15\\ P(Y=2) &= P(X=0,Y=2)+P(X=1,Y=2)\\ &= 0.05+0.30=0.35\\ P(Y=3) &= P(X=0,Y=3)+P(X=1,Y=3)\\ &= 0.23+0.27=0.50\end{aligned}$$

Joint distributions need not be restricted to only two random variables. We can build joint distributions of any number of random variables and can compute marginal probabilities in the same way that we do for a joint distribution of two random variables.

The rules, laws, and theorems of probability that we learned in chapter one carry over to random variables. This is because we can consider a new sample space of values that our random variable can take. Suppose we build two random variables X that maps outcomes to the the values -1,0,1 and Y that maps outcomes to the values 1,2,3. We can define a new sample space $\mathcal{G} = \{(x,y) | x \in supp(X) \text{ and } y \in supp(Y)\}$ With our new sample space, we can now discuss statements like $P(X = x|Y = y)$.

Example: If we continue with our pharmaceutical example, we can build a new sample space $\mathcal{G} = \{(0,1),(0,2),(0,3),(1,1),(1,2),(1,3)\}$ and compute, for example, $P(X = 1|Y = 2) = \frac{P(X=1,Y=2)}{P(Y=2)} = \frac{0.05}{0.35} = 0.14$.

We can define the conditional probability of the value of one random variable given another as

$$P(A=a|B=b) = \frac{P(A=a,B=b)}{P(B=b)} \tag{2.22}$$

and so define statistical independence between two random variables, A and B, as

$$P(A=a|B=b) = P(A=a) \text{ for all } a \in supp(A), b \in supp(B) \tag{2.23}$$

We can also translate the law of total probability from events to random variables. For a partition $[Y = y_1] \cup [Y = y_2] \cup [Y = y_3] \cup \cdots \cup [Y = y_n]$ of the event that the random variable $X = x$,

$$\begin{aligned}P(X=x) = P(X=x|Y=y_1)p(y_1) + P(X=x|Y=y_2)p(y_2)\\ +P(X=x|Y=y_3)p(y_3)\cdots + P(X=x|Y=y_n)p(y_n)\end{aligned} \tag{2.24}$$

By thinking of the values of a random variable as outcomes in a new sample space, we can apply our past intuition and the past mechanics of events, sets to a collection of random variables.

2.8 Functions of a random variable

There are times that we may wish to summarize the behavior of a random variable. One common way to describe how probability is distributed among values in the support of a random variable is by computing some function of that random variable.

2.8.1 Multiplying and adding a constant to a random variable

Suppose we have defined a random variable X and wish to study a new random variable $Y = g(X)$ that is a function of X. The values and the probabilities assigned to X, along with the function g will determine the values in the support of Y and assigned probabilities to these values.

2.8.1.1 Translation

Define a random variable X with p.m.f. f_X and a new random variable $Y = X + c$ where c is a constant. Then the support for Y, $supp(Y) = \{x + c | x \in supp(X)\}$ and the p.m.f. of Y, $f_Y(y)$ is

$$\begin{aligned} f_Y(y) &= P(Y = y) && (2.25)\\ &= P(X + c = y) && (2.26)\\ &= P(X = y - c) && (2.27)\\ &= f_X(y - c) && (2.28) \end{aligned}$$

In otherwords, the probability distribution on Y is $\mathbb{P} = \{(y, p) \mid y = x + c \text{ and } P(X = x) = p\}$. This type of transformation of the random variable X to Y is called a **translation**.

Example: Suppose we define a random variable Z such that $supp(Z) = \{0, 1, 2, 3, 4, 5\}$ and

$$f_Z(z) = \begin{cases} 0.10 & \text{if } z = 0 \\ 0.05 & \text{if } z = 1 \\ 0.20 & \text{if } z = 2 \\ 0.12 & \text{if } z = 3 \\ 0.07 & \text{if } z = 4 \\ 0.58 & \text{if } z = 5 \end{cases} \quad (2.29)$$

We can define a new random variable $Y = Z - 4$. This random variable will have $supp(Y) = \{-4, -3, -2, -1, 0, 1\}$ and

$$f_Y(y) = \begin{cases} 0.10 & \text{if } y = -4 \\ 0.05 & \text{if } y = -3 \\ 0.20 & \text{if } y = -2 \\ 0.12 & \text{if } y = -1 \\ 0.07 & \text{if } y = 0 \\ 0.58 & \text{if } y = 1 \end{cases} \tag{2.30}$$

2.8.1.2 Scaling

Define a random variable X with p.m.f. f_X and a new random variable $Y = c \cdot X$ where c is a constant. Then the support for Y, $supp(Y) = \{c \cdot x | x \in supp(X)\}$ and the p.m.f. of Y, $f_Y(y)$ is

$$\begin{aligned} f_Y(y) &= P(Y = y) && (2.31) \\ &= P(cX = y) && (2.32) \\ &= P(X = y/c) && (2.33) \\ &= f_X(y/c) && (2.34) \end{aligned}$$

When we multiply or divide a random variable by a constant to produce a new random variable this is called **scaling**.

Example: Suppose we define a random variable Z such that $supp(Z) = \{0, 1, 2, 3, 4, 5\}$ and

$$f_Z(z) = \begin{cases} 0.10 & \text{if } z = 0 \\ 0.05 & \text{if } z = 1 \\ 0.20 & \text{if } z = 2 \\ 0.12 & \text{if } z = 3 \\ 0.07 & \text{if } z = 4 \\ 0.58 & \text{if } z = 5 \end{cases} \tag{2.35}$$

We can define a new random variable $Y = 4 \cdot Z$. This random variable will have $supp(Y) = \{0, 4, 8, 12, 16, 20\}$ and

$$f_Y(y) = \begin{cases} 0.10 & \text{if } y = 0 \\ 0.05 & \text{if } y = 4 \\ 0.20 & \text{if } y = 8 \\ 0.12 & \text{if } y = 12 \\ 0.07 & \text{if } y = 16 \\ 0.58 & \text{if } y = 20 \end{cases} \tag{2.36}$$

2.8.2 General transformation of a discrete random variable

We explored translation and scaling as specific ways to use one random variable to generate another. However, there are many functions we can apply to a random variable X to create a new random variable Y, and we need a more general method to compute the probability mass function for Y.

Build a random variable X and define Y to be a new random variable such that $Y = g(X)$. Then the support for Y is $supp(Y) = \{y|y = g(x) \text{ and } x \in supp(X)\}$, and

$$f_Y(y) = f_X(x_1) + f_X(x_2) + \cdots f_X(x_n) \tag{2.37}$$

where $x_1, x_2, \cdots x_n \in g^{-1}(y)$. If Y is a function of X then the probability that the random variable Y will assign to the value y is the sum of the probabilities the random variable X assigns to the set of values $x_1, x_2, \cdots, x_n$ that are mapped to y by the function g.

Example: Let X be the random variable with p.m.f

$$f_X(x) = \begin{cases} 0.01 & \text{if } x = -2 \\ 0.05 & \text{if } x = -1 \\ 0.50 & \text{if } x = 0 \\ 0.34 & \text{if } x = 1 \\ 0.10 & \text{if } x = 2 \end{cases} \tag{2.38}$$

and also build a new random variable $Y = X^2$. Here the function g is $g(x) = x^2$. The support of Y is

$$\begin{aligned} supp(Y) &= \{-2^2, -1^2, 0^2, 1^2, 2^2\} && (2.39) \\ &= \{4, 1, 0, 1, 4\} && (2.40) \\ &= \{0, 1, 4\} && (2.41) \end{aligned}$$

To compute $P(Y = 0)$ we need to sum up the probabilities of all values in $supp(X)$ that g maps to 0. In our case the only value mapped to 0 is the value 0, and so $P(Y = 0) = P(X = 0) = 0.50$. To compute $P(Y = 1)$ we need to sum the probabilities of all values that g maps to the value 1. In this case two values in X map to 1, the values -1 and 1, and so $P(Y = 1) = P(X = -1) + P(X = 1) = 0.05 + 0.34 = 0.39$. Finally, $P(Y = 4) = P(X = -2) + P(X = 2) = 0.11$ (why?). The pm.f for Y is

$$f_Y(y) = \begin{cases} 0.50 & \text{if } y = 0 \\ 0.39 & \text{if } y = 1 \\ 0.11 & \text{if } y = 4 \end{cases} \tag{2.42}$$

2.8.3 Expectation

Suppose we build a random variable X with a corresponding probability mass function f_X.

The **expected value** of a random variable X is computed as

$$\begin{aligned} \mathbb{E}\,(X) &= P(X = x_1)x_1 + P(X = x_2)x_2 + \cdots + P(X = x_n)x_n & (2.43)\\ &= f(x_1)x_1 + f(x_2)x_2 + \cdots + f(x_n)x_n & (2.44) \end{aligned}$$

where $x_1, x_2, \cdots, x_n$ are all values in the $supp(X)$.

An intuitive definition of the expected value is that $\mathbb{E}(X)$ is a weighted average of all values in the support of X where the weight for x_i is the probability of x_i. The expected value of X will be close to values in $supp(X)$ with high probability.

Example: Build a random variable Y with support $supp(Y) = \{-1, 0, 1\}$ and $f_Y = \{(-1, 0.2), (0, 0.5), (1, 0.3)\}$. The expected value of Y is $\mathbb{E}(Y) = 0.2(-1) + 0.5(0) + 0.3(1) = 0.1$.

2.8.3.1 Properties of the expectation

The expectation is a linear function, that is $\mathbb{E}(aY + b) = a\mathbb{E}(Y) + b$. We can show this by defining a random variable $Z = aY + b$ and asking

$$\begin{aligned} \mathbb{E}(Z) &= z_1 f_Z(z_1) + z_2 f_Z(z_2) + \cdots z_n f_Z(z_n) & (2.45)\\ &= (ay_1 + b) f_Z(z_1) + (ay_2 + b) f_Z(z_2) + \cdots (ay_n + b) f_Z(z_n) & (2.46)\\ &= a\left[y_1 f_Z(z_1) + y_2 f_Z(z_2) + \cdots y_n f_Z(z_n)\right] \\ &\quad + b\left[f_Z(z_1) + f_Z(z_2) + \cdots + f_Z(z_n)\right] & (2.47)\\ &= a\left[y_1 f_Z(z_1) + y_2 f_Z(z_2) + \cdots y_n f_Z(z_n)\right] + b \;\; \text{(why?)} & (2.48)\\ &= a\left(y_1 f_Y(y_1) + y_2 f_Y(y_2) + \cdots y_n f_Y(y_n)\right) + b & (2.49)\\ &= a\mathbb{E}(Y) + b & (2.50) \end{aligned}$$

The step (2.49) deserves some attention. Values z_i are equal to $ay_i + b$, they are mapped from the values y_i and so the probability that Z equals z_i is equivalent to the probability that Y equals y_i.

2.8.4 Second moment and variance

Define a random variable Y with $supp(Y) - \{y_1, y_2, \cdots, y_n\}$, then **variance** is the following function of Y

$$\begin{aligned} V(Y) = [y_1 - \mathbb{E}(Y)]^2\, P(Y = y_1) + [y_2 - \mathbb{E}(Y)]^2\, P(Y = y_2) + \\ \cdots + [y_n - \mathbb{E}(Y)]^2\, P(Y = y_n) \end{aligned} \tag{2.51}$$

The variance can be thought of as the squared distance of each value in the support of the random variable Y from the expected value weighted by the probability of each value. In some sense, the variance attempts to measure the squared distance from the expected value.

Example: Define a random variable Z with probability mass function

$$f_Z(z) = \begin{cases} 0.14 & \text{if } z = 0 \\ 0.39 & \text{if } z = 1 \\ 0.21 & \text{if } z = 2 \\ 0.26 & \text{if } z = 4 \end{cases} \tag{2.52}$$

To compute $V(Z)$ we need to first compute the expected value of Z or $\mathbb{E}(Z)$:

$$\mathbb{E}(Z) = f_Z(0) \cdot 0 + f_Z(1) \cdot 1 + f_Z(2) \cdot 2 + f_Z(4) \cdot 4 \tag{2.53}$$

$$= 0.14 \cdot 0 + 0.39 \cdot 1 + 0.21 \cdot 2 + 0.26 \cdot 4 \tag{2.54}$$

$$= 1.85 \tag{2.55}$$

Now we can compute the variance

$$\begin{aligned} V(Z) = (0 - 1.85)^2 \cdot f_Z(0) + (1 - 1.85)^2 \cdot f_Z(1) + \\ (2 - 1.85)^2 \cdot f_Z(2) + (4 - 1.85)^2 \cdot f_Z(4) \end{aligned} \tag{2.56}$$

$$\begin{aligned} &= (0 - 1.85)^2 \cdot 0.14 + (1 - 1.85)^2 \cdot 0.39 + \\ &(2 - 1.85)^2 \cdot 0.21 + (4 - 1.85)^2 \cdot 0.26 \end{aligned} \tag{2.57}$$

$$= 1.97 \tag{2.58}$$

2.8.5 Just in time summation

Summing a sequence of values is performed so frequently in mathematics and statistics that we have developed a special notation that simplifies sums.

Given a sequences of values, $x_1, x_2, x_3, \cdots, x_n$, that we wish to sum define the following operator to represent that sum

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n \tag{2.59}$$

2.8.6 How the expected value earned its name

2.8.6.1 Markov inequality

$$P(X > a) < \frac{\mathbb{E}(X)}{a} \tag{2.60}$$

Suppose we define a random variable X with a support that takes only non-negative numbers.

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{x_i \in supp(X)} x_i f_X(x_i) && (2.61)\\
\mathbb{E}(X) &= \sum_{x_i \leq a} x_i f_X(x_i) + \sum_{x_i > a} x_i f_X(x_i) && (2.62)\\
\mathbb{E}(X) &> \sum_{x_i > a} x_i f_X(x_i) && (2.63)\\
\mathbb{E}(X) &> \sum_{x_i > a} a f_X(x_i) && (2.64)\\
\mathbb{E}(X) &> a \sum_{x_i > a} f_X(x_i) && (2.65)\\
\mathbb{E}(X) &> aP(X > a) && (2.66)\\
\frac{\mathbb{E}(X)}{a} &> P(X > a) && (2.67)\\
& && (2.68)
\end{aligned}$$

2.8.6.2 Chebychev's inequality

\nobreak

2.8.7 Covariance

\nobreak

2.8.8 Correlation

\nobreak

2.9 A Model

\nobreak

2.10 Exercises

1. Suppose $\mathcal{G} = \{a, b, c\}$ and $P(\{a\}) = 0.2$, $P(\{b\}) = 0.3$, $P(\{c\}) = 0.5$. Define a random variable X such that $X(a) = 1$, $X(b) = 1$, and $X(c) = 0$. Define a second random variable Y such that $Y(a) = 0$, $Y(b) = 1$, $Y(c) = 2$.

   (a) Compute $P(X = 1)$

   (b) Compute $P(X = 0)$

   (c) What is $supp(X)$?

   (d) What new sample space does X generate?

   (e) Compute $P(Y = 1)$

   (f) Compute $P(Y = 0)$

   (g) What is $supp(Y)$?

   (h) What new sample space does Y generate?

2. Let $\mathcal{G} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$, the set of all positive integers. Further define a random variable K with the following probability mass function

$$\begin{aligned} f(k) &= \left(\frac{1}{2}\right)^k \text{when } k \leq 10 \\ f(11) &= 0.009 \end{aligned}$$

   (a) Is the pmf f a valid probability distribution? Why or why not?

   (b) What value of K is assigned the highest probability?

   (c) Please define the cumulative mass function for the random variable K

3. Define a random variable with $supp(Y) = \{-3, -2, -1, 0, 1, 2, 3\}$ and cu-

mulative mass function

$$F(y) = \begin{cases} 0.10 & \text{when } y = -3 \\ 0.24 & \text{when } y = -2 \\ 0.36 & \text{when } y = -1 \\ 0.50 & \text{when } y = 0 \\ 0.67 & \text{when } y = 1 \\ 0.78 & \text{when } y = 2 \\ 1.00 & \text{when } y = 3 \end{cases}$$

(a) What is $P(Y \leq -1)$

(b) What is $P(Y = -1)$

(c) What is $P(Y = 1)$

(d) Please define the p.m.f for the random variable Y.

(e) Graph the c.m.f

(f) Graph the p.m.f

4. Define the following joint distribution of random variable Q and R that are mapped from the sample space $\mathcal{G}$

Q	R	prob
2	1	0.01
2	2	0.075
2	3	0.13
1	1	0.1
1	2	0.05
1	3	0.17
0	1	0.05
0	2	0.30
0	3	0.115

(a) What is the implied support of Q?

(b) What is the implied support of R?

(c) Compute P(Q=1,R=2)

(d) Compute the marginal probabilities for Q

(e) Compute the marginal probabilities for R

(f) Compute $P(R = 1|Q = 0)$

(g) The random variable Q is called **statistically independent** from R if for every value $q \in supp(Q)$ and for every value $r \in R$ the following is true $P(Q = q|R = r) = P(Q)$. Is Q statistically independent from R? Why or why not?

5. For two events A and B that are statistically independent, we found that $P(A \cap B) = P(A)P(B)$. Please derive an equivalent expression for two random variables by applying the definition of conditional probability and statistical independence to random variables.

6. Define the following joint distribution of random variable Q and R that are mapped from the sample space $\mathcal{G}$

Q	R	prob
2	1	0.01
2	2	0.075
2	3	0.13
1	1	0.1
1	2	0.05
1	3	0.17
0	1	0.05
0	2	0.30
0	3	0.115

   (a) Please compute $\mathbb{E}(Q)$

   (b) Please compute $\mathbb{E}(R)$

   (c) Please compute $V(Q)$

   (d) Please compute $V(R)$

7. *Kimberly Palaguachi-Lopez—-Class of 2024:* Consider the following joint distribution of random variable Z and L that are mapped from the sample space G

Z	L	prob
1	0	0.05
2	0	0.15
3	0	0.19
2	1	0.02
1	1	0.12
3	1	0.04
2	0.5	0.10
3	0.5	0.13
1	0.5	0.20

   (a) Please compute the $supp(Z)$?

   (b) Please compute the $supp(L)$?

(c) Solve for $P(Z = 1, L = 1)$

(d) Solve for the respective marginal probabilities of Z

(e) Solve for the respective marginal probabilities of L

(f) Is the probability distribution for the random variable Z a valid probability distribution? Why or why not? Is the probability distribution for L a valid probability distribution? Why or why not?

8. Suppose X is a random variable with support $supp(X) = \{-2, -1, 0, 1, 2, 3\}$ and $Y = |X|$. Further assume the c.m.f of X is

$$F_X(x) \begin{cases} 0.05 & \text{if } x = -2 \\ 0.15 & \text{if } x = -1 \\ 0.35 & \text{if } x = 0 \\ 0.65 & \text{if } x = 1 \\ 0.95 & \text{if } x = 2 \\ 1.00 & \text{if } x = 3 \end{cases} \tag{2.69}$$

(a) Define the support of Y

(b) Build the p.m.f of Y

(c) Build the c.m.f of Y

9. Compute $\sum_{i=-5}^{i=5} i^2/2$

10. Simplify the $V(X)$ into the equation $E(X^2) - [E(X)]^2$. Hint: Write down the definition of variance using the expectation, then expand the squared terms and simplify.

11. Define the random variable A with pmf

$$f_B(b) = \begin{cases} 0.52 & \text{if } b = 0 \\ 0.12 & \text{if } b = 1 \\ 0.34 & \text{if } b = 2 \end{cases} \tag{2.70}$$

(a) Compute $\mathbb{E}(B)$

(b) Compute $V(B)$

(c) Use Chebychev's inequality to make a statement about $P(B \geq b)$

3

Probability distributions: templates

thomas mcandrew

Lehigh University

CONTENTS

3.1 Introduction

There exist several probability distributions that have been studied in depth. Templates like this allow us to model data generated from an experiment quickly. Random variable templates are so useful that several computer languages have optimized how to compute probabilities, expected values, and higher moments.

We will study parametric distributions for random variables. A set of **parameters—**constants— determine how our random variable assigns probabilities to outcomes.

3.2 Discrete distributions

3.2.1 The Bernoulli distribution

The Bernoulli distribution assigns probabilities to a random variable who support contains the values 0 and 1. The Bernoulli distribution has a single parameter, often denoted θ, that controls how probabilities are assigned to these two values 0 and 1.

To communicate that a random variable Z has a Bernoulli distribution with parameter θ, we write

$$Z \sim \text{Bernoulli}(\theta) \tag{3.1}$$

The parameter θ can take any value between 0 and 1 inclusive, or $\theta \in [0,1]$. The allowable values a set of parameters can take is called the **parameter space**.

The support of Z is $supp(Z) = \{0,1\}$, and the probability mass function for the random variable Z is

$$f_Z(z) = \begin{cases} \theta & \text{if } z = 1 \\ 1-\theta & \text{if } z = 0 \end{cases} \tag{3.2}$$

We can use the probability mass function to compute the expectation

$$\begin{aligned} \mathbb{E}(Z) &= f_Z(1) \cdot 1 + f_Z(0) \cdot 0 && (3.3) \\ &= f_Z(1) \cdot 1 = f_Z(1) && (3.4) \\ &= \theta, && (3.5) \end{aligned}$$

and we can use the probability mass function to compute the variance

$$\begin{aligned} V(Z) &= (1-\theta)^2 f_Z(1) + (0-\theta)^2 f_Z(0) && (3.6) \\ &= (1-\theta)^2\theta + \theta^2(1-\theta) && (3.7) \\ &= \theta(1-\theta)\left[(1-\theta)+\theta\right] && (3.8) \\ &= \theta(1-\theta) && (3.9) \end{aligned}$$

Example: Define $Z \sim$ Bernoulli(0.45). Then $supp(Z) = \{0, 1\}$, $P(Z = 0) = 0.55$, the $P(Z = 1) = 0.45$, $\mathbb{E}(Z) = 0.45$ and $V(Z) = 0.45(0.55) = 0.25$.

Example: A clinical trial enrolls patients and follows them for one year. The clinical team wants to understand the proportion of patients that experience or do not experience an adverse event. We could model whether each patient either experiences or does not experience an adverse event using a Bernoulli distribution. Define Z_i as a Bernoulli distributed random variable for the i^{th} patient in the study. When Z_i = 1 the i^{th} patient experienced an adverse event and 0 otherwise.

3.2.2 The Geometric distribution

If a random variable X has a **Geometric** distribution then the support is the values $supp(X) = \{1, 2, 3, 4, 5, \cdots\}$ and the probability mass function is

$$f_X(x) = p(1-p)^{x-1} \tag{3.10}$$

A random variable that has a geometric distribution has a single parameter p that can take any value between 0 and 1 inclusive, or $p \in [0, 1]$.

The expected value and variance are

$$\mathbb{E}(X) = \frac{1}{p} \tag{3.11}$$

$$V(X) = \frac{1}{p}\left(\frac{1-p}{p}\right) \tag{3.12}$$

A random variable that follows the Geometric distribution often corresponds to an experiment where there is a p probability that an event occurs (often called a success), a $(1-p)$ probability an event does not occur (often called a failure). We assume that each experiment is independent from the previous experiments. The Geometric distribution assigns a probability to the the number of times the experiment is repeated until the event occurs (i.e. the number of repeated experiments until success).

Example: Suppose we want to study the number of times a diagnostic test needs to be implemented until that test detects the presence of viral infection. Further assume we have found in a laboratory setting that the probability this test detects a viral infection is 0.70. We could model the number of attempts until detection as a random variable A, and the rv A would follow a geometric distribution with parameter value equal to 0.70 or $A \sim$ Geom(0.70).

3.2.3 The Poisson distribution

If a random variable X has a Poisson distribution then the support of X is all non-negative integers or $supp(X) = \{0, 1, 2, 3, 4, ...\}$, and the probability mass function is

$$f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{3.13}$$

where $x!$ is read "x factorial" and is defined as

$$x! = x(x-1)(x-2)(x-3)\cdots(2)(1) \tag{3.14}$$

For example, $5! = (5)(4)(3)(2)(1) = 60$. The parameter space for the single parameter λ is all positive real numbers or $\lambda \in (0, \infty)$.

The expected value and variance are

$$\mathbb{E}(X) = \lambda \tag{3.15}$$

$$V(X) = \lambda \tag{3.16}$$

A random variable that follows a Poisson distribution often corresponds to an experiment where the quantity of interest is a rate. A Poisson random variable assigns probabilities to the number of occurrences of an event in a given time period.

Example: The owner of a cafe wants records the number of espressos they produce each day and wants to characterize the probability they produce 0, 1, 2, etc. espressos. For one month the owner records the number of espressos produced per day and find on average that they produce 25 per day. We can model the number of espressos per day as a random variable $X \sim Pois(25)$.

3.2.4 The Binomial distribution

A random variable X distributed Binomial(N, θ) has as support $supp(X) = \{0, 1, 2, 3, 4, 5, \cdots, N\}$, and the probability mass function is

$$f_X(x) = \binom{N}{x}\theta^x(1-\theta)^{N-x} \tag{3.17}$$

where $\binom{N}{x}$ is called a binomial coefficient and is defined as $\binom{N}{x} = \frac{N!}{x!(N-x)!}$. The binomial coefficient is often read "N choose x" and counts the number of ways one can choose x items from a set of N items where the order that the x items is chosen does not matter. For example, $\binom{10}{4}$ counts the number of ways to choose 4 items from a set of 10 items where the order we selected each of the four items does not matter.

The expected value and variance of X are

$$\mathbb{E}(X) = N\theta \tag{3.18}$$

$$V(X) = N\theta(1-\theta) \tag{3.19}$$

Given N observations, the binomial distribution assigns probabilities to the number of observations that experience an outcome of interest where we assume that the probability any single observation experiences the event is θ.

Example: Imagine we randomize 200 patients in a clinical trial where 100 are enrolled to receive a novel treatment and 100 are enrolled to receive a control treatment. In the treatment group, 10 patients experience an adverse event from the treatment and in the control group 15 patients experience an adverse event. In previous work we found that the probability any one patient experiences an adverse event in the treatment group is 0.02 and in the control group is 0.04. We can define a random variable $T \sim \text{Bin}(100, 0.02)$ that assigns a probability to the number of patients who experience an adverse event and define a random variable $C \sim \text{Bin}(100, 0.04)$ that assigns probabilities to the number of patients who experience an event in the control group.

3.2.5 The Discrete Uniform distribution

A random variable X has a uniform discrete distribution $U(\alpha, \beta)$ is this random variable has a support equal to $supp(X) = \{\alpha, \alpha+1, \alpha+2, \cdots, \beta\}$ and a probability mass function equal to

$$f_X(x) = \frac{1}{N} \tag{3.20}$$

$$N = \beta - \alpha + 1 \tag{3.21}$$

where N count the number of outcomes between α and β inclusive.

The expect6ed value and variance of X are

$$\mathbb{E}(X) = \frac{\alpha+\beta}{2} \tag{3.22}$$

$$V(X) = \frac{N^2-1}{12} \tag{3.23}$$

The parameters α and β for a uniform discrete distribution can take any integer value so long as the constraint $\alpha < \beta$ is satisfied.

Example: A game of dice is played between two friends. The die that roll has six sides. We can model the probability the first player rolls any of the one through six values as $F \sim U(1, 6)$ and the probability the second player rolls values between one and six as $S \sim U(1, 6)$.

3.3 Continuous distributions

Up until this point we have characterized discrete sample spaces and random variables that are defined on discrete sample spaces, called discrete random variables.

A **continuous sample space**, $\mathcal{G}$, is a subset of the real line ($\mathbb{R}$). There are an infinite, uncountable number of outcomes possible in a continuous sample space and this introduces some oddities when we try to assign probabilities on a sample space like this.

A **continuous random variable** is a function from a continuous sample space to the real line.

3.3.1 Discrete to continuous: Points to intervals

Consider a discrete sample space with outcomes $\mathcal{G} = \{0, 1/4, 2/4, 3/4, 1\}$, and define a random variable X that has a uniform discrete distribution with parameters 0 and 1 or $X \sim U(0, 4)$.Because there are 5 outcomes the probability of each outcome is $1/5$.

Let's increase the size of our discrete sample space to contain 100 values between 0 and 1 by setting $\mathcal{G} = \{0, 1/99, 2/99, 3/99, \cdots, 98/99, 1\}$. If we define a uniformly distributed random variable on this space then the probability of each outcome is $1/100$.

As we increase the number of points in our sample space, the probability assigned to each individual point will shrink towards zero. In a continuous space the probability of any individual value is zero.

Because the $P(X = x) = 0$ for a continuous random variable defined on a continuous $\mathcal{G}$, we need a new way to define probabilities for a continuous random variables.

3.3.2 Densities

For a continuous random variable X with support $supp(X) = A$, where A is an interval of the real line, we will assign non-negative probabilities to any interval that is a subset of A according to a probability density function f_X.

A probability density function f_X will help us compute the probability over an interval in A. A probability density must have values that are non-negative (i.e. on or above the horizontal line x=0 in an xy-plane) and the area under this probability density function for the set A must equal one.

The probability that a random variable with pdf f_X assigns to the interval $(a, b) \subset A$ is the area under f_X from a to b.

<picture>

3.3.3 Uniform continuous distribution

For a discrete random variable, X, the discrete uniform distribution from α to β assigns the same probability to every outcome $\alpha, alpha + 1, \cdots, \beta$.

For a continuous random variable Y, the continuous uniform distribution from a to b or $Y \sim U(a, b)$ will assign the same probability to any interval of the same length between the values a and b.

The probability density function on the set $A = (a, b)$ is

$$f_Y(y) = \frac{1}{b - a} \tag{3.24}$$

To compute the probability of an interval, lets say (c, d) where $c > a$ and $d < b$, we need to compute the area under $f_Y(y)$ from c to d.

<picture>

The area under $f_Y(y)$ is equivalent to the area of a rectangle with width $d - c$ and height $\frac{1}{b-a}$. This area equals $\frac{d-c}{b-a}$ and so

$$P(c < Y < d) = \frac{d - c}{b - a} \tag{3.25}$$

We can also define the cumulative density function (cdf), $F_Y(y)$, as

$$F_Y(y) = P(Y < y) = P(a < Y < y) = \frac{y - a}{b - a} \tag{3.26}$$

The cdf is an important function for continuous random variables because this function inputs a value y and returns the probability over all values less than y—an interval.

<picture>

The expected value and variance are

$$\mathbb{E}(Y) = \frac{a + b}{2} \tag{3.27}$$

$$V(Y) = \frac{(b - a)^2}{12} \tag{3.28}$$

3.3.4 Just-in-time integration

We recognized that the area under the probability density function of a continuous uniform random variable $f_Y(y)$ is the same as the area of a rectangle.

But there are several continuous random variables with non-linear densities that are not easy to compute. We need a special tool and special notation to handle computing areas under these densities.

Suppose we define a continuous random variable Z that has two parameters, μ and σ, with the following probability density function

$$f_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} \tag{3.29}$$

defined over all negative and positive numbers or $\mathbb{R}$.

<pciture>

The area under this density is not at all straight forward to compute. However, we may be able to use the area of several rectangles—areas that we can compute—to approximate the area under this more complicated curve. Approximating the area under a curve with a set of simpler curves is called **Riemann integration**.

Lets suppose we are interested in

$$P(-1 < Z < 1) \tag{3.30}$$

which is the area under $f_Z(z)$ from the value -1 to the value 1.

<picture>

One way to approximate this area is to split up the interval $I = (-1, 1)$ into (lets say) 20 pieces $P = \{(-1, -9/10), [-9/10, -8/10), \cdots, [1/10, 2/10), [2/10, 3/10], \cdots [9/10, 1)\}$. For each smaller interval P_i we can choose a single point z_i, where z_{-10} is a point between -1 and -9/10, the point z_{-9} is a point between -9/10 and -8/10 and so on. We can create rectangles with width 1/10 and height $f(z_i)$ and sum these rectangles to approximate the area under f_Z.

<picutre>

The area of these rectangles is

$$P(-1 < z < 1) \approx \sum_{i=-10}^{10} \frac{1}{10} f(z_i) \tag{3.31}$$

As we split this interval into more and more intervals the rectangles will better and better approximate the area under f_Z.

<picture>

If these exists a number S such that the more splits we create the approximate area gets closer to S, then S is called the integral of f_Z from -1 to 1 and is the exact area under f_Z.

We write

$$\int_{z=-1}^{z=1} f_Z(z)dz \tag{3.32}$$

to represent the area under the probability density f_Z between the values -1 and 1.

Example: For a continuous random variable $X \sim U(a,b)$ the integral from a to a value x is the area of a rectangle. $\int_{x=a}^{x=x} f_X(x)dx = \frac{x-a}{b-a}$.

Example: The cumulative density function evaluate at a value y for a random variable Y is the area under the probability density function from the smallest possible value in the support of Y to y. This can be represented with an integral $F_Y(y) = \int_{y=-\infty}^{y=y} f_Y(y)dy$.

3.3.5 Normal Distribution

A continuous random variable X has a Normal distribution with parameters μ and σ if the support of X is all real numbers and the probability density function is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \tag{3.33}$$

The normal distribution has several important properties that apply **only to the normal distribution**.

3.3.5.1 The expected value and variance are equal to distinct parameters

The expected value and variance of X are

$$\mathbb{E}(X) = \mu \tag{3.34}$$

$$V(X) = \sigma^2 \tag{3.35}$$

For most continuous distributions the expected value and variance are a combination of all the paramter values. The normal distribution is unique in that the expected value only contains a single parameter μ and the variance only contains a unique parameter σ.

3.3.5.2 Shifting and scaling by a constant is easy

If a random variable $Y = c + X$ where $X \sim \mathcal{N}(\mu, \sigma^2)$ then the distribution of Y is $Y \sim N(\mu + c, \sigma^2)$. If a random variable $Y = cX$ where $X \sim \mathcal{N}(\mu, \sigma^2)$ then the distribution of Y is $Y \sim N(\mu, c^2\sigma^2)$.

These shifting and scaling properties are unique to the normal distribution.

3.4 Quantiles and the quantile function

A $(100 \times p)$ **quantile** is a value x such that

$$P(X < x) \leq p \tag{3.36}$$
$$P(X \leq x) \geq p \tag{3.37}$$

where $p \in [0, 1]$.

If X is a discrete random variable the $P(X < x)$ may not equal $P(X \leq)$. However when X is a continuous random variable the $P(X < x)$ will equal $P(X \leq x)$ and so for a continuous random variable the $(100 \times p)$ quantile is a value x such that

$$P(X < x) = p \tag{3.38}$$

Example: Define $X \sim \mathcal{N}(0, 1)$.Then The 97.5th quantile for X is the value 1.96 because $P(X < 1.96) = 0.975$. The 2.5th quantile for X is the value -1.96 because $P(X < -1.96) = 0.025$. The 90th quantile for X is the value 1.28 because $P(X < 1.28) = 0.90$ and so the value -1.28 is the 10th quantile because $P(X < -1.28) = 0.10$.

Example: Define $Y \sim \text{Binom}(10, 0.3)$. Then the 90th quantile is the value 5. This is because $P(Y \leq 5) = 0.95 \geq 0.90$ and $P(Y < 5) = 0.85 \leq 0.90$.

We can also define a **Quantile function** for a random variable X that takes as input values of $p \in [0, 1]$ and returns the value x such that

$$P(X < x) \leq p \tag{3.39}$$
$$P(X \leq x) \geq p \tag{3.40}$$

Often we can think of the quantile function as the inverse of the cumulative mass (density) function. For a random variable Y, the cmf or cdf takes as input values in the support of Y and returns the probability that $P(Y \leq y)$.

Instead, the Quantile function takes as input a probability and returns the value such that $P(Y \leq y) \geq p$ and $P(Y < y) \leq p$.

3.5 Exercises

1. Let $X \sim \text{Bernoulli}(0.2)$
   (a) P(X=0) = ?
   (b) P(X=1) = ?
   (c) Please compute $\mathbb{E}(X)$
   (d) Please compute $V(X)$
   (e) Define the $supp(X)$
2. Let $Y \sim \text{Bernoulli}(\theta)$, and show that $P(Y = 1) = \mathbb{E}(Y)$
3. Let $Y \sim \text{Bernoulli}(\theta)$, and show that $V(Y) \leq \mathbb{E}(Y)$
4. Let $Y \sim \text{Bernoulli}(\theta)$, and let Z be the following function of Y:

$$Z(y) = \begin{cases} 1 & \text{if } y = 0 \\ 0 & \text{if } y = 1 \end{cases} \tag{3.41}$$

   What probability distribution does Z follow and why?
5. Design an experiment (short description) and define a random variable Y that may follow a geometric distribution. In the context of your experiment, how would you communicate $\mathbb{E}(Y)$ to another without statistical expertise?
6. Define a random variable R with a binomial distribution ($R \sim \text{Bin}(10, 0.2)$).
   (a) Compute $\mathbb{E}(R)$
   (b) Compute $V(R)$
   (c) Describe to someone who may not have statistical expertise what $P(R = 3)$ means? Be sure to include assumptions about the Binomial distribution and how the parameters N, θ relate to this probability.
   (d) For what value of θ is $V(R)$ the highest? Why does this make sense intuitively?
7. Suppose $Y \sim \text{Pois}(2)$
   (a) Compute $P(Y = 2)$

   (b) Compute $P(Y \leq 2)$

   (c) Compute $P(Y > 2)$

8. $X \sim \mathcal{N}(\mu, \sigma^2)$

   (a) Let $Y = X - \mu$. What is the distribution of Y?

   (b) Let $Z = Y/\sigma$. What is the distribution of Z?

   (c) Compute $P(Z = 0)$

9. Suppose we define a new random variable W with support $supp(W) = [0, 1]$ and probability density function

$$f_W(w) = 2w \tag{3.42}$$

   (a) Compute $P(W < 1/2) = \int_0^{1/2} f_W(w)\, dw$

   (b) Compute $P(W < 1) = \int_0^{1} f_W(w)\, dw$

10. Let X be a continuous random variable. Is $P(X \leq x) = P(X < x)$? Why or why not?

Part II

Statistics

4

Law of large numbers, Method of Moments, and the Central limit theorem

thomas mcandrew

Lehigh University

CONTENTS

4.1 Introduction

The law of large numbers (LLN) and the central limit theorem underpin a large portion of statistical theory. We will see that the LLN will link proportions to probabilities, expected values to the sample mean, and allow us to estimate parameters of a random variable given a dataset.

The method of moments will be the primary tool we use to derive parameter estimates for parameters for a given distribution of a random variable. This method follows directly from the LLN.

Finally, we will explore the Central limit theorem which is the foundation of much of generating hypothesis tests and building confidence intervals.

4.2 Independent and identically distributed and sample

To model a dataset, $\mathcal{D}$, we will need to make assumptions about how our data set was generated.

A common assumption for a dataset with n data points $\mathcal{D} = (x_1, x_2, \cdots, x_n)$ is that each data point was generated from a random variable $x_1 \sim X_1, x_2 \sim X_2, \cdots, x_n \sim X_n$. The collection of random variables $X_1, X_2, \cdots, X_n)$ is called a **sample** and the dataset $\mathcal{D}$ above is typically called one **realized sample** or a **realization** of the sample.

If we additionally assume that all the random variables are **identically distributed**, or that they have the same distribution $X_1, X_2, \cdots, X_n \sim f$, and we assume that for any i and j that X_i and X_j are independent then we call our sample a **random sample**.

Because pairwise independence between random variables and that all random variables are identically distributed are two common assumptions, they are abbreviated as **i.i.d** which stands for "independent and identically distributed".

4.3 Properties of the Expected value and Variance of a sum

To explore the law of large numbers, we will need another property of the expectation. The expected value of the sum of random variables is the sum of the expected value of each individual random variable.

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) \tag{4.1}$$

When we combine the properties of expected values we have already learned, we can state

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y) \tag{4.2}$$

and this is true for any number of random variables

$$\mathbb{E}\left(\sum_{i=1}^{N} a_i X_i \right) = a_i \sum_{i=1}^{N} \mathbb{E}(X_i) \tag{4.3}$$

We start at the definition of the expected value to derive this property

$$\begin{aligned}
\mathbb{E}(X+Y) &= \sum_x \sum_y (X+Y) f_{X+Y}(x,y) && (4.4)\\
&= \sum_x \sum_y X f_{X+Y}(x,y) + Y f_{X+Y}(x,y) && (4.5)\\
&= \sum_x \sum_y X f_{X+Y}(x,y) + \sum_x \sum_y Y f_{X+Y}(x,y) && (4.6)\\
&= \sum_x X f_X(x) + \sum_y Y f_Y(y) && (4.7)\\
&= \mathbb{E}(X) + \mathbb{E}(Y) && (4.8)
\end{aligned}$$

Example: Suppose $X \sim \text{Geom}(1/4)$ and $Y \sim \text{Bernoulli}(1/2)$ then $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) = \frac{1}{1/4} + \frac{1}{2} = 4\frac{1}{2}$.

The variance is an expected value and so it follows that

$$V(X+Y) = V(X) + V(Y) \tag{4.9}$$

and we can state more generally that for two independent random variables

$$V(aX+bY) = a^2 V(X) + b^2 V(Y) \tag{4.10}$$

Example: Let $X \sim \text{Binomial}(N, \theta)$ and $Y \sim \text{Binomial}(M, \gamma)$ then $V(X - Y) = V(X + (-1)Y) = V(X) + (-1)^2 V(Y) = N\theta(1-\theta) + M\gamma(1-\gamma)$.

4.4 Law of large numbers

The intuition behind the law of large numbers can be characterized by the following experiment: you are asked to flip a fair coin and record the whether the coin is heads up or tails up. After 10 flips you are asked to compute the proportion of heads up flips, after 50 flips you are asked to compute this proportion, after 100, 1,000, 10,000 flips you are asked to compute the proportion of heads up flips. We expect that if this coin is fair that the proportion of flips with heads face up will get closer and closer to 0.50.

The law of large numbers attempts to describe this intuition.

Define a sequence of random variables $X_1, X_2, \cdots, X_n$ such that any pair of random variables, X_i and X_j are independent (that is $P(X_i = x|X_j = y) = P(X_i = x)$. Finally, transform this sequence into a single random variable $\overline{X}$ that is equal to $\overline{X} = \dfrac{X_1 + X_2 + X_3 + \cdots X_N}{N}$.

Then the law of large numbers (LLN) states that given any small number ϵ that is greater than 0 as n grows towards infinity $(n \to \infty)$

$$P(|\overline{X}_n - \mu| > \epsilon) \to 0 \tag{4.11}$$

where $\mu = \mathbb{E}(\overline{X})$.

We can picture a distribution $Z_n = |\overline{X}_n - \mu|$ that depends on n and as n increase the random variable Z_n assigns more and more probability to the value 0. <pic>

The LLN has many applications.

Proportions approximate probabilities

Let a sequence of Bernoulli distributed random variables be (i) pairwise independent and (ii) be distributed the same $X_1, X_2, \cdots, X_N \sim \text{Bern}(\theta)$.

Lets take a look at the transformation $\overline{X}$. Consider a sequence of two random variables with a Bernoulli distribution. Then define

$$S = X_1 + X_2 \tag{4.12}$$

The expected value of S is

$$\begin{align} \mathbb{E}(S) &= \mathbb{E}(X_1 + X_2) \tag{4.13}\\ &= \mathbb{E}(X_1) + \mathbb{E}(X_2) \tag{4.14}\\ &= \theta + \theta \tag{4.15}\\ &= 2\theta \tag{4.16} \end{align}$$

and the expected value of $\overline{X}$ is then

$$\mathbb{E}(\overline{X}) = \frac{\mathbb{E}(S)}{2} = \frac{2\theta}{2} = \theta \tag{4.17}$$

To gain intuition about what a sample from $\overline{X}$ looks like, assume we collect data points from n random variables $x_1 \sim X_1, x_2 \sim X_2, \cdots, x_n \sim X_n$. An example of a single data point d could be

$$d = (0, 1, 0, 1, 0, 0, 1, 1, 1, 0) \tag{4.18}$$

and the sample mean $\overline{x}$, one sample from $\overline{X}$, is then

$$S = \sum_{i=1}^{10} d_i \tag{4.19}$$

$$\overline{x} = \frac{X}{n} \tag{4.20}$$

However, the value of S can be counted in two different ways. We can sum up d_i one by one in order, or we can multiply zero by the frequency of zeros in our dataset, multiply 1 by the frequency of ones in our dataset, and sum both of these products.

$$S = \sum_{i=1}^{10} d_i \tag{4.21}$$

$$= 0 + 1 + 0 + 1 + 0 + 0 + 1 + 1 + 1 + 0 \tag{4.22}$$

$$S = N(0) \cdot 0 + N(1) \cdot 1 \tag{4.23}$$

where $N(x)$ is the number of values x in our dataset.

If we use the above method to sum our data points the sample mean simplifies to

$$\overline{x} = \frac{N(0) \cdot 0 + N(1) \cdot 1}{N} \tag{4.24}$$

$$= \frac{N(1)}{N}. \tag{4.25}$$

The sample mean is the proportion of 1s (p_n) from our sample of n data points, and the law of large number then says that

$$P(|\overline{X} - \mu| > \epsilon) \to 0 \tag{4.26}$$

$$P(|p_n - \theta| > \epsilon) \to 0 \tag{4.27}$$

or, roughly, that the sample proportion approaches θ, the true probability that a 1 will appear.

The Relationship between the expected value and the sample mean

Define a discrete random variable $X_1, X_2, \cdots, X_n \sim f_X$, $supp(X_i) = \{1, 2, 3\}$

Lets assume $n = 10$ and take a closer look at S. One sample from our $n = 10$ random variables could be $d = (2, 1, 3, 2, 3, 2, 3, 1, 2, 2)$ and we can compute the sum $S = 2 + 1 + 3 + 2 + 3 + 2 + 3 + 1 + 2 + 2$. However, we could have also summed these ten numbers by multiplying each unique number in the dataset by the frequency this number occurs and then summing these products: $S = N(1) \cdot 1 + N(2) \cdot 2 + N(3) \cdot 3$, where $N(x)$ is the number of times the value x appears. For our specific example $S = 2 \cdot 1 + 5 \cdot 2 + 3 \cdot 3$. The sample mean is then

$$\overline{x} = \frac{S}{N} \tag{4.28}$$

$$= \frac{N(1) \cdot 1 + N(2) \cdot 2 + N(3) \cdot 3}{N} \tag{4.29}$$

$$= \frac{N(1)}{N}1 + \frac{N(2)}{N}2 + \frac{N(3)}{N}3 \tag{4.30}$$

We know from our discussion of the LLN and Bernoulli distributed random variables that $\frac{N(1)}{N}$ is the proportion of 1s , $p_n(1)$, and this proportion will approach the true probability of a 1 as we collect more data (as $n \to \infty$).

Let's rewrite the above sample mean computation

$$\overline{x} = \frac{N(1)}{N}1 + \frac{N(2)}{N}2 + \frac{N(3)}{N}3 \tag{4.31}$$

$$= p_n(1)1 + p_n(2)2 + p_n(3)3 \tag{4.32}$$

$$\tag{4.33}$$

As $n \to \infty$

$$\overline{x} = \frac{N(1)}{N}1 + \frac{N(2)}{N}2 + \frac{N(3)}{N}3 \tag{4.34}$$

$$= f_X(1)1 + f_X(2)2 + f_X(3)3 \tag{4.35}$$

but the above is the expected value of X, $\mathbb{E}(X)$, and so as n gets larger the sample mean will become closer to the true expected value.

4.4.1 LLN for transformations

Earlier we saw that by definition if we define a random variable X and a function g then

$$\mathbb{E}\left[g(X)\right] = \sum_{i \in supp(X)} g(x_i) f_X(x_i) \tag{4.36}$$

Because the above is equivalent to the expectation of a transformed random variable $Y = g(X)$ the LLN has a similar result for any transformation of a random variable.

As $n \to \infty$,

$$P(|\overline{Y}_n - \mu_Y| > \epsilon) \to 0 \tag{4.37}$$

where $\mu_Y = \mathbb{E}(Y) = \mathbb{E}\left[g(X)\right]$.

The above implies that the LLN applies to any transformation of a random variable, including

$$S = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} \tag{4.38}$$

where $\mu = \mathbb{E}[g(X)]$.

Let us find a simple expression for $\mathbb{E}(S)$.

$$\begin{aligned}
\mathbb{E}(S) &= \mathbb{E}\left[\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}\right] && (4.39)\\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[(X_i - \mu)^2\right] && (4.40)\\
&= \frac{1}{N}\sum_{i=1}^{N}V(X) && (4.41)\\
&= \frac{1}{N}NV(X) = V(X) && (4.42)
\end{aligned}$$

Because $\mathbb{E}(S) = V(X)$, the LLN states that for any $\epsilon > 0$ as $n \to \infty$

$$P(|S - V(X)| > \epsilon) \to 0 \tag{4.43}$$

or that S approaches the true variance of X.

4.5 Statistics, Estimators, and the Method of Moments

Given a random sample $(X_1, X_2, X_3, \cdots, X_n)$ or collection of random variables that are i.i.d, a **statistic** is a function of our sample

$T(X_1, X_2, X_3, \cdots, X_n)$. Because a statistic is a function of a collection of random variables, our statistic too can be characterized by a probability distribution over potential values of the statistic.

For a random sample $(X_1, X_2, X_3, \cdots, X_n)$ all with the same probability density function $f_X(\theta)$ where θ is a parameter, an *estimator*, $T(X_1, X_2, \cdots, X_n)$ of θ is (i) a statistic that (ii) is meant get closer to θ as $n \to \infty$.

An **estimate** is an estimator applied to a realized sample, and an estimate for a parameter is often denoted as the parameter with a "hat". For example, an estimate of the parameter β would bve denoted $\hat{\beta}$.

One way to develop an estimator for a parameter, or for several parameters, is to use the **Method of Moments**. For a random sample $(X_1, X_2, \cdots, X_n)$ where each random variable follows the same probability density function with k parameters $X_i \sim f_X(x|\theta_1, \theta_2, \theta_k)$, the Method of moments generates estimators for all k parameters by solving this set of equations

$$\mathbb{E}(X) = g_1(\theta_1, \theta_2, \cdots, \theta_k) \tag{4.44}$$

$$\mathbb{E}(X^2) = g_2(\theta_1, \theta_2, \cdots, \theta_k) \tag{4.45}$$

$$\vdots \tag{4.46}$$

$$\mathbb{E}(X^k) = g_k(\theta_1, \theta_2, \cdots, \theta_k), \tag{4.47}$$

replacing each exact, expected value by sample means. The expected value $\mathbb{E}(X)$ is replaced by $\overline{X} = \sum_{i=1}^{N} X_i/N$, the expected value $\mathbb{E}(X^2)$ is replaced by its sample mean $\overline{X^2} = \sum_{i=1}^{N} X_i^2/N$, etc.

Lets look at an example.

MoM estimator for the Bernoulli

Suppose we are studying the probability of survival among patients who are treated with dialysis and experienced a myocardial infarction (see `10.1056/NEJM199809173391203` for a manuscript on this topic). We collect data on 50 patients and model whether they survive or not as a random sample $(X_1, X_2, X_3, \cdots, X_{50})$ where $X_i \sim \text{Bern}(\theta)$ and the value 1 represents survival and zero otherwise. We follow patients for a year and record a dataset $\mathcal{D} = (1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, \cdots)$ that contains 36 ones and 14 zeros.

Because the Bernoulli distribution has a single parameter we only need a single equation that related the expected value and our parameter θ. Because the Bernoulli distribution has a single parameter we only need a single equation that related the expected value and our parameter θ.

The expected value is $E(X) = \theta$ and so we can derive a MoM estimator by replacing the exact $\mathbb{E}(X)$ with the sample mean $\overline{X}$. Our estimator for θ is $\hat{\theta} = \overline{X} = \sum_{i=1}^{N} x_i/N = 36/50 = 72\%$.

MoM estimator for the Uniform Continuous

Suppose $(Y_1, Y_2, \cdots, Y_{10})$ is a random sample such that $Y_i \sim U(\alpha, \beta)$. We collect a dataset (realized sample) $\mathcal{D} = (0.22, 0.45, 0.12, 0.32, 0.56, 0.73, 0.91, 0.51, 0.11, 0.67)$. To develop an estimate for α and an estimate for β using the method of moments, we will need to look at the following two equations:

$$\mathbb{E}(X) = \frac{\alpha + \beta}{2} \tag{4.48}$$

$$\mathbb{E}(X^2) = g(\alpha, \beta) \tag{4.49}$$

On first glance, the second equation sounds difficult to solve. However, we found an expression for the variance that depends on $\mathbb{E}(X^2)$ and on $\mathbb{E}(X)$ in chapter 2, exercise 9.

$$V(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \tag{4.50}$$

$$\mathbb{E}(X^2) = V(X) + [\mathbb{E}(X)]^2 \tag{4.51}$$

We know that $V(X) = \frac{(b-a)^2}{12}$ and so

$$\begin{aligned}
\mathbb{E}(X^2) &= V(X) + [\mathbb{E}(X)]^2 && (4.52)\\
&= \frac{(\beta - a)^2}{12} + \left[\frac{(\alpha + \beta)}{2}\right]^2 && (4.53)\\
&= \frac{\beta^2 + \alpha^2 - 2\alpha\beta}{12} + \frac{\alpha^2 + \beta^2 + 2\alpha\beta}{4} && (4.54)\\
&= \frac{\beta^2 + \alpha^2 + \alpha\beta}{3} && (4.55)
\end{aligned}$$

Now our MoM equations are

$$\mathbb{E}(X) = \frac{\alpha + \beta}{2} \tag{4.56}$$

$$\mathbb{E}(X^2) = \frac{\beta^2 + \alpha^2 + \alpha\beta}{3} \tag{4.57}$$

and we will replace $\mathbb{E}(X)$ with the sample mean $\overline{X}$ and replace $\mathbb{E}(X^2)$ with $\overline{X^2}$, and then solve for α and β.

$$\overline{X} = \frac{\alpha + \beta}{2} \tag{4.58}$$

$$\overline{X^2} = \frac{\beta^2 + \alpha^2 + \alpha\beta}{3} \tag{4.59}$$

From the first equation we find $\alpha = 2\overline{X} - \beta$. We can plug the above into the second equation to find

$$\overline{X^2} = \frac{\beta^2 + (2\overline{X} - \beta)^2 + (2\overline{X} - \beta)\beta}{3} \tag{4.60}$$

$$3\overline{X^2} = \beta^2 + 4\overline{X}^2 + \beta^2 - 4\overline{X}\beta + 2\overline{X}\beta - \beta^2 \tag{4.61}$$

$$3\overline{X^2} = \beta^2 + 4\overline{X}^2 - 2\overline{X}\beta \tag{4.62}$$

$$0 = \beta^2 - (2\overline{X})\beta + (4\overline{X}^2 - 3\overline{X^2}) \tag{4.63}$$

$$\hat{\beta} = 2\overline{X} \pm \frac{\sqrt{(2\overline{X})^2 - 4(1)(4\overline{X}^2 - 3\overline{X^2})}}{2} \tag{4.64}$$

$$\hat{\beta} = 2\overline{X} \pm \frac{\sqrt{4\overline{X}^2 - (16\overline{X}^2 - 12\overline{X^2})}}{2} \tag{4.65}$$

$$\hat{\beta} = 2\overline{X} \pm \frac{\sqrt{-12\overline{X}^2 + 12\overline{X^2}}}{2} \tag{4.66}$$

$$\hat{\beta} = 2\overline{X} \pm \frac{2\sqrt{3}\sqrt{\overline{X^2} - \overline{X}^2}}{2} \tag{4.67}$$

$$\hat{\beta} = 2\overline{X} \pm \sqrt{3}\sqrt{\overline{X^2} - \overline{X}^2} \tag{4.68}$$

$$\tag{4.69}$$

and so for our problem above $\hat{\beta} = 1.37$ and so $\alpha = -0.45$.

MoM estimator for the Poisson

Suppose $(R_1, R_2, \cdots, R_{5))}$ is a random sample such that $R_i \sim \text{Pois}(\lambda)$, and we are given the following dataset $\mathcal{D} = (3, 7, 10, 12, 45)$.

We can develop an estimator for λ by using the LLN and MoM:

$$\frac{1}{n}(R_1 + R_2 + \cdots + R_n) \to \mathbb{E}(R) \tag{4.70}$$

as $n \to \infty$.

The MoM of moments estimator is

$$\begin{aligned} \frac{1}{n}(r_1 + r_2 + \cdots + r_n) &\approx \mathbb{E}(R) && (4.71)\\ \frac{1}{n}(r_1 + r_2 + \cdots + r_n) &= \lambda && (4.72)\\ \hat{\lambda} &= \frac{1}{n}\sum_{i=1}^{n} r_i && (4.73) \end{aligned}$$

as $n \to \infty$.

In our example, our estimate for the true λ would be

$$\hat{\lambda} = \frac{1}{5}(3 + 7 + 10 + 12 + 45) = 15.4 \tag{4.74}$$

We can compute an estimated expected value, variance, estimated probabilities, etc by replacing the our true parameter value λ with our estimate $\hat{\lambda}$. For example, an estimate for the expected value is

$$\mathbb{E}(\hat{R}) = \hat{\lambda} = 15.4, \tag{4.75}$$

an estimated probability mass function

$$f_R(r; \hat{\lambda}) = e^{-\hat{\lambda}}\frac{\hat{\lambda}^r}{r!} \tag{4.76}$$

4.6 Central limit theorem

Given a random sample $(X_1, X_2, \cdots, X_n)$, we used the LLN and method of moments to produce point estimates—a single number—for a parameter that assigned probabilities to value of X_i. However the LLN only guarantees that our estimator will grow closer to the true parameter value as $n \to \infty$. The LLN does not describe how close we are to the true parameter value for any specific n.

The Central limit theorem (CLT) describes the statistical distribution of the sample mean $\bar{X}$, allowing us to describe how close we are to the true parameter value of interest.

For a random sample $(X_1, X_2, \cdots, X_n)$ the **Central Limit Theorem** states

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \tag{4.77}$$

or that

$$\sum_{i=1}^{n} X_n \sim \mathcal{N}\left(n\mu, n\sigma^2\right) \tag{4.78}$$

where $\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$, $\mu = \mathbb{E}(X)$ and $\sigma^2 = V(X)$.

CLT for the Bernoulli

Suppose we wish to study a random sample of n Bernoulli distributed random variables $(Y_1, Y_2, \cdots, Y_n) \sim \text{Bern}(\theta)$.

We know that $\mathbb{E}(Y) = \theta$ and $V(Y) = \theta(1-\theta)$, and so

$$\overline{Y}_n \sim \mathcal{N}\left(\theta, \frac{\theta(1-\theta)}{n}\right) \tag{4.79}$$

We saw in the previous section on the LLN and MoM that an estimate for θ is the sample mean so then

$$\hat{\theta} \sim \mathcal{N}\left(\theta, \frac{\theta(1-\theta)}{n}\right) \tag{4.80}$$

However, we still cannot compute probabilities that $\hat{\theta}$ is in some interval because the Normal distribution on the right hand side involves the true, unknowable, θ.

$$\hat{\theta}(Y) - \theta \sim \mathcal{N}\left(0, \frac{\theta(1-\theta)}{n}\right) \tag{4.81}$$

here we introduce the notation $\hat{\theta}(Y)$ to reinforce that this is a random variable and not a single value.

To approximate the difference $\hat{\theta}(Y) - \theta$ we can replace θ with our MoM estimate $\overline{y}$.

$$\hat{\theta}(Y) - \theta \sim \mathcal{N}\left(0, \frac{\overline{Y}_n(1-\overline{Y}_n)}{n}\right) \tag{4.82}$$

We can use the above to approximate the probability that the difference between our estimated $\hat{\theta}$ and the true θ is inside some interval. Suppose we collect the data $\mathcal{D} = (1,1,0,1,0,1,1,1,0,0,1,1,0,1,0)$. Then the CLT says

$$\hat{\theta}(Y) - \theta \sim \mathcal{N}\left(0, \frac{\overline{y}_n(1-\overline{y}_n)}{n}\right) \tag{4.83}$$

$$\hat{\theta}(Y) - \theta \sim \mathcal{N}\left(0, \frac{\frac{9}{15}\cdot\frac{6}{15}}{15}\right) \tag{4.84}$$

$$\tag{4.85}$$

and we can compute for example the probability that our estimate is 0.1 away from the truth.

$$P(-0.1 < \hat{\theta}(Y) - \theta < 0.1) = F(0.1) - F(-0.1) \approx 0.57 \tag{4.86}$$

In other words, there is a 0.57 probability that our estimate is within 0.1 of the true parameter value θ.

4.7 Confidence intervals: the CLT in action

Given a random sample $(X_1, X_2, \cdots, X_n,)$, a **confidence interval** for a parameter θ is a pair of statistics $(L(X), U(X))$ such that $P(\theta \in [L(X), U(X)]) =$

$1-\alpha$. The number α is called the **confidence coefficient** and we often say that the interval $[L(X), U(X)]$ is a $1-\alpha$ confidence interval for the parameter θ.

A confidence interval provides a range of values that the true parameter is likely to be between.

Example: Suppose we decide to study the prevalence of tuberculosis among people who have diabetes mellitus. Our model supposes we can represent each person in our sample with a random variable $X \sim \text{Bern}(\theta)$ where θ is the probability of an active tuberculosis infection. Given a random sample $(X_1, X2, \cdots, X_n)$ where $X_i \sim \text{Bern}(\theta)$, we find 34 people had an active infection and 3,456 did not. If we generate a 95% confidence interval for the parameter θ that is [0.05%, 2%] then there are tow claims. Claim one: The confidence interval applied to a random sample that follows all of our assumptions will have a 95% probability that the parameter value is within our confidence interval. Claim two: the confidence interval that was produced from our data either does or does not contain the true parameter value, and we are 95% confident that this interval contains the true parameter value.

4.7.1 Z values

Suppose we assumed a random sample generated an observed dataset $\mathcal{D}$

$$(X_1, X_2, \cdots, X_n) \tag{4.87}$$

$$X_i \sim f \tag{4.88}$$

The following transformation

$$Z = \frac{X - \mathbb{E}(X)}{\sqrt{V(X)}} \tag{4.89}$$

is called **standardizing** a random variable.

We can show that a standardized variable has an expected value of zero and variance of one using expected value and variance properties.

$$\mathbb{E}(Z) = \mathbb{E}\left(\frac{1}{V(X)}\left[X - \mathbb{E}(X)\right]\right) \tag{4.90}$$

$$= \frac{1}{\sqrt{V(X)}}\left[\mathbb{E}(X) - \mathbb{E}(X)\right] \tag{4.91}$$

$$= 0 \tag{4.92}$$

The expected value of our standardized random variable Z is zero.

$$V(Z) = V\left(\frac{1}{\sqrt{V(X)}}\left[X - \mathbb{E}(X)\right]\right) \tag{4.93}$$

$$= \frac{1}{V(X)}\left[V(X) - 0\right] \tag{4.94}$$

$$= 1 \tag{4.95}$$

The variance of our standardized random variable Z is one.

In practice, we can transform our individual data points $x_1, x_2, \cdots, x_n$ into standard, or "z" scores by subtracting from each point x_i the sample mean and sample standard deviation.

Example: Suppose we are given a dataset $\mathcal{D} = (1.05, 4.20, -2.80, 1.34)$. To transform these four points into four z-scores we need to follow three steps: (i) compute the sample mean, (ii) compute the sample standard deviation, and (iii) subtract from each point the sample mean and divide this difference by the sample standard deviation.

The sample mean is

$$\overline{x} = (1.05 + 4.20 - 2.80 + 1.34)/4 = 0.95$$

.

The sample standard deviation is

$$s = \left[(1.05 - 0.95)^2 + (4.20 - 0.95)^2 + (2.80 - 0.95)^2 + (1.34 - 0.95)^2\right]/4 = 2.49$$

.

Our four z-scores are then

$$z = ((1.05 - 0.95)/2.49, (4.20 - 0.95)/2.49, (-2.80 - 0.95)/2.49, (1.34 - 0.95)/2.49)$$

Z-scores are useful for characterizing the probability of sample points in a dataset and for eliminating measurement units from a set of sample points. The z score for a sample point z_i

$$z_i = \frac{x_i - \mathbb{E}(X)}{\sqrt{V(X)}} \tag{4.96}$$

can be interpreted as the number of standard deviations the sample point x_i is from the expected value of the random variable X_i.

Chebychev's inequality tells us that values generated by a random variable X are likely close to the expected value, and so large z-scores are improbable while z-scores close to zero are probable. A z-score is meant to map values to the probability that this observation occurred. Large z values should denote a small probability that this event occurred and small z values should denote a large probability that this event occurred.

Z-scores also eliminate measurement units (u) from an observation by mapping u to units of standard deviation. Mapping the units of two or more variables to the number of standard deviations from their mean is a convenient way to compare two variables (i) where one variable has large values and the second has small values (ii) where one variable has a large variance and the second has a small variance. Large, small values and large, small variances associated with a variable are often because of the chosen units.

4.7.2 Using the CLT to build a confidence interval

The method of moments (MoM) has given us an algorithm for estimating parameter values, however the MoM only provides a single estimate. Because we compute our estimated parameter values from a sample of data, we know that our estimate is not exact and will vary for every data set we collect.

A confidence interval for a parameter, or parameters, is a set of statistics—a lower bound (L) and upper bound (U)—where we expect the true parameter values to lie with high probability. Confidence intervals take into account the size of our sample dataset, the variability in our estimate of parameter values, and should accompany a point estimate for a parameter.

Example:

More formally, given a random sample $(X_1, X_2, \cdots, X_n)$, a confidence interval for the parameter θ is a pair of statistics $[L(X), U(X)]$ such that $P(\theta \in [L(X), U(X)]) = 1 - \alpha$. The value α is called the confidence coefficient and the interval $[L(X), U(X)]$ is called a $1 - \alpha$ confidence interval.

One way to build a confidence interval is to use the Central Limit Theorem. The CLT states that

$$\overline{X} \sim \mathcal{N}\left(\mu, \sigma^2\right) \tag{4.97}$$

where μ is $\mathbb{E}(X)$ and σ^2 is $V(X)$.

Suppose $(X_1, X_2, \cdots, X_n)$. The CLT states that

$$\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n) \tag{4.98}$$

We can subtract from our random variable $\overline{X}$ the expected value μ

$$\overline{X} - \mu \sim \mathcal{N}(0, \sigma^2/n) \tag{4.99}$$

and divide by $\frac{\sigma}{\sqrt{n}}$ to find that

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1) \tag{4.100}$$

$$Z = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim \mathcal{N}(0, 1) \tag{4.101}$$

where $\mu = \mathbb{E}(X)$ and $\sigma^2 = V(X)$.

The translated and scaled quantity above, a new random variable Z, has a standard normal distribution. A random variable Z has a **standard normal** distribution if $Z \sim \mathcal{N}(0, 1)$. That is, a normal distribution with expected value zero and variance one.

One way to form a $1 - \alpha$ confidence for a parameter μ is to find two values a and b such that $P(a < Z < b) = 1 - \alpha$ and then solve for μ.

Lets suppose we found these two values a and b

$$P(a < Z < b) = 1 - \alpha \tag{4.102}$$

$$P\left(a < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < b\right) = 1 - \alpha \tag{4.103}$$

$$P\left(a \cdot \frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < b \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{4.104}$$

$$P\left(-a \cdot \frac{\sigma}{\sqrt{n}} > \mu - \overline{X} > -b \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{4.105}$$

$$P\left(\overline{X} - a \cdot \frac{\sigma}{\sqrt{n}} > \mu > \overline{X} - b \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{4.106}$$

The normal distribution is symmetric about 0 and so if $P(Z > a) = \alpha$ then $P(Z < -a) = \alpha$. To bound Z such that $P(a < Z < b) = 1 - \alpha$ we will split $1 - \alpha$ evenly.

Suppose we find values a and b such that

$$P(Z < a) = \alpha/2 \tag{4.107}$$

$$P(Z > b) = \alpha/2 \tag{4.108}$$

$$\tag{4.109}$$

then

$$P([Z < a] \cup [a < Z < b] \cup [Z > b]) = 1 \tag{4.110}$$
$$P(Z < a) + P(a < Z < b) + P(Z > b) = 1 \tag{4.111}$$
$$P(a < Z < b) = 1 - P(Z < a) - P(Z > b) \tag{4.112}$$
$$P(a < Z < b) = 1 - \alpha/2 - \alpha/2 \tag{4.113}$$
$$P(a < Z < b) = 1 - \alpha \tag{4.114}$$

We define the symbol z_α as the value such that the probability that a random variable Z with a standard normal distribution, $\mathcal{N}(0,1)$, is less than z_α equals α.

$$P(Z < z_\alpha) = \alpha. \tag{4.115}$$

So then

$$P\left(\overline{X} - a \cdot \frac{\sigma}{\sqrt{n}} > \mu > \overline{X} - b \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{4.116}$$
$$P\left(\overline{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} > \mu > \overline{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{4.117}$$
$$P\left(\overline{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} > \mu > \overline{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{4.118}$$
$$P\left(\overline{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{4.119}$$

Often a $1 - \alpha$ confidence for μ is written

$$\overline{X} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{4.120}$$

Example: Suppose $(Y_1, Y_2, \cdots, Y_n)$ and $Y_i \sim \text{Pois}(\lambda)$. The variance for Y_i, $\sigma^2 = V(Y)$ is λ. Then a $1 - \alpha$ confidence interval for λ is $\overline{Y} \pm z_{1-\alpha/2}\sqrt{\frac{\lambda}{n}}$. Given a dataset $\mathcal{D} = (2, 5, 3, 8, 10)$ we can estimate $\hat{\lambda} = \overline{y}$ and so a sample confidence interval is $\overline{y} \pm z_{1-\alpha/2}\sqrt{\frac{\overline{y}}{n}} = 5.6 \pm z_{1-\alpha/2}\sqrt{\frac{5.6}{5}}$

Below is a table of common z values

A 95% confidence interval will use the z values -1.96 and 1.96, a 90% confidence interval will use the z values -1.64 and 1.64, and a 80% confidence interval will use the values -1.28 and 1.28.

4.7.3 Hypothesis Testing

The goal of statistical hypothesis testing is to support or deny a claim about a parameter. We first generate a set of statements about a parameter—called a

α	z_α
0.025	-1.96
0.050	-1.64
0.100	-1.28
0.500	0
0.900	1.28
0.950	1.64
0.975	1.96

hypotheses—that can we test. Second, we collect data that will help us make a decision about whether one statement or the other is true. Third, we decide which statement can be refuted.

A **hypothesis** is typically a pair of statements about a parameter. The **null hypothesis (**H_0**)** is a statement about a parameter value that allows us to characterize data generated by that parameter.

Example: Suppose the following random sample $(X_1, X_2, \cdots, X_n)$ X_i Bern(θ). Then a null hypothesis might be $\theta = 0.50$. This statement about a parameter would allow us to characterize samples of data $(x_1, x_2, \cdots, x_n)$ we may see from our random sample $(X_1, X_2, \cdots, X_n)$.

The **alternative hypothesis (**H_1**)** is a statement about ta parameter value that is contrary to the null hypothesis.

Example: Suppose the sample sample and distribution as in the above example. Then an alternative hypothesis may be $\theta > 0.50$. This statement about a parameter would also allow us to characterize samples of data $(x_1, x_2, \cdots, x_n)$ we may see from our random sample $(X_1, X_2, \cdots, X_n)$.

We will say that a **hypothesis** must contain a null and alternative statement.

Example: A hypothesis may be $H_0 : \theta = 0.50$ vs $H_1 : \theta > 0.50$

Though we plan to collect data and create a test that will support or deny the above hypothesis, we may make a mistake. We will never know the true parameter value from any finite given random sample and so our test will be imperfect. Mistakes in hypothesis testing are categorized into two types: types I and II. A **type I error** (informally called a false positive) is when we decide to accept the alternative hypothesis (H_1) however the null hypothesis is the truth. A **type II error** (informally called a false negative) is when we decide to accept the null hypothesis however the alternative hypothesis is true. Related to type I and II errors is the **power** of a hypothesis. The **power** of a hypothesis test is the probability that we decide on the alternative hypothesis given that the alternative hypothesis is true.

Define a random sample $(X_1, X_2, X_3, \cdots, X_n)$ and X_i f. We define the **space**, $\mathcal{S}$, of our random sample as $supp(X_1) \times supp(X_2) \times \cdots \times supp(X_n)$. The space of a random sample is the set of all possible tuples that our random sample could generate, or all possible data sets.

A **hypothesis test** is a decision criteria such that we reject the null hypothesis if a realization $(x_1, x_2, \cdots, x_n)$ from our random sample $(X_1, X_2, \cdots, X_n)$ satisfies some criteria. The set C of all possible realizations that lead us to reject the null hypothesis is called the **critical region**.

Example: Suppose $(X_1, X_2, \cdots, X_n)$ and $X_i \sim \mathcal{N}(\mu, 1)$. We can form the hypothesis $H_0 : \mu = 0$ vs $H_1 : \mu > 0$. Intuitively we may decide to collect realizations $\mathcal{D} = (x_1, x_2, \cdots, x_n)$, compute $\overline{x} = \sum_{i=1}^{n} x_i/n$, and decide to reject the null hypothesis if $\overline{x} > c$. The set of all samples $(x_1, x_2, \cdots, x_n)$ where $\overline{x} > c$ would be our **critical region**.

4.7.3.1 Z-test for Normally distributed data

A two-sided one-sample Z test supposes that given a random samples $(X_1, X_2, X_3, \cdots, X_n)$ the following null and alternative hypotheses

$$H_0 : \mu = \mu_0 \tag{4.121}$$

$$H_1 : \mu \neq \mu_0 \tag{4.122}$$

where μ_0 is a constant, we can form the following statistic—called a test statistic

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \tag{4.123}$$

A **test statistic** is a statistic that is used to help determine whether a sample is or is not in a pre-specified critical region.

If the H_0 was true, we would expect $\overline{X}$ to be close to the value μ_0, then our test statistic Z should be close to the value zero. When Z is far from zero then we would reject H_0 and suspect that H_1 is the more plausible conclusion.

Then we can define a critical region (C) as

$$C = \{(x_1, x_2, \cdots, x_n) |\ |Z| > c\} \tag{4.124}$$

The value c can be chosen by limiting our potential type I error, the error that we accept H_1 given that H_0 is true. If H_0 is true then the probability that we accept H_1 is all the possible samples such that

$$P(|Z| > c) \tag{4.125}$$

$$P(-c > Z > c) \tag{4.126}$$

or samples where Z is smaller than negative c and larger than positive c.

Lets limit our type I error to α so that we need to find a value c such that

$$P(-c > Z > c) \leq \alpha \tag{4.127}$$

We will split the probability α in half so that

$$P(Z > c) \leq \frac{\alpha}{2} \tag{4.128}$$

$$P(Z < -c) \leq \frac{\alpha}{2} \tag{4.129}$$

We know from the CLT that the random variable Z has a standard normal distribution if $\mathbb{E}(X) = \mu_0$, in other words, if H_0 is true. This means that we can compute the probability that

$$P(Z > c) = \frac{\alpha}{2}. \tag{4.130}$$

This is by definition the value $z_{1-\alpha/2}$.

Example: Suppose we collect a dataset $\mathcal{D} = (-0.43, 1.58, -2.37, -0.11, -0.15, 2.23, -0.87, -1.24, 1.55, 0.38)$. We assume that the data was generated from a random sample $(G_1, G_2, \cdots, G_{10})$ where $G_i \sim \mathcal{N}(\mu, \sigma^2)$. We wish to test the following hypotheses:

$$H_0 : \mu = 0 \tag{4.131}$$

$$H_1 : \mu > 0 \tag{4.132}$$

with the critical region $C = \{(g_1, g_2, \cdots, g_{10}) \, | Z > 2\}$

We first compute our Z sample statistic $Z = \frac{\bar{x}-\mu_0}{\sigma^2/\sqrt{n}} = \frac{0.057-0}{0.18/\sqrt{10}} = 1.01$. We used the MoM estimate for σ^2. Because our sample Z statistic is $1.01 < 2$ we cannot reject the null hypothesis.

4.7.3.2 Z-test for a Bernoulli sample

The Z-test can be applied to any random sample. For example, suppose we take a random sample $(Y_1, Y_2, \cdots, Y_n)$ where $Y_i \sim \text{Bern}(\theta)$.

We can build a Z test statistic for hypotheses of the form

$$H_0 : \theta = \theta_0 \tag{4.133}$$

$$H_1 : \theta \neq \theta_0 \tag{4.134}$$

as

$$Z = \frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} \tag{4.135}$$

$$= \frac{\overline{X} - \theta_0}{\sqrt{\hat{\theta}(1-\hat{\theta})/n}} \tag{4.136}$$

$$\tag{4.137}$$

where $\hat{\theta}$ is the MoM estimator for θ.

4.7.3.3 The p-value

\nobreak

4.8 Exercises

1. Suppose we collect the following dataset $D = (10, 2, 6, 7, 17, 3, 1, 1, 6, 5, 4, 1)$ and further we assume that each data point d_i was generated by sampling from a sequence of i.i.d random variables $d_1 \sim X_1, d_2 \sim X_2, \cdots, d_{12} \sim X_{12}$ where $X_i \sim \text{Geom}(p)$.

   (a) State the LLN for the above problem

   (b) Compute the sample mean of D

   (c) What can we say about the sample mean, using the LLN, if we collect 13,14,$\cdots$ data points?

2. Suppose we collect the following sample $\mathcal{D} = (10, 2, 6, 7, 17, 3, 1, 1, 6, 5, 4, 1)$ and further we assume that each data point d_i was generated by sampling from a sequence of i.i.d random variables $d_1 \sim X_1, d_2 \sim X_2, \cdots, d_{12} \sim X_{12}$ where $X_i \sim \text{Pois}(\lambda)$.

   (a) State the LLN for the above problem

   (b) Compute the sample variance

   (c) What can we say about the sample variance, using the LLN, if we collect 13,14,$\cdots$ data points?

3. Define independent and identically distributed random variables $B_1, B_2, B_3, B_4, B_5 \sim \text{Geom}(p)$. Suppose we sample these rvs and collect $d = (12, 9, 1, 5, 3)$.

   (a) Use the method of moments to develop an estimator for p.

   (b) Estimate p from the above sample $\mathcal{D}$.

(c) What is the estimated expected value?

(d) What is the estimated variance?

(e) What is the estimated $P(B_1 = 1)$

4. Assume $(Y_1, Y_2, Y_3, \cdots, Y_n)$ are a random sample where $Y_i \sim \mathcal{N}(\mu, \sigma^2)$. Further assume we collected the data set $\mathcal{D} = (1.30, -2.11, 1.44, 0.35, -0.40, 0.61, -1.06, -0.86, -0.60, 0.19)$.

   (a) Use the method of moments to develop an estimator for μ and for σ^2.

   (b) Estimate μ and σ^2 from the above sample $\mathcal{D}$.

   (c) What is the estimated expected value?

   (d) What is the estimated variance?

   (e) What is the estimated $P(Y_1 = 2)$

5. Lets assume we decide to study the incubation period for the influenza virus. The incubation period is defined as the number of days between when the virus infects a host and when that host becomes symptomatic. We collect from 5 individuals the date they came in contact with someone who was infected with influenza and the date they themselves were symptomatic.

$$\mathcal{D} = \begin{bmatrix} \text{Date of contact} & \text{Date of Symptoms} \\ 02/20 & 02/21 \\ 04/12 & 04/17 \\ 03/22 & 03/25 \\ 10/24 & 10/26 \\ 07/15 & 07/18 \end{bmatrix} \tag{4.138}$$

   (a) Provide a statistical setup for this data. Define a sample, assumptions about the sample, and a distribution for each random variable that is a part of the sample

   (b) Estimate the parameters in your statistical setup using the Method of Moments

   (c) Describe your results

6. Suppose $(Y_1, Y_2, \cdots, Y_n)$ is a random sample where $Y_i \sim \text{Pois}(\lambda)$.

   (a) Define an estimator for λ using the MoM

   (b) Characterize the distribution of $\hat{\lambda}$ using the CLT.

7. The Aedes mosquito is a vector for yellow and dengue fever, the Zika virus, and Chikungunya. Tracking the daily incidence of Aedes mosquitoes in a given location is one signal associated with the incidence of these four diseases. Mosquitoes are routinely captured and counted and the daily frequency of mosquitoes is reported to departments of public health.

   Suppose we capture and count the number of mosquitoes in a specific county for two weeks, and compile this data in the following dataset $\mathcal{D} = (124, 98, 188, 212, 100, 34, 90, 99, 46, 176, 67, 94, 344, 67)$ where one data point represents the number of mosquitoes collected in one day.

   (a) Define a set of random variables that we will use to model the daily frequency of mosquitoes. Include notation for the random sample and define a common Poisson distribution for all random variables.

   (b) Use the method of moments to estimate parameter value (λ) from the given dataset $\mathcal{D}$

   (c) Estimate the probability that we observe 124 mosquitoes

   (d) Estimate the probability that we observe 120 to 125 mosquitoes

   (e) Estimate the expected value and the variance

   (f) Let's rework the above model and assume that each random variable follows a Normal distribution. Use the method of moments to estimate parameters value (μ, σ^2) from the given dataset $\mathcal{D}$.

   (g) Estimate the expected value and the variance for the Normal model

8. A Randomized Control trial (RCT) is a common study design to compare the efficacy and safety of a novel device. Suppose the sponsor for a new cardiovascular device wishes to compare the safety of their device, which they define as the percent of patients who survive 7 days after the procedure, and the efficacy, which they define as the percent of patients who survive at one year after the procedure.

   The trial enrolls 50 patients to receive the device (the device group) and 50 patients who receive optimal medical therapy (the control group). The trial enrolls these 100 patients over the course of two years. At year 3 all 100 patients have been contacted at one year after their procedure (called a patient's one year followup).

   We find that in the device group 8 patients did not survive at or after 7 days from the date of their procedure and 18 patients did not survive one year after the date of their procedure. In the control group 9 patients did not survive at or after 7 days and 33 patients did not survive one year after the procedure.

(a) Define a set of random variables that we will use to model the number of patients in the device group who survive 7 days after the procedure out of a total of 50 patients and a second set of random variables to model the number of patients in the control group who survive 7 days after the procedure out of a total of 50 patients. Include mathematical notation for the random sample and a common distribution for both sets of random variables.

(b) Use the Method of Moments to estimate the parameters for the distribution you chose for survival at or after 7 day for patients in the device group. Report the parameter estimate.

(c) Use the Method of Moments to estimate the parameters for the distribution you chose for survival at or after 7 days for patients in the control group. Report the parameter estimate.

(d) What can you conclude about the safety of the novel device compared to the safety of optimal medical therapy?

(e) Use the Method of Moments to estimate the parameters for the distribution you chose for survival at one year for patients in the device group. Report the parameter estimate.

(f) Use the Method of Moments to estimate the parameters for the distribution you chose for survival at one year for patients in the control group. Report the parameter estimate.

(g) What can you conclude about the efficacy of the novel device compared to the efficacy of optimal medical therapy?

9. Suppose a random sample $(Y_1, Y_2, \cdots, Y_{10})$ generates the following dataset $\mathcal{D} = (0, -2.3, 0.4, 9.2, 10, -9.8, 3, 3, 3, 0)$. Assume that $X_i \sim \mathcal{N}(\mu, \sigma^2)$ and compute the corresponding ten z-scores.

10. Suppose we collect the following dataset $\mathcal{D} = (6.51, -1.11, 7.29, 0.23, 3.45, 0.85, -0.42, 5.66, 0.04, -1.31)$. Further, lets assume $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

(a) Please compute a $1 - \alpha$ confidence interval for μ. The symbol $z_{1-\alpha/2}$ should appear in this interval.

(b) Please compute a 95% confidence interval for μ.

(c) Please compute a 80% confidence interval for μ.

(d) Why is your 95% confidence interval larger than your 80% confidence interval?

11. Suppose we collect the following dataset $\mathcal{D} = (4, 0, 1, 2, 8, 13, 0, 1, 0, 3)$. Further, lets assume $X_i \sim \text{Geom}(p)$.
    (a) Please compute a $1 - \alpha$ confidence interval for $\frac{1}{p}$. The symbol $z_{1-\alpha/2}$ should appear in this interval.
    (b) Please compute a 95% confidence interval for $\frac{1}{p}$.
    (c) Please compute a 80% confidence interval for $\frac{1}{p}$.
12. Suppose we collect the following dataset $\mathcal{D} = (0, 1, 1, 1, 1, 1, 1, 0, 1, 1)$. Further, lets assume $X_i \sim \text{Bernoulli}(\theta)$.
    (a) Please compute a $1 - \alpha$ confidence interval for θ. The symbol $z_{1-\alpha/2}$ should appear in this interval.
    (b) Please compute a 95% confidence interval for θ.
    (c) Please compute a 80% confidence interval for θ.
13. Suppose we collect the following dataset $\mathcal{D} = (6, 6, 2, 7, 3, 3, 5, 5, 7, 5)$. Further, lets assume $X_i \sim \text{Binomial}(50, \theta)$.
    (a) Please compute a $1 - \alpha$ confidence interval for θ. The symbol $z_{1-\alpha/2}$ should appear in this interval.
    (b) Please compute a 95% confidence interval for θ.
    (c) Please compute a 80% confidence interval for θ.
14. Suppose we collect the following dataset $\mathcal{D} = (1, 4, 2, 1, 1, 4, 3, 2, 3, 0)$. Further, lets assume $X_i \sim \text{Poisson}(\lambda)$.
    (a) Please compute a $1 - \alpha$ confidence interval for λ. The symbol $z_{1-\alpha/2}$ should appear in this interval.
    (b) Please compute a 95% confidence interval for λ.
    (c) Please compute a 80% confidence interval for λ.
15. Consider two random samples $(Y_1, Y_2, \cdots, Y_{n_1})$ and $(X_1, X_2, \cdots, X_{n_2})$ where $Y_i \sim \text{Bern}(\theta_y)$ and $X_i \sim \text{Bern}(\theta_x)$. Please compute a $1 - \alpha$ confidence interval for $\overline{Y} - \overline{X}$.
16. Suppose we collect the following dataset $\mathcal{D}_y = (0, 1, 1, 0, 1, 0, 1, 1, 1)$ corresponding to a random sample $(Y_1, Y_2, \cdots, Y_9)$ where $Y_i \sim \text{Bern}(\theta_y)$ and a dataset $\mathcal{D}_x = (0, 1, 0, 0, 0, 0)$ corresponding to a random sample $(X_1, X_2, \cdots, X_6)$ where $X_i \sim \text{Bern}(\theta_x)$.
    (a) Please compute a 95% confidence interval for $\theta_y - \theta_x$.
    (b) Please compute a 95% confidence interval for $\theta_x - \theta_y$.
    (c) Please compute a 80% confidence interval for $\theta_y - \theta_x$.

17. Suppose we collect the following dataset $\mathcal{D}_y = (3, 3, 2, 1, 3, 2, 1, 4, 0, 3)$ corresponding to a random sample $(Y_1, Y_2, \cdots, Y_{10})$ where $Y_i \sim \text{Pois}(\lambda_y)$ and a dataset $\mathcal{D}_x = (1, 0, 3, 0, 1, 2, 3, 1)$ corresponding to a random sample $(X_1, X_2, \cdots, X_8)$ where $X_i \sim \text{Pois}(\lambda_x)$.

    (a) Please compute a 95% confidence interval for $\lambda_y - \lambda_x$.

    (b) Please compute a 80% confidence interval for $\lambda_y - \lambda_x$.

18. Suppose we collect the following dataset $\mathcal{D}_y = (-2.02, 0.78, -0.20, -0.47, 0.11, -2.57, 1.27, 3.17, 0.60)$ corresponding to a random sample $(Y_1, Y_2, \cdots, Y_9)$ where $Y_i \sim \mathcal{N}(\mu_y, 2)$ and a dataset $\mathcal{D}_x = (2.46, -6.00, -3.67, -1.48, -3.60, 1.72, -1.28, -2.93, -1.70)$ corresponding to a random sample $(X_1, X_2, \cdots, X_9)$ where $X_i \sim \mathcal{N}(\mu_x, 2)$.

    (a) Please compute a 95% confidence interval for $\mu y - \mu_x$.

    (b) Please compute a 80% confidence interval for $\mu_y - \mu_x$.

19. Suppose we collect a dataset $\mathcal{D} = (2, 5, 1, 10, 5, 8, 3)$. Assume that this data was generated from a random sample $(Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_7)$ where $Y_i \sim \text{Pois}(\lambda)$.

    Our null and alternative hypotheses for the parameter λ are

    $$H_0 : \lambda = 4 \tag{4.139}$$

    $$H_1 : \lambda > 4 \tag{4.140}$$

    (a) Please compute a Z test statistic for the above sample.

    (b) Given a critical region $C = \{(y_1, y_2, \cdots, y_7) \,|\, Z > 2.0\}$, would we reject the null hypothesis?

    (c) Please provide an intuitive explanation for the critical region C.

    (d) If we developed a new critical region C_1 and $C_1 \subset C$ would our type I error for C_1 be smaller or larger than the type I error for C?

20. Suppose we are asked to analyze the primary safety endpoint for a clinical trial at a point when half of the patients were enrolled and followed for one year. The trial plans to enroll a total of 70 patients. At one year we find that 12% of patient experienced the safety endpoint, defined as a combination of all cause mortality, myocardial infarction, or stroke. The trial is assumed safe if the true proportion of events is smaller than 10%.

    (a) Please develop a statistical setup for the above dataset. Include a sequence of random variables and define a common distribution for this sample.

    (b) Derive a method of moments estimator for the above parameter(s) of your common distribution.

(c) Compute an estimate of your parameters using your MoM estimator.

(d) Compute a 95% confidence interval for your parameter(s).

(e) State a reasonable null and alternative hypothesis to test if your parameter is below 0.10.

(f) Please compute a Z test statistic for the above sample and your hypotheses.

(g) Given a critical region $C = \{(y_1, y_2, \cdots) \,|\, Z < -1.96\}$, would we reject the null hypothesis?

(h) Would the pvalue that corresponds to this test statistic be greater than or smaller than 0.05? Why?

5

Likelihood theory

thomas mcandrew

Lehigh University

CONTENTS

5.1 Introduction

There are many different strategies for estimating parameters when given a sample $(X_1, X_2, X_3, \cdots, X_n)$, and a single relation of all the random variables in this sample—what we call a dataset ($\mathcal{D}$). Each strategy for estimating parameters has advantages and disadvantages. Up until this point we learned a single strategy called *The Method of Moments*. Lets explore a second, popular strategy for estimating parameters given a sample and a model called *The likelihood approach*.

The likelihood approach has as advantages that estimators are often simpler to compute than the method of moments, we can construct a straightforward method to compute the variance of estimators which is based on the curvature of a likelihood functions, and choosing parameters that "best fit" the data is an intuitive approach.

5.2 Probability of a dataset

Suppose that we are provided a single dataset $\mathcal{D} = (y_1, y_2, \cdots, y_n)$, and we assume that this dataset was generated from a sample $(Y_1, Y_2, Y_3, \cdots, Y_n)$.

The **Probability of our dataset**—$P(\mathcal{D})$—is defined as

$$\begin{aligned} P(\mathcal{D}) &= P(Y_1 = y_1 \cap Y_2 = y_2 \cap \cdots \cap Y_n = y_n) && (5.1)\\ &= P(Y_1 = y_1, Y_2 = y_2, \cdots, Y_n = y_n) && (5.2) \end{aligned}$$

where we use a comma to denote the $\cap$ symbol. The probability of our dataset is the probability that random variable Y_1 generates the realized value y_1, Y_2 generates y_2 , and so on, all simultaneously.

If we assume that our sample is independent and identically distributed (i.i.d.) then we can simplify $P(\mathcal{D})$.

We need one fact first.

5.2.0.1 Conditional probability chains

Consider computing the probability of a sample of size three (X_1, X_2, X_3).

$$P(X_1 = x_1 \cap X_2 = x_2 \cap X_3 = x_3) \tag{5.3}$$

We can imagine that inside the **event** $X_1 = x_1 \cap X_2 = x_2 \cap X_3 = x_3$ there exists two events.

$$[X_1 = x_1] \cap [X_2 = x_2 \cap X_3 = x_3] \tag{5.4}$$

Here, we partitioned the above event that contained three random variables into two events: one event that contains a single random variable and a second event that contains two random variables.

Now we can use an identity that we learned in the past to break this probability into two. For two events A and B, recall

$$P(A \cap B) = P(B|A)P(A). \tag{5.5}$$

Then

$$P([X_1 = x_1] \cap [X_2 = x_2 \cap X_3 = x_3]) = \tag{5.6}$$
$$P(X_1 = x_1 | [X_2 = x_2 \cap X_3 = x_3]) P(X_2 = x_2 \cap X_3 = x_3). \tag{5.7}$$

We could apply the same technique above to the event $X_2 = x_2 \cap X_3 = x_3$ and find that

$$P(X_1 = x_1 | [X_2 = x_2 \cap X_3 = x_3]) P(X_2 = x_2 \cap X_3 = x_3) = \tag{5.8}$$
$$P(X_1 = x_1 | [X_2 = x_2 \cap X_3 = x_3]) P(X_2 = x_2 | X_3 = x_3) P(X_3 = x_3) \tag{5.9}$$

The process of breaking a single probability into several probabilities in this way is sometimes called producing a conditional chain of probabilities.

5.2.1 Probability of an i.i.d. sample

Given an i.i.d. sample $(Y_1, Y_2, \cdots, Y_n)$ the probability of this sample is

$$P(Y_1 = y_1, Y_2 = y_2, \cdots Y_n = y_n) \tag{5.10}$$

We can break this probability into several conditional probabilities.

$$P(Y_1 = y_1, Y_2 = y_2, \cdots Y_n = y_n) = \tag{5.11}$$
$$P(Y_1 | Y_2 \cap Y_3 \cap \cdots Y_n) P(Y_2 | Y_3 \cap \cdots Y_n) P(Y_3 | Y_4 \cap \cdots Y_n) \cdots \tag{5.12}$$
$$P(Y_{n-1} | Y_n) P(Y_n) \tag{5.13}$$

But because we assume our sample is (mutually) independent then

$$P(Y_1 | Y_2 \cap Y_3 \cap \cdots Y_n) P(Y_2 | Y_3 \cap \cdots Y_n) P(Y_3 | Y_4 \cap \cdots \tag{5.14}$$
$$Y_n) \cdots P(Y_{n-1} | Y_n) P(Y_n) = P(Y_1) P(Y_2) \cdots P(Y_n) \tag{5.15}$$

To recap, the probability of a dataset generated by an i.i.d. sample is

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \cdots X_n = x_n) = \tag{5.16}$$
$$P(X_1 = x_1) P(X_2 = x_2) P(X_3 = x_3) \cdots P(X_n = x_n) \tag{5.17}$$
$$= \prod_{i=1}^{N} P(X_i = x_i) \tag{5.18}$$

Example: Suppose an i.i.d. sample $(Y_1, Y_2, \cdots, Y_n)$ generated a dataset $\mathcal{D} = (y_1, y_2, \cdots, y_n)$. Further assume $Y_i \sim \text{Bern}(\theta)$ for $i = 1$ to N. Then

$$
\begin{aligned}
P(\mathcal{D}|\theta) &= P(Y_1 = y_1)P(Y_2 = y_2)\cdots P(Y_n = y_n) && (5.19)\\
&= \left[\theta^{y_1}(1-\theta)^{1-y_1}\right]\left[\theta^{y_2}(1-\theta)^{1-y_2}\right]\cdots\left[\theta^{y_n}(1-\theta)^{1-y_n}\right] && (5.20)\\
&= \prod_{i=1}^{N}\left[\theta^{y_i}(1-\theta)^{1-y_i}\right] && (5.21)\\
&= \theta^{\sum_{i=1}^{N} y_i}(1-\theta)^{N-\sum_{i=1}^{N} y_i} && (5.22)
\end{aligned}
$$

5.3 The likelihood function

The probability of our dataset $P(\mathcal{D})$ depends on a set of parameters (θ), and we can write $P(\mathcal{D}|\theta)$. Different choices for our parameter values will result in different values of $P(\mathcal{D}|\theta)$.

Example: (Bernoulli example continued). Suppose that we are given the dataset $\mathcal{D} = (1, 0, 1, 0, 0, 1, 1)$ and we assume that these 7 datapoints were generated from an i.i.d. Bernoulli model. Then, as we saw above,

$$P(\mathcal{D}|\theta) = \theta^{\sum_{i=1}^{N} y_i}(1-\theta)^{N-\sum_{i=1}^{N} y_i} \tag{5.23}$$

. For our dataset, we can fill in the y_i values to find

$$P(\mathcal{D}|\theta) = \theta^4(1-\theta)^{7-4} \tag{5.24}$$

We can now compute the probability of our data for different choices of θ

$$
\begin{aligned}
P(\mathcal{D}|\theta = 0.25) &= 0.25^4(1-0.25)^3 = && (5.25)\\
P(\mathcal{D}|\theta = 0.50) &= 0.50^4(1-0.50)^3 = && (5.26)\\
P(\mathcal{D}|\theta = 0.75) &= 0.75^4(1-0.75)^3 = && (5.27)
\end{aligned}
$$

The **likelihood function**, $\mathcal{L}$, is the probability of our observed data as a function of a set of parameter values

$$\mathcal{L}(\theta|\mathcal{D}) = P(\mathcal{D}|\theta) \tag{5.28}$$

The difference between $P(\mathcal{D}|\theta)$ and $\mathcal{L}(\theta|\mathcal{D})$ is the argument. For $P(\mathcal{D}|\theta)$ we expect, given a set of fixed parameter values, to compute the probability of different datasets. For $\mathcal{L}(\theta|\mathcal{D})$ we expect, given a fixed, observed dataset $\mathcal{D}$, to compute the probability of $\mathcal{D}$ for different values of θ.

The likelihood function allows us to investigate the plausability of different parameter values after we observed a dataset. The likelihood need not be a probability distribution, and often is not.

Example: Suppose we observe for 7 days the number of patients who arrive at a specific hospital and are diagnosed with influenza. The data we collect is $\mathcal{D} = (10, 13, 4, 17, 34, 11, 3)$. We choose to consider our data a random sample $(X_1, X_2, \cdots, X_7)$ where $X_i \sim \text{Pois}(\lambda)$. Then

$$
\begin{aligned}
P(\mathcal{D}|\lambda) &= \prod_{i=1}^{7} f(x_i|\lambda) && (5.29)\\
&= \prod_{i=1}^{7} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} && (5.30)\\
&= \frac{e^{-7\lambda}\lambda^{10+13+\cdots+3}}{10!13!4!\cdots 3!} && (5.31)
\end{aligned}
$$

If we consider the above as a function of λ—$\mathcal{L}(\lambda|\mathcal{D})$ then if $\mathcal{L}(\lambda_1|\mathcal{D}) > \mathcal{L}(\lambda_2|\mathcal{D})$ then the parameter value λ_1 may be a more plausible parameter value compared to λ_2.

One way to use the likelihood function is to attempt to find the most plausible parameter value, the value that maximizes the likelihood. However this value may be difficult to find. Lets explore one strategy for simplifying the search for this optimal parameter value and a second strategy that can help us compute this optimal value.

5.4 The logarithm and the log likelihood

With a likelihood function in hand, one aim may be to find the set of parameter values that maximizes this function—the set of parameter values that are more plausible given an observed dataset. However, the likelihood function $\mathcal{L}$ can in many cases be difficult to compute because it is a sequence of multiplications. Our aims is to find a related function $\ell\ell$ such that parameter values that maximize $\mathcal{L}$ also maximize this "easier to work with" $\ell\ell$ function.

A function g is **monotone increasing** if for two values $x > y$ then $g(x) \geq g(y)$. If a monotone function g is applied to a second function f, the values that maximize f also maximize this new function.

Hypothesis: Suppose the value x^* maximizes the function f. That is, for all x the value $f(x^*) \geq f(x)$. Then if g is a monotone increasing function then x^* also maximizes $g(f(x))$.

Evidence: We know that $f(x^*) \geq f(x)$. g is a monotone increasing function and so by definition for $x \geq y$, $g(x) > g(y)$. We can use as an input to g the values $f(x^*)$ and $f(x)$ to find that because $f(x^*) \geq f(x)$ then $g(f(x^*)) \geq g(f(x))$.

The above property of a monotone increasing function is useful if we can find such a function that simplifies our likelihood—a sequence of multiplications. The logarithm is a monotone increasing function (i.e. if $x \geq y$ then $\log(x) \geq \log(y)$), and the logarithm transforms multiplications (hard) into sums (easy).

The **log likelihood** function is the natural logarithm of the likelihood function

$$\ell\ell(\theta|\mathcal{D}) = \log\left[\mathcal{L}(\theta|\mathcal{D})\right] \tag{5.32}$$

The loglikelihood simplifies the likelihood by transforming a large sequence of multiplications into sums.

There is one more principle about maximizing functions that will be useful. Suppose we find the value (x^*) that maximizes the function

$$g(x) = f(x) + c \tag{5.33}$$

then that value x^* also maximizes $f(x)$. Also, suppose we find the value (x^*) that maximizes the function

$$g(x) = c \cdot f(x) \tag{5.34}$$

then that value x^* also maximizes $f(x)$.

The above two principles state that if we are interested in finding the maximum value of a function g then we can consider a simpler function f which does not contain unrelated constants.

Example: The value x^* that maximizes the function $g(x) = -x^2 + 3$ is the same value that maximizes the function $f(x) = -x^2$.

Example: The value x^* that maximizes the function $g(x) = \frac{e^{-x^2}}{4}$ is the same value that maximizes the function $f(x) = e^{-x^2}$.

Example: We found in the previous section that the likelihood for the number of admitted patients with influenza was

$$\mathcal{L}(\lambda|\mathcal{D}) = \frac{e^{-7\lambda}\lambda^{10+13+\cdots+3}}{10!13!4!\cdots 3!}. \tag{5.35}$$

The log likelihood of the above is, the simpler,

$$\ell\ell(\lambda|\mathcal{D}) = \log\left[\frac{e^{-7\lambda}\lambda^{10+13+\cdots+3}}{10!13!4!\cdots 3!}\right] \tag{5.36}$$

$$= -7\lambda + (10+13+\cdots+3)\log(\lambda) - \left[\log(10!) + \log(13!) + \cdots + \log(3!)\right] \tag{5.37}$$

We see that the last term of factorials does not involve our parameter value λ. This term is then a constant and we could instead aim to find the maximium of the simpler function

$$\begin{aligned} \ell\ell(\lambda|\mathcal{D}) &= -7\lambda + (10 + 13 + \cdots + 3)\log(\lambda) && (5.38) \\ &= -7\lambda + 83\log(\lambda) && (5.39) \end{aligned}$$

In some cases, we may be able to find the parameter value, or set of parameter values, that maximize the log likelihood function. To explore this topic we will need to briefly learn about the derivative of a continuous function.

5.5 Just in time derivatives

The derivative describes the tangential slope of a function at a point and can serve as a linear approximation to a more complicated function. We will see how we can use the derivative to setup and solve an equation that will return a parameter value that maximizes the log likelihood. This estimate of our parameter will be called a **maximum likelihood estimate**.

5.5.1 Sequences and limits

A sequence is defined as a function whose domain is the natural numbers $\mathbb{N} = \{1, 2, 3, 4, 5, 6, \cdots\}$. Often as annotate the value associated with the 1, $f(1)$ as a_1, the value associated with 2, $f(2)$, as a_2, and so on.

Example: The following is a sequence $f(x) = 2^{-x}$ for $x \in \mathbb{N}$.The first few values of this sequence are $f(1) = a_1 = 1/2$, $f(2) = a_2 = 1/4$, $f(3) = a_3 = 1/8$, and so on. We can also describe the above sequence using this notation $\{2^{-n}\}_{n=1}^{\infty}$.

Intuitively, we may wish to know how a sequence behaves as it is enumerated. A sequence may increase to infinity, decrease to negative infinity, oscillate or have some other periodic behavior, or a sequence may move closer and closer to a single value. If a sequence $a_{n}{}_{n=1}^{\infty}$ gets closer and closer to a single value l then we say that the sequence has a **limit** and that limit is the value l.

The more formal definition of a limit is the following: For a sequence $\{a_n\}_{n=1}^{\infty}$, given any value $\epsilon > 0$ the value l is a limit if you can find an integer N such that $|a_N - l| < \epsilon$ for integers greater than N (i.e. $N+1$, $N+2, \cdots$).

We may also define a limit L of a function at the value a if given a sequence of values $a_1, a_2, \cdots a_n$ that approach the value a, the sequence $f(a_1), f(a_2), f(a_3), \cdots$ approaches the value L. We call L the limit of f at the value a and write

$$\lim_{x \to a} f(x) = L \tag{5.40}$$

Intuitively, the above says that as the input of our function approaches the value a, the function values approach the value L.

To solve our problem of finding the parameter value that maximizes the log likelihood, we need to learn about a special type of limit called the derivative.

5.5.2 Derivative of a function

The derivative of a function at the value a, or $f'(a)$, is

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a} \tag{5.41}$$

For intuition, we can imagine building a sequence

$$a_1, a_2, a_3, \cdots, a_n \to a \tag{5.42}$$

that approaches the value a and computing the associated sequence

$$\frac{f(a_1) - f(a)}{a_1 - a}, \frac{f(a_2) - f(a)}{a_2 - a}, \frac{f(a_3) - f(a)}{a_3 - a}, \cdots, \frac{f(a_n) - f(a)}{a_n - a} \to f'(a) \tag{5.43}$$

One common sequence that approaches a is the sequence $a + \{h_n\}_{n=1}^{\infty}$ and where the sequence h approaches 0. Then the definition of the derivative would be

$$\begin{aligned} f'(a) &= \lim_{h \to 0} \frac{f(a+h) - f(a)}{a + h - a} && (5.44) \\ &= \lim_{h \to 0} \frac{f(a+h) - f(a)}{h} && (5.45) \end{aligned}$$

This is because as the sequence $\{h_n\}$ approaches zero, the sequence $\{a + h_n\}$ approaches the value a.

Example: Let $g(x) = x^2$. The derivative of g at the value 1 is by definition

$$\begin{aligned}
g'(1) &= \lim_{h \to 0} \frac{g(1+h) - g(1)}{h} && (5.46)\\
&= \lim_{h \to 0} \frac{(1+h)^2 - 1^2}{h} && (5.47)\\
&= \lim_{h \to 0} \frac{h^2 + 2h + 1 - 1}{h} && (5.48)\\
&= \lim_{h \to 0} \frac{h^2 + 2h}{h} && (5.49)\\
&= \lim_{h \to 0} \frac{h + 2}{1} && (5.50)\\
&= 2 && (5.51)
\end{aligned}$$

For any value c we find that the derivative of g is

$$\begin{aligned}
g'(c) &= \lim_{h \to 0} \frac{g(c+h) - g(c)}{h} && (5.52)\\
&= \lim_{h \to 0} \frac{(c+h)^2 - c^2}{h} && (5.53)\\
&= \lim_{h \to 0} \frac{h^2 + 2hc + c^2 - c^2}{h} && (5.54)\\
&= \lim_{h \to 0} \frac{h^2 + 2hc}{h} && (5.55)\\
&= \lim_{h \to 0} \frac{h + 2c}{1} && (5.56)\\
&= 2c && (5.57)
\end{aligned}$$

In some sense, the derivative $g'(c)$ too is a function, the function $g'(c) = 2c$.

5.5.3 When the derivative equals zero

When the derivative of a function at the point x^* is 0 the point x^* is either a minimum or a maximum value. This is the key to finding a parameter value that maximizes our log likelihood. If we can compute the derivative of our log likelihood $\ell\ell'(\theta)$ and find the value θ such that $\ell\ell' = 0$ then we know we have found either a parameter value that minimizes or maximizes our log likelhood, which then minimizes or maximizes our likelihood, which suggests that this parameter value is the most plausible value given our observed data. Phew.

Lets prove that when the derivative equal zero then we are at a maximum. *Hypothesis*: If $f(x^*) > f(x)$ for any possible input x then $f'(x^*) = 0$.

Let us build a two sequences, one sequence $\{x_1, x_2, \cdots\}$ that approaches x^* such that all the values are smaller than x^* and a second sequence $\{y_1, y_2, \cdots\}$ such that all values are larger than x^*.

If we choose any x_n in the sequence, the following must be true:

$$\frac{f(x_n) - f(x^*)}{x_n - x^*} \geq 0 \qquad \text{bc } f(x^*) \text{ is a max and } x_n < x^* \tag{5.58}$$

$$\tag{5.59}$$

If we choose any y_n in the sequence, the following must also be true:

$$\frac{f(y_n) - f(y^*)}{y_n - x^*} \leq 0 \qquad \text{bc } f(x^*) \text{ is a max and } y_n > y^* \tag{5.60}$$

$$\tag{5.61}$$

If we assume that there exists a derivative value $f'(x^*)$ then that value must be greater than or equal to zero and also less than or equal to zero. The only options then is $f'(x^*) = 0$.

5.5.4 derivative "rules"

There are several rules that can make computing derivatives of complicated functions easier.

- Constant Rule: if $f(x) = c$ where c is a constant then $f'(x) = 0$
  - Example: if $f(x) = 72.67$ then

$$f'(x) = 0 \tag{5.62}$$

- Power Rule: if $f(x) = x^n$ then $f'(x) = nx^{(n-1)}$
  - Example: if $f(x) = x^{3/2}$ then $f'(x) = \frac{3}{2}x^{3/2-1} = \frac{3}{2}x^{1/2}$
- Exp Rule: if $f(x) = e^x$ then $f'(x) = e^x$
- log Rule: if $f(x) = \log(x)$ then $f'(x) = \frac{1}{x}$
- Addition Rule: if $f(x) = g(x) + h(x)$ then $f'(x) = g'(x) + h'(x)$
  - Example: if $f(x) = x^2 + \frac{1}{3}x^{-1}$ then

$$f'(x) = \left(x^2\right)' + \left(\frac{1}{3}x^{-1}\right)' \tag{5.63}$$

$$f'(x) = 2x^1 + \frac{-1}{3}x^{-2} \tag{5.64}$$

- "Chain" rule: if $f(x) = g(h(x))$ then $f'(x) = g'(h(x)) \cdot h'(x)$
  - Example: Suppose $f(x) = (x+2)^2$ then g(h(x)) = $h(x)^2$ and $h(x) = x + 2$.

$$h'(x) = 1 \tag{5.65}$$
$$g'(h(x)) = 2h(x) \tag{5.66}$$

Therefore

$$f'(x) = g'(h(x)) \cdot h'(x) \tag{5.67}$$
$$= 2h(x) \cdot 1 \tag{5.68}$$
$$= 2(x+2) \tag{5.69}$$

  - Example: Suppose $f(x) = \sqrt{\log(x)}$ then g(h(x)) = $h(x)^{1/2}$ and $h(x) = \log(x)$.

$$h'(x) = \frac{1}{x} \tag{5.70}$$
$$g'(h(x)) = \frac{1}{2}h(x)^{-1/2} \tag{5.71}$$

Therefore

$$f'(x) = g'(h(x)) \cdot h'(x) \tag{5.72}$$
$$= \frac{1}{2}h(x)^{-1/2} \cdot \frac{1}{x} \tag{5.73}$$
$$= \frac{\log(x)^{-1/2}}{2x} \tag{5.74}$$
$$= \frac{1}{2x\sqrt{\log(x)}} \tag{5.75}$$

5.6 Computing the maximum likelihood estimator

Given a i.i.d sample $(X_1, X_2, \cdots, X_n)$ such that $X \sim f(x|\theta)$ we can compute a maximum likelihood estimator of the parameter θ by following these steps: (i) form the likelihood function (ii) convert the likelihood to a log likelihood (iii) compute the derivative of the log likelihood $\ell\ell'(\theta)$ (iv) find the parameter value θ such that $\ell\ell'(\theta) = 0$.

Example: Given $(X_1, X_2, X_3, \cdots, X_n)$, the distribution for each $X_i \sim \text{Bern}(\theta)$, and the dataset $\mathcal{D} = (0, 1, 1, 0, 1)$ we can compute the maximum likelihood estimate for θ.

1. The likelihood of $\mathcal{D}$ is

$$\begin{aligned}
\ell(\theta) &= f(0) \cdot f(1) \cdot f(1) \cdot f(0) \cdot f(1) \\
&= \left(\theta^{x_1}(1-\theta)^{1-x_1}\right) \left(\theta^{x_2}(1-\theta)^{1-x_2}\right) \left(\theta^{x_3}(1-\theta)^{1-x_3}\right) \left(\theta^{x_4}(1-\theta)^{1-x_4}\right) \left(\theta^{x_5}(1-\theta)^{1-x_5}\right) \\
&= \left(\theta^{0}(1-\theta)^{1-0}\right) \left(\theta^{1}(1-\theta)^{1-1}\right) \left(\theta^{1}(1-\theta)^{1-1}\right) \left(\theta^{0}(1-\theta)^{1-0}\right) \left(\theta^{1}(1-\theta)^{1-1}\right) \\
&= (1-\theta)\,(\theta)\,(\theta)\,(1-\theta)\,(\theta) \\
&= \theta^3(1-\theta)^2
\end{aligned}$$

2. Compute the log likelihood

$$\begin{aligned}
\ell\ell(\theta) &= \log\left[\ell(\theta)\right] \\
&= \log\left[\theta^3(1-\theta)^2\right] \\
&= \log\left[\theta^3\right] + \log\left[(1-\theta)^2\right] \\
&= 3\log\left[\theta\right] + 2\log\left[(1-\theta)\right]
\end{aligned}$$

3. Compute the derivative of the log likelihood

$$\begin{aligned}
\ell\ell(\theta)' &= (3\log[\theta] + 2\log[(1-\theta)])' \\
&= (3\log[\theta])' + (2\log[(1-\theta)])' && \text{Addition rule} \\
&= 3\,(\log[\theta])' + 2\,(\log[(1-\theta)])' && \text{Constants rule} \\
&= 3\frac{1}{\theta} + 2\,(\log[(1-\theta)])' && \text{Derivative of log} \\
&= 3\frac{1}{\theta} + 2\,(\log[z])' && \text{let } z = (1-\theta) \\
&= 3\frac{1}{\theta} + 2\frac{1}{z}(z)' && \text{Chain rule} \\
&= 3\frac{1}{\theta} + 2\frac{1}{1-\theta}(1-\theta)' && \text{Replace z} \\
&= 3\frac{1}{\theta} + 2\frac{1}{1-\theta}(1)' - (\theta)' && \text{Addition rule} \\
&= 3\frac{1}{\theta} + 2\frac{1}{1-\theta}0 - (\theta)' && \text{Derivative of a constant} \\
&= 3\frac{1}{\theta} + 2\frac{1}{1-\theta}(-1) && \text{Derivative of } \theta \\
&= 3\frac{1}{\theta} - 2\frac{1}{1-\theta} \\
&= \frac{3}{\theta} - \frac{2}{1-\theta}
\end{aligned}$$

4. Set $\ell\ell(\theta)' = 0$ and solve for θ

$$\begin{aligned}
\ell\ell'(\theta) &= \frac{3}{\theta} - \frac{2}{1-\theta} = 0 \\
\frac{3}{\theta} &= \frac{2}{1-\theta} \\
3 - 3\theta &= 2\theta \\
3 &= 5\theta \\
\theta &= 3/5 \\
\hat{\theta} &= 3/5
\end{aligned}$$

For our observed data $\mathcal{D}$ and our assumed model that the random variables generating each data point are i.i.d and distributed Bern(θ) our maximum likelihood estimate for θ is $\hat{\theta}_{\text{mle}} = 3/5$.

Example: Given $(Y_1, Y_2, Y_3, \cdots, Y_n)$, the distribution for each $Y_i \sim \text{Geom}(p)$, and the dataset $\mathcal{D} = (7, 19, 1, 11)$ we can compute the maximum likelihood estimate for the parameter p.

1. The likelihood of $\mathcal{D}$ is

$$\begin{aligned}
\ell(p) &= f(9) \cdot f(19) \cdot f(1) \cdot f(11) && (5.76) \\
&= p(1-p)^8 \cdot p(1-p)^{18} \cdot p(1-p)^0 \cdot p(1-p)^{10} && (5.77) \\
&= p^4(1-p)^{8+18+0+10} && (5.78) \\
&= p^4(1-p)^{36} && (5.79) \\
& && (5.80)
\end{aligned}$$

2. Form the log likelihood

$$\begin{aligned}
\ell\ell(p) &= \log\left[\ell(p)\right] && (5.81) \\
&= \log\left[p^4(1-p)^{36}\right] && (5.82) \\
&= \log\left[p^4\right] + \log\left[(1-p)^{36}\right] && (5.83) \\
&= 4\log\left[p\right] + 36\log\left[(1-p)\right] && (5.84)
\end{aligned}$$

3. Compute the derivative of the log likelihood

$$\ell\ell'(p) = 4\frac{1}{p} + 36\log(h(x)) \text{ where } h(x) = 1 - p \tag{5.85}$$

$$= \frac{4}{p} + 36\frac{1}{h(x)} \cdot -1 \tag{5.86}$$

$$= \frac{4}{p} - 36\frac{1}{1-p} \tag{5.87}$$

$$= \frac{4}{p} - \frac{36}{1-p} \tag{5.88}$$

4. Set the derivative equal to zero and solve for the parameter

$$\ell\ell'(p) = \frac{4}{p} - \frac{36}{1-p} = 0 \tag{5.89}$$

$$= \frac{4p(1-p)}{p} - \frac{36p(1-p)}{1-p} = 0 \tag{5.90}$$

$$= 4(1-p) - 36p = 0 \tag{5.91}$$

$$= 4 - 4p - 36p = 0 \tag{5.92}$$

$$= 4 = 40p \tag{5.93}$$

$$p = 4/40 = 1/10 \tag{5.94}$$

For our observed data $\mathcal{D}$ and our assumed model that the random variables generating each data point are i.i.d and distributed Geom(p), our maximum likelihood estimate for p is $\hat{p}_{\text{mle}} = \frac{1}{10}$.

5.7 Fisher information

A maximum likelihood estimate for a parameter describes what single parameter value most probably generated our observed dataset. However, intuitively, if we had a collected a different dataset of the same size then we would likely have arrived at a different maximum likelihood estimate. Repeating the steps: (i) generate dataset of same size, (ii) compute maximum likelihood estimate, (iii) store estimate would lead to many estimates of the same model. To better characterize our model, we will need in addition to a single maximum likelihood estimate a measure of how that estimate may vary.

The Fisher Information $\mathcal{I}(\theta)$ about a parameter θ is defined as

$$\mathcal{I}(\theta) = \mathbb{E}\left(\left[\ell'(x|\theta)\right]^2\right) = \mathbb{V}\left(\ell(\theta)\right) \tag{5.95}$$

the intuition behind the Fisher information is that log likelihood function that are very peaked around their maximum likelihood estimate will have large Fisher information values. On the contrary, if a log-likelihood is very flat near it's maximum likelihood estimate then the Fisher information will be small.

In some sense, the Fisher information describes the curvature of the log likelihood around the maximum likelihood estimate and that curvature is also related to the variability of our estimate.

We will not explore this topic in class, but it is often easier to compute an equivalent definition of the Fisher information

$$-\,\mathbb{E}(\ell\ell''(\theta|X)) = \mathbb{V}(\ell\ell'(\theta|X)). \tag{5.96}$$

This definition says that in order to compute the Fisher information we will need to compute the second derivative of the log likelihood, compute the expedctation, and then multiply that second derivative by negative one.

Lets look at an example of using this alternative definition to compute the Fisher information for a Bernoulli model.

Example: Suppose a sample $(X_1, X_2, X_3, \cdots, X_n)$ that we assume $X_i \sim$ Bernoulli(θ). Then the log likelihood for a single data point is

$$\begin{aligned}
\ell(\theta) &= \log\left[f(d_1|\theta)\right] && (5.97)\\
&= \log\left[\theta^{d_1}(1-\theta)^{1-d_1}\right] && (5.98)\\
&= d_1\log(\theta) + (1-d_1)\log(1-\theta) && (5.99)\\
& && (5.100)
\end{aligned}$$

and the derivative of the log likelihood is

$$\ell'(\theta) = \frac{d_1}{\theta} - \frac{1-d_1}{1-\theta} \tag{5.101}$$

To compute the second derivative, we take the derivative of $\ell'(\theta)$.

$$\begin{aligned}
\ell'(\theta) &= d_1\theta^{-1} - (1-d_1)(1-\theta)^{-1} && (5.102)\\
\ell''(\theta) &= -d_1\theta^{-2} - (1-d_1)(1-\theta)^{-2} && (5.103)
\end{aligned}$$

The above if the second derivative of the log likelihood for a single datapoint. However that data point was generated from a random variable X_1 and so we can then treat $\ell''(\theta)$ as a transformation of the random variable X_1.

$$\begin{aligned}
\ell''(X|\theta) &= -X\theta^{-2} - (1-X)(1-\theta)^{-2} && (5.104)\\
\ell''(X|\theta) &= -\frac{X}{\theta^2} - \frac{1-X}{(1-\theta)^2} && (5.105)\\
& && (5.106)
\end{aligned}$$

The above transformation is a complicated function, but the fisher information only requires that we understand as single quantity from this transformation—the expected value.

$$\mathbb{E}\left[\ell''(X|\theta)\right] = -\frac{\mathbb{E}(X)}{\theta^2} - \frac{1-\mathbb{E}(X)}{(1-\theta)^2} \tag{5.107}$$

$$= -\frac{\theta}{\theta^2} - \frac{1-\theta}{(1-\theta)^2} \tag{5.108}$$

$$= -\frac{1}{\theta} - \frac{1}{(1-\theta)} \tag{5.109}$$

The final steps asks that we take the negative of the quantity above

$$\mathcal{I}(\theta|X) = \frac{1}{\theta} + \frac{1}{(1-\theta)} \tag{5.110}$$

$$= \frac{1-\theta+\theta}{\theta(1-\theta)} \tag{5.111}$$

$$= \frac{1}{\theta(1-\theta)} \tag{5.112}$$

5.7.1 Fisher for a sample of size n

The Fisher information for a sample of size n $(X_1, X_2, \cdots, X_n)$ is equal to

$$\mathcal{I}(x_1, x_2, \cdots, x_n) = n\mathcal{I}(\theta) \tag{5.113}$$

Intuitively, a sample of size 8 is twice as informative as a sample of size 4.

5.8 CLT-like thm that involves the Fisher information

An important theorem that will allow us to construct confidence intervals and hypothesis tests about our parameters states the following:

$$\sqrt{n}(\hat{\theta} - \theta) \sim \mathcal{N}\left(0, \frac{1}{\mathcal{I}(\theta)}\right) \tag{5.114}$$

where $\hat{\theta}$ is our maximum likelihood estimate, θ is our true, and unknown, parameter, and $\mathcal{I}$ is the information about θ.

This theorem

5.9 Homework

1. Suppose that we collected a dataset with a single observation $\mathcal{D} = \{12\}$. Please compute the probability of $\mathcal{D}$ if we assume our data was generated by a random variable X

   (a) Such that $X \sim \text{Geom}(p)$. Assume we know the parameter value p.

   (b) Such that $X \sim \text{Pois}(\lambda)$. Assume we know the parameter value λ.

   (c) Such that $X \sim \mathcal{N}(\mu, \sigma^2)$. Assume we know the parameter values (μ, σ^2).

2. Suppose that we collected a dataset with a two observations $\mathcal{D} = \{12, 2\}$. Please compute the likelihood function for $\mathcal{D}$ if we assume our data was generated by a random variable X

   (a) Such that $X \sim \text{Geom}(p)$. Assume we know the parameter value p.

   (b) Such that $X \sim \text{Pois}(\lambda)$ Assume we know the parameter value λ.

   (c) Such that $X \sim \text{Binom}(20, \theta)$. Assume we know the parameter value θ.

3. Suppose that we collected a dataset with a n observations $\mathcal{D} = \{x_1, x_2, \cdots, x_n\}$. Please compute the log likelihood if we assume our data was generated by a random variable X

   (a) Such that $X \sim \text{Geom}(p)$. Assume we know the parameter value p.

   (b) Such that $X \sim \text{Pois}(\lambda)$ Assume we know the parameter value λ.

   (c) Such that $X \sim \text{Binom}(N, \theta)$. Assume we know the parameter value θ.

   (d) Such that $X \sim \mathcal{N}(\mu, \sigma^2)$. Assume we know the parameter values

4. Suppose you collect a dataset $\mathcal{D} = (x_1, x_2, \cdots, x_n)$ that you assume is generated from an i.i.d. sample $X_1, X_2, \cdots, X_n$, where $X_i \sim f(x|\theta)$ and θ is a parameter. You decide to compute the loglikelihood

$$\ell\ell(\theta|\mathcal{D}) = \sum_{i=1}^{n} \log\left[f(x_i|\theta)\right] \tag{5.115}$$

   Because we are interested in finding the value θ^* that maximizes $\ell\ell$ then multiplying the log likelihood by a constant will not change the optimal θ^*

$$\frac{1}{n}\ell\ell(\theta|x_1, x_2, \cdots, x_n) = \frac{1}{n}\sum_{i=1}^{n} \log\left[f(x_i|\theta)\right] \tag{5.116}$$

   (a) The law of large numbers states that the above quantity will approach?

5. Suppose that we are asked to help better understand the burden of influenza circulating in in the state of Pennsylvania. Over the course of 6 weeks we collect the following number of cases of reported influenza $\mathcal{D} = (143, 12, 124, 56, 66)$.

   (a) Please propose a model for this data and a short description for why you chose this model.

   (b) Please compute the likelihood for your model parameters given the data

   (c) Please compute the log likelihood.

6. Consider the sequence $a_n = 2^{-n}$ for $n = 1$ to ∞.

   (a) Write down the first 3 items in this sequence

   (b) Please compute
   $$\lim_{n\to\infty} a_n$$

7. Does the infinite sequence $\{-1, 1, -1, 1, -1, 1, \cdots\}$ have a limit? Why or why not?

8. Consider the finite sequence $a_n = 2^{-n}$ for $n = 1$ to 100, a sequence that ends at 2^{-100}.

   (a) Does this sequence have a limit point? Why or why not?

9. Compute the derivative of

   (a) $f(x) = e^{2x}$

   (b) $f(x) = \log(\frac{1}{2x})$

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

   (d) $f(x) = 10$

   (e) $f(x) = -3x$

   (f) $f(x) = \frac{1}{2}x^2$

   (g) $f(x) = \log(x)$

   (h) $f(x) = \log(x^2)$

   (i) $f(x) = \log(10)$

10. Consider the function $g(x) = -x^2 + 3$.

   (a) Compute the derivative of g

   (b) Does the x value that maximizes g not depend on the value of c? Why or why not?

11. Consider the function $g(x) = f(x) + c$ where c is a constant.

   (a) Compute the derivative of g

   (b) Does the x value that maximizes g depend on the value of c? Why or why not?

12. Consider the function $g(x) = cf(x)$ where c is a constant.

   (a) Compute the derivative of g

   (b) Does the x value that maximizes g depend on the value of c? Why or why not?

13. Assume a set of random variables $Z_1, Z_2, \cdots, Z_n$ such that $Z_i \sim \exp(\beta)$ generated an observed dataset. An exponential distribution assign a density over continuous numbers from 0 to infinity. The probability density function is

$$f(z) = \beta e^{-\beta z}$$

   Suppose we collect a dataset $\mathcal{D} = (2, 4, 0.5, 3)$.

   (a) Please compute the log likelihood of an arbitrary dataset $\mathcal{D} = (z_1, z_2, z_3 \cdots, z_n)$

   (b) Compute the derivative of the log likelihood above

   (c) Set the derivative of the log likelihood equal to zero and solve for β (ie find the parameter value that maximizes the log likelhiood).

   (d) Apply the maximum likelihood estimate you found to the dataset $\mathcal{D} = (2, 4, 0.5, 3)$.

14. Assume a set of random variables $Y_1, Y_2, \cdots, Y_n$ such that $Y_i \sim \text{Geom}(p)$ generated an observed dataset $\mathcal{D}$. Please compute the maximum likelihood estimate for p. Include the log likelihood, derivative, and solution when setting the derivative to zero:

   (a) Given a dataset $\mathcal{D} = \{12, 1, 0, 3, 8\}$

   (b) Given a dataset $\mathcal{D} = \{y_1, y_2, \ldots, y_n\}$

15. Assume a set of random variables $A_1, A_2, \cdots, A_n$ such that $A_i \sim \text{Bern}(\theta)$ generated an observed dataset. Please compute the maximum likelihood estimate for p. Include the log likelihood, derivative, and solution when setting the derivative to zero:

   (a) Given a dataset $\mathcal{D} = \{0, 0, 0, 0\}$

   (b) Do you think the estimate you computed above is reasonable? Why or why not?

   (c) Given a dataset $\mathcal{D} = \{a_1, a_2, \ldots, a_n\}$

16. Assume a set of random variables $A_1, A_2, \cdots, A_n$ such that $A_i \sim \text{Pois}(\lambda)$ generated an observed dataset. Please compute the maximum likelihood estimate for p. Include the log likelihood, derivative, and solution when setting the derivative to zero:

    (a) Given a dataset $\mathcal{D} = \{11, 12, 2, 10\}$

    (b) Given a dataset $\mathcal{D} = \{a_1, a_2, \ldots, a_n\}$

17. Assume that we model a sample $(X_1, X_2, \cdots, X_n)$ with a Poisson distribution ($X \sim \text{Pois}(\lambda)$). Please derive the Fisher information $\mathcal{I}(\lambda)$

18. The number of influenza A and influenza B cases each week is recorded by the Pennsylvania Department of health. For the year 2022 the number of cases in Northampton county, PA is Please pose a model for the number of influenza A cases, and compute the maximum likelihood estimate for any parameters in this model. Explain why you chose your particular model and explain parameter estimates. In addition, use the Fisher information about your parameter to build a 95CI around this estimated parameter value.

Flu A	Flu B
100	0
89	0
10	10
45	12
12	98

6

Single covariate linear regression and conditional expected value and variance

thomas mcandrew

Lehigh University

CONTENTS

6.1 Introduction

Regression is likely one of the most important concepts in modern statistical data analysis. Thousands if not tens of thousands scientific analyses depend on regression to understand the relationship between one or more variables and an outcome of interest and to make predictions about future outcomes given a new, not yet seen observation in nature. The regression approach to model building begins with a univariate distribution—often a familiar distribution—and then allows the parameters of that distribution to depend on observed data.

6.2 A new data, sampling setup

Previous to this chapter our typical statistical setup assumed that our data was a set of the form $\mathcal{D} = \{x_1, x_2, x_3, \cdots, x_n\}$ where we collected a single piece of information from each observation. Our statistical model then assumed that these observations were generated from corresponding random variables $X_1, X_2, \cdots, X_n$ that were independent and identically distributed.

Our new setup will consider collecting one additional piece of information from each observation

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \cdots (x_n, y_n)\} \tag{6.1}$$

Our statistical setup will assume that the y data in the second position of each tuple above is the outcome of interest and that this data was sampled from a set of corresponding random variables. We will not consider the x data as being generated from corresponding random variables. The x data is thought of as fixed. Our statistical setup is

$$(x_1, Y_1), (x_2, Y_2), (x_3, Y_3), \cdots, (x_n, Y_n) \tag{6.2}$$

where we assume that Y_i are independent and identically distributed. We will assign to the random variables Y a probabilistic distribution.

Example: We are asked to collect data on patients who are admitted to the hospital to test the hypothesis that those who are hospitalized with influenza are more likely to be under the age of 3 or over the age of 65. A single observation will then contain two pieces of information: (i) whether the patient admitted to the hospital has confirmed influenza and (ii) whether the patient is younger than 3 or older than 65 or in between these ages. Our dataset will look like $\mathcal{D} = \{(r_1, i_1), (r_2, i_2), (r_3, i_3), \cdots, (r_n, i_n)\}$ where the variable r is the value one when a patient is younger than 3 or older than 65 and the value zero otherwise, and the variable i is the value one when the patient was infected with influenza and the value zero otherwise.

6.3 Conditional distributions

We must make concise what we mean when we suppose that values of x impact our outcome of interest Y. Let us define the distribution of the random Y, conditional on the observation x, or $Y|x$, as a probability mass function or density function that produces different probabilities over the support of Y for different values of the observation x.

Example: Assume that Y has a geometric distribution with the following modification

$$P(Y = y|x) = g(x)(1 - g(x))^{y-1} \tag{6.3}$$

where the function $g(x)$ must be between the values 0 and 1 (Why?). We could write the above as $Y|x \sim \text{Geom}(g(x))$ and say that the distribution of Y is conditional on the value of x. When $g(x) = 0.2$ then $P(Y = 5) = 0.08$ and when $g(x) = 0.9$ then $P(Y = 5) = 0.00009$.

6.4 Conditional expectation and variance

For a discrete random variable, the conditional expectation of a random variable $Y|x$ is defined as

$$\mathbb{E}(Y|x) = \sum_{y \in supp(Y)} yp(y|x) \tag{6.4}$$

and for a continuous random variable is defined as

$$\mathbb{E}(Y|x) = \int_{y \in supp(Y)} yf(y|x)\ dy \tag{6.5}$$

where f is the conditional probability density corresponding to $Y|x$.

Because the distribution of Y depends on x then the expected value too will depend on the observed x value. However, we assume that the observation x is not random, the x value is a fixed observation. This assumption makes computing the conditional expectation easier.

Example: Suppose that $supp(Y) = \{-1, 0, 1\}$ and that x can take the values 0 or 1. Further assume the conditional probability distribution Then

$$\begin{aligned}
\mathbb{E}(Y|x = 0) &= \sum_{y \in \{-1,0,1\}} yp(y|x = 0) && (6.6)\\
&= (-1)0.1 + (0)0.50 + (1)0.40 && (6.7)\\
&= -0.1 + 0.40 = 0.30 && (6.8)
\end{aligned}$$

	x=0	x=1
y=-1	0.10	0.40
y=0	0.50	0.30
y=1	0.40	0.30

TABLE 6.1
Conditional distribution of $Y|x$

and

$$\begin{aligned} \mathbb{E}(Y|x=1) &= \sum_{y\in\{-1,0,1\}} yp(y|x=1) && (6.9)\\ &= (-1)0.4 + (0)0.30 + (1)0.30 && (6.10)\\ &= -0.4 + 0.30 = -0.10 && (6.11) \end{aligned}$$

For a discrete random variable, the conditional variance of a random variable $Y|x$ is defined as

$$\mathbb{V}(Y|x) = \sum_{y\in supp(Y)} \left[y - \mathbb{E}(Y|x)\right]^2 p(y|x) \tag{6.12}$$

and for a continuous random variable is defined as

$$\mathbb{V}(Y|x) = \int_{y\in supp(Y)} \left[y - \mathbb{E}(Y|x)\right]^2 f(y|x)\, dy \tag{6.13}$$

where f is the conditional probability density corresponding to $Y|x$.

Example: Lets continue with the above example and compute $V(Y|x)$.

$$\begin{aligned} \mathbb{V}(Y|x=0) &= \sum_{y\in\{-1,0,1\}} [y - \mathbb{E}(Y|x=0)]^2 p(y|x=0) && (6.14)\\ &= (-1-0.30)^2 0.1 + (0-0.30)^2 0.50 + (1-0.30)^2 0.40 && (6.15)\\ &= 0.41 && (6.16) \end{aligned}$$

6.5 Parameters as a function of observations

Single covariate linear regression assumes that given a sample $(Y_1, x_1), (Y_2, x_2), (Y_3, x_3), \cdots, (Y_n, x_n)$ that the conditional distribution of Y given x follows a normal distribution

$$Y|x \sim \mathcal{N}\left(\mu(x), \sigma^2\right) \tag{6.17}$$

where $\mu(x) = \beta_0 + \beta_1 x$, and β_0 and β_1 are parameters that take values in $\mathcal{R}$. The above is sometimes called **the probability form** of linear regression. In probability form we emphasize that our target of interest is modeled as a random variable Y and make explicit that this random variable is expected to be distributed normal.

Because β_0 and β_1 are constant, and because x is a fixed, constant observation then $\mu(x) = \beta_0 + \beta_1 x$ is a constant. Remember that it is easy to translate normal distributions by a constant. Lets choose to add and subtract the constant $(\mu(x))$

The **conditional expectation for single covariate regression** can be computed by recognizing that $\mu(x)$ is a constant,

$$\mathbb{E}(Y|x) = \mu(x) = \beta_0 + \beta_1 x, \tag{6.18}$$

and the **conditional variance for single covariate regression** can also be computed as

$$\mathbb{V}(Y|x) = \sigma^2 \tag{6.19}$$

We see that the variance does not depend on x and remains constant.

$$Y|x \sim \mathcal{N}(\mu(x), \sigma^2) \tag{6.20}$$

$$Y|x = Y|x + \mu(x) - \mu(x) \tag{6.21}$$

Lets define a new random variable $\epsilon = Y|x - \mu(x)$. Then

$$Y|x = \mu(x) + \epsilon \tag{6.22}$$

$$\tag{6.23}$$

We can look closer at the distribution for ϵ The random variable ϵ is a translation of $Y|x$.

$$\epsilon \sim \mathcal{N}(\mu(x), \sigma^2) - \mu(x) \tag{6.24}$$

$$\epsilon \sim \mathcal{N}(\mu(x) - \mu(x), \sigma^2) \tag{6.25}$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{6.26}$$

Finally we see that we can rewrite our linear regression as

$$Y|x = \mu(x) + \epsilon \tag{6.27}$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{6.28}$$

This form for linear regression—the conditional expected value plus an error term (ϵ)—is often called the **model form** for linear regression. Model form emphasizes, $\mu(x)$, often thought of as how you think x and Y are associated with one another.

Linear regression supposes that there exists a linear relationship between the random variable Y and the observations x. One way to investigate whether or not a linear relationship between Y and x appears reasonable is to draw a scatter plot.

Given a dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)\}$ A scatter plot first draws a x and y axis on the cartesian coordinate plane. The second step iterates through each data point (x_i, y_i) and places a dot on the plane at (x_i, y_i). If it appears that a straight line may well explain the association between x and $\mathbb{E}(Y|x)$ then a linear regression model could be fit to this data.

6.6 The residual, diagnostics, and LINE assumptions

Single covariate linear regression is a model to help us understand the relationship between two covariates that we call, y and x. Rarely, if ever, will our model perfectly describe our observed data. In this case, we will need to evaluate how well our model fits our observed data and if single covariate regression is a reasonable, or un-reasonable, model. Evaluating how well our statistical model describes an observed dataset is called **diagnostics**.

Assume that we can find optimal estimates for our parameters β_0, β_1, and σ^2.

6.6.1 Residuals, fitted values, and Normality

The random variable ϵ that we defined for the model form of single covariate linear regression is exact, assuming that we know all of our parameter values. However, in most cases we will not be given the exact parameter values. Instead we are handed a dataset and asked to estimate our parameters. In this case, we will never know the true distribution of our error term ϵ_i that corresponds to a specific x_i value. We can estimate the error term ϵ at the value x given: (i) estimates for our parameters and (ii) a datapoint (x, y) from our dataset. Given a datapoint $d = (x, y)$, a **residual** is the difference between the truly observed y and an estimate of y, $\hat{y}$, or $\hat{\epsilon} = y - \hat{y}$.

Given estimates for our parameters $(\beta_0, \beta_1, \sigma^2)$, a **fitted value** to a datapoint $d = (x, y)$ is $\hat{y} = \hat{\beta_0} + \hat{\beta_1}x$.

Residuals and fitted values can be used to test the assumptions of our regression model. If our regression captures the variability in our dataset well then the residuals $\hat{\epsilon}_i$ should behave similarly to the true error terms $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

6.6.1.1 Linearity

One assumption about single covariate linear regression to evaluate is that the relationship between x and the conditional expected value of Y, $\mathbb{E}(Y|x)$ is linear. Single covariate regression assumes that $\mathbb{E}(Y|x) = \beta_0 + \beta_1 x$ and that $\mathbb{E}(\epsilon) = 0$ for any choice of x.

To test for linearity we can plot for each x value the corresponding residual $\hat{\epsilon} = y - (\beta_0 + \beta_1 x)$. If single covariate regression sufficiently explains the relationship between our random variables Y and fixed data x then we could expect that this plot has residuals values centered around zero for all choice of x. If single covariate regression does not fit the data well then our residuals will not be centered around zero.

Example: Suppose we collect the following dataset and we compute estimates for our parameters $\hat{\beta_0} = 0.88$, $\hat{\beta_1} = 0.64$, and $\hat{\sigma^2} = 0.12$.

x	y
0.44	1.23
0.50	1.19
0.83	1.26
0.41	1.26
0.03	0.87
0.20	0.87
0.98	1.45
0.37	1.28
0.80	1.31
0.99	1.69

TABLE 6.2
Data simulated from the model $y \sim \mathcal{N}(1 + 0.5x, 0.15)$

We can compute 10 fitted values and 10 residuals corresponding to our 10 collected data points

By plotting our fitted values $\hat{y}$, which depend on x, against our residuals we can visualize whether or not our residuals $\hat{\epsilon}$ are centered around zero. If the residuals show a pattern that does not appear to be centered around zero than the relationship between x and Y may not be linear.

x	y	$\hat{y}$	$\hat{\epsilon}$
0.44	1.23	1.16	0.07
0.50	1.19	1.20	-0.01
0.83	1.26	1.41	-0.15
0.41	1.26	1.14	0.12
0.03	0.87	0.90	-0.03
0.20	0.87	1.01	-0.14
0.98	1.45	1.51	-0.06
0.37	1.28	1.12	0.16
0.80	1.31	1.39	-0.08
0.99	1.69	1.51	0.18

TABLE 6.3
Data simulated from the model $y \sim \mathcal{N}(1 + 0.5x, 0.15)$

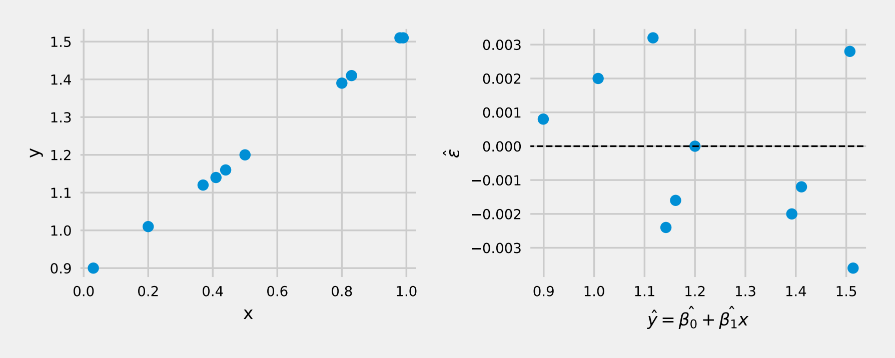


FIGURE 6.1
(Left) Scatter plot of (x, y) pairs from the above dataset. (Right) Scatterplot of residuals versus fitted values.

We see from the plot of x vs y that the relationship between x and y looks linear. In addition, the scatter plot of the fitted values versus residuals also looks like the residuals show not particular pattern and are centered around zero for all fitted values.

6.6.1.2 Equal variance

A plot of fitted values versus residuals can also be used to assess whether the assumption that the error (ϵ) is distributed $\mathcal{N}(0, \sigma^2)$ according to a normal distribution with **the same variance for all values of** x. This assumption is often called the equal variances assumption.

Below we generated 100 (x, y) pairs of observed data points. We estimated our parameter values (more on this later) and then computed fitted values and residuals. Again we plotted a scatterplot of x vs y and a plot of fitted values vs residuals. We see from the residuals vs fitted values that it appears that the residuals move further away from zero as x values increase. This indicates that the **equal variances** assumption may not best describe our data.

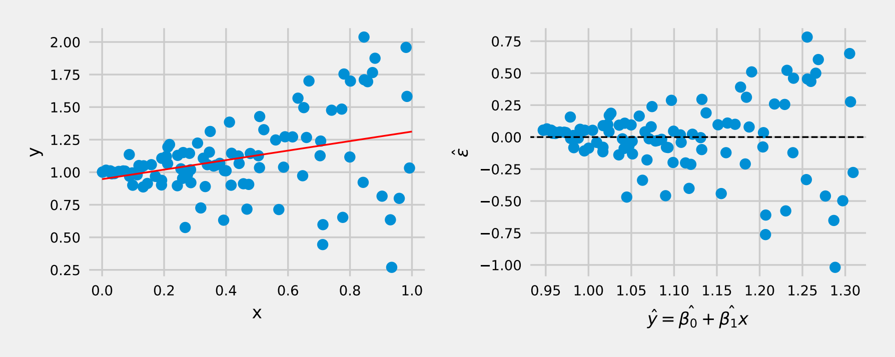


FIGURE 6.2
(Left) Scatter plot of 100 (x, y) pairs from the above dataset with the line (red) $\mathbb{E}(\hat{Y}|x) = \hat{\beta_0} + \hat{\beta_1}x$. (Right) Scatterplot of residuals versus fitted values.

6.6.2 Independence

The independence between data points $d_i = (x_i, y_i)$ and $d_j = (x_j, y_j)$ for any (i, j) can be tested with a sequence vs residual plot. For this plot we graph residuals versus a variable that indicates when in time the data was collected. For example, we may decide to plot the time the data point was collected vs residual, or we can rank the data by the time it was collected and plot the rank vs residual. For single covariate linear regression, we expect that there does not exist a relationship between when the data was collected and the residuals. We assume that errors ϵ are independent from one another. If there is a relationship between when the data was collected and residuals then this may indicate the data points are not independent.

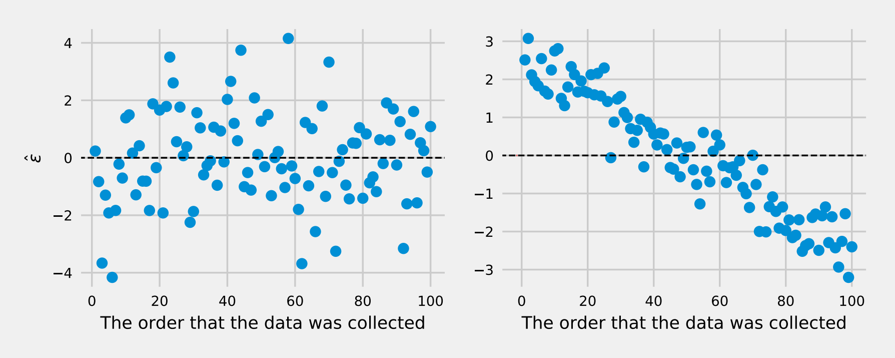


FIGURE 6.3
(Left) Scatter plot of 100 (order, $\hat{\epsilon}$) pairs from a dataset where data was generated independently. (Right) Scatterplot of the order that the data was collected versus residuals for data that was generated dependently.

6.7 Sums of squares

Up until this point we assumed that we can compute estimates for β_0 and for β_1, but we have not yet discussed methods to compute estimates for these parameters. The first method we will explore is called **minimizing the sums of squares** or is often also called **least squares**.

Given a dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)\}$ and a single covariate regression model $Y_i = \mu(x) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$, the least squares estimates for β_0 and β_1 are the values that minimize

$$SS(\beta_0, \beta_1) = \sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{6.29}$$

The function $SS(\beta_0, \beta_1)$ is called the "sums of squares" function. To minimize $SS(\beta_0, \beta_1)$ we will

1. Treat β_1 as a constant, treat β_0 as a variable, take the derivative and set that equation to zero.
2. Treat β_0 as a constant, treat β_1 as a variable, take the derivative and set that equation to zero.
3. Solve the above two equations for β_0 and β_1.

6.7.1 β_0 as the variable and β_1 as a constant

Because the SS function is a sum of similar terms, we can compute the derivative for one term and then add each term together.

$$\begin{aligned} SS(\beta_0)' &= \left\{ \sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_i)]^2 \right\}' \qquad (6.30) \\ &= \sum_{i=1}^{N} \left\{ [y_i - (\beta_0 + \beta_1 x_i)]^2 \right\} \qquad (6.31) \end{aligned}$$

The goal then is to compute the derivative of

$$SS(\beta_0) = [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{6.32}$$

We will need to use the chain rule

$$\begin{aligned} SS(\beta_0) &= [h(\beta_0)]^2 \qquad (6.33) \\ h(\beta_0) &= y_i - (\beta_0 + \beta_1 x_i) \qquad (6.34) \\ & \qquad (6.35) \end{aligned}$$

The derivative of

$$\begin{aligned} SS(\beta_0) &= [h(\beta_0)]^2 \qquad (6.36) \\ & \qquad (6.37) \end{aligned}$$

is

$$\begin{aligned} SS(\beta_0) &= 2\,[h(\beta_0)] \qquad (6.38) \\ & \qquad (6.39) \end{aligned}$$

and the derivative of

$$\begin{aligned} h(\beta_0) &= y_i - (\beta_0 + \beta_1 x_i) \qquad (6.40) \\ & \qquad (6.41) \end{aligned}$$

is

$$h(\beta_0)' = -1, \tag{6.42}$$

and so the derivative of $SS(\beta_0)$ is

$$\begin{aligned} SS(\beta_0)' &= \sum_{i=1}^{N} 2\,[h(\beta_0)] \times (-1) \qquad (6.43) \\ &= \sum_{i=1}^{N} -2\,[y_i - (\beta_0 + \beta_1 x_i)] \qquad (6.44) \end{aligned}$$

We can set this equation equal to zero and solve for β_0 to find the least squares estimate for β_0

$$\sum_{i=1}^{N} -2\left[y_i - (\beta_0 + \beta_1 x_i)\right] = 0 \tag{6.45}$$

$$\sum_{i=1}^{N} \left[y_i - (\beta_0 + \beta_1 x_i)\right] = 0 \tag{6.46}$$

$$\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \beta_0 - \sum_{i=1}^{N} \beta_1 x_i = 0 \tag{6.47}$$

$$\sum_{i=1}^{N} y_i - N\beta_0 - \sum_{i=1}^{N} \beta_1 x_i = 0 \tag{6.48}$$

$$\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \beta_1 x_i = N\beta_0 \tag{6.49}$$

$$\sum_{i=1}^{N} y_i - \beta_1 \sum_{i=1}^{N} x_i = N\beta_0 \tag{6.50}$$

$$\frac{\sum_{i=1}^{N} y_i}{N} - \beta_1 \frac{\sum_{i=1}^{N} x_i}{N} = \beta_0 \tag{6.51}$$

$$\bar{y} - \beta_1 \bar{x} = \beta_0 \tag{6.52}$$

The **least squares estimate for** β_0 is

$$\hat{\beta_0}_{\text{SS}} = \bar{y} - \beta_1 \bar{x} \tag{6.53}$$

6.7.2 β_1 as the variable and β_0 as a constant

Lets take the same approach for computing an estimate for β_1.

$$SS(\beta_1)' = \left\{\sum_{i=1}^{N} \left[y_i - (\beta_0 + \beta_1 x_i)\right]^2\right\}' \tag{6.54}$$

$$= \sum_{i=1}^{N} \left\{\left[y_i - (\beta_0 + \beta_1 x_i)\right]^2\right\} \tag{6.55}$$

The goal then is to compute the derivative of

$$SS(\beta_1) = \left[y_i - (\beta_0 + \beta_1 x_i)\right]^2 \tag{6.56}$$

We will need to use the chain rule

$$SS(\beta_1) = \left[h(\beta_1)\right]^2 \tag{6.57}$$

$$h(\beta_1) = y_i - (\beta_0 + \beta_1 x_i) \tag{6.58}$$

The derivative of

$$SS(\beta_1) = [h(\beta_1)]^2 \tag{6.59}$$

is

$$SS(\beta_1) = 2\,[h(\beta_1)] \tag{6.60}$$

and the derivative of

$$h(\beta_1) = y_i - (\beta_0 + \beta_1 x_i) \tag{6.61}$$

is

$$h(\beta_0)' = -x_i, \tag{6.62}$$

and so the derivative of $SS(\beta_1)$ is

$$\begin{aligned} SS(\beta_1)' &= \sum_{i=1}^{N} 2\,[h(\beta_1)] \times (-x_i) && (6.63) \\ &= \sum_{i=1}^{N} -2x_i\,[y_i - (\beta_0 + \beta_1 x_i)] && (6.64) \end{aligned}$$

We can set this equation equal to zero and solve for β_1 to find the least squares estimate for β_1

$$\begin{aligned} \sum_{i=1}^{N} -2x_i\,[y_i - (\beta_0 + \beta_1 x_i)] = 0 && (6.65) \\ \sum_{i=1}^{N} -x_i\,[y_i - (\beta_0 + \beta_1 x_i)] = 0 && (6.66) \\ \sum_{i=1}^{N} -x_i y_i + x_i\beta_0 + \beta_1 x_i^2 = 0 && (6.67) \\ \sum_{i=1}^{N} -x_i y_i + \beta_0 \sum_{i=1}^{N} x_i + \beta_1 \sum_{i=1}^{N} x_i^2 = 0 && (6.68) \\ \beta_1 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i - \beta_0 \sum_{i=1}^{N} x_i && (6.69) \\ && (6.70) \end{aligned}$$

6.7.3 Solving for β_0 and β_1

At this point we have two equations and two unknowns:

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \tag{6.71}$$

$$\beta_1 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i - \beta_0 \sum_{i=1}^{N} x_i \tag{6.72}$$

$$\tag{6.73}$$

We can solve the above equations and find that the least squares estimates for β_0 and for β_1 are

$$\hat{\beta_1} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{6.74}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x} \tag{6.75}$$

$$\tag{6.76}$$

Example: Suppose we collect the dataset

x	y
0.38	0.72
0.04	1.00
0.79	0.19
0.48	0.69
0.68	0.56

We can compute the least squares estimates for β_0 and β_1 by following these steps:

Compute $\bar{x}$ and $\bar{y}$

$$\bar{x} = \frac{0.38 + 0.04 + 0.79 + 0.48 + 0.68}{5} = 0.47 \tag{6.77}$$

$$\bar{y} = \frac{0.72 + 1.00 + 0.19 + 0.69 + 0.56}{5} = 0.63 \tag{6.78}$$

$$\tag{6.79}$$

Augment the above table with a column that computes $x_i - \bar{x}$ and $y_i - \bar{y}$

x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$
0.38	0.72	-0.09	0.09
0.04	1.00	-0.43	0.37
0.79	0.19	0.31	-0.44
0.48	0.69	0.01	0.05
0.68	0.56	0.21.	-0.07

Add a column that computes $(x_i - \bar{x})(y_i - \bar{y})$ and that computes $(x_i - \bar{x})^2$

x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
0.38	0.72	-0.09	0.09	-0.01	0.01
0.04	1.00	-0.43	0.37	-0.16	0.18
0.79	0.19	0.31	-0.44	-0.14	0.10
0.48	0.69	0.01	0.05	0.00	0.0
0.68	0.56	0.21.	-0.07	-0.02	0.04

Compute the sums of the last two columns and divide the second to last column by the last column to estimate β_1

$$\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = -0.01 - 0.16 - 0.14 + 0.00 - 0.02 = -0.33 \tag{6.80}$$

$$\sum_{i=1}^{N}(x_i - \bar{x})^2 = 0.01 + 0.18 + 0.10 + 0.0 + 0.04 = 0.33 \tag{6.81}$$

$$\hat{\beta_1} = -0.33/0.33 = -1 \tag{6.82}$$

Use your estimate for β_1 to estimate β_0

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x} \tag{6.83}$$

$$\hat{\beta_0} = 0.63 - (-1)(0.47) = 1.1 \tag{6.84}$$

6.8 Maximum likelihood estimates for β_0, β_1, and σ^2

We must estimate three parameters for single covariate linear regression: $\beta_0, \beta_1, \sigma^2$. We will take a maximum likelihood approach to estimate these three covariates in two steps. Step one: compute the log-liklihod for a typical univariate normal distribution. Step two: convert this loglikelihood to single covariate linear regression.

Give a sample $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \cdots, (x_n, y_n)\}$ assume that we model Y as $Y_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$ where μ does not depend on x observations.

Then the likelihood can be computed as

$$\begin{aligned}
\ell(\mu, \sigma^2) &= f(y_1) \cdot f(y_2) \cdot f(y_3) \cdots f(y_n) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{1}{2\sigma^2}\left(y_1 - \mu\right)^2\right] \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{1}{2\sigma^2}\left(y_2 - \mu\right)^2\right] \\
&\cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{1}{2\sigma^2}\left(y_3 - \mu\right)^2\right] \cdots \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{1}{2\sigma^2}\left(y_n - \mu\right)^2\right]
\end{aligned}$$

Lets simplify this equation by computing the log likelihood.

$$\begin{aligned}
\ell\ell(\mu, \sigma^2) &= -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \left[\frac{1}{2\sigma^2}\left(y_1 - \mu\right)^2\right] - \frac{1}{2}\log\left(2\pi\sigma^2\right) - \left[\frac{1}{2\sigma^2}\left(y_2 - \mu\right)^2\right] \\
&- \frac{1}{2}\log\left(2\pi\sigma^2\right) - \left[\frac{1}{2\sigma^2}\left(y_3 - \mu\right)^2\right] \cdots - \frac{1}{2}\log\left(2\pi\sigma^2\right) - \left[\frac{1}{2\sigma^2}\left(y_n - \mu\right)^2\right] \\
&= \sum_{i=1}^{n} -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \left[\frac{1}{2\sigma^2}\left(y_i - \mu\right)^2\right] \\
&= -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \sum_{i=1}^{n}\left[\frac{1}{2\sigma^2}\left(y_i - \mu\right)^2\right] \\
&= -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\left(y_i - \mu\right)^2\right]
\end{aligned}$$

Our log likelihood is a function of two parameters. To maximize for μ, we pretend that our log likelihood is a function of just μ and treat σ^2 as a constant. To maximize for σ^2, we pretend that our log likelihood is a function of σ^2 and treat μ as a constant.

The maximum likelihood estimate (mle) for μ is

$$\hat{\mu}_{\text{mle}} = \sum_{i=1}^{N} y_i / N \tag{6.85}$$

and the mle for σ^2 is

$$\hat{\sigma^2}_{\text{mle}} = \sum_{i=1}^{N} \frac{\left(y_i - \hat{\mu}_{\text{mle}}\right)^2}{N} \tag{6.86}$$

6.9 Maximum likelihood estimate for σ^2 and the MSE

\nobreak

6.10 Homework

1. Are the following two datasets equal to one another? Why or why not?

$$\mathcal{D}_1 = \{(3,1),(0.1,10),(-1,1),(1,0)\} \tag{6.87}$$
$$\mathcal{D}_2 = \{(0.1,10),(1,0),(3,1),(-1,1)\} \tag{6.88}$$

2. Suppose that you are asked to explore in a group of patients who have a family history of heard disease the relationship between the number of packs of cigarettes smoked by the patient and Creatine Kinase-MB (CK-MB), an enzyme found in heart muscle that when elevated indicates damage to the heart. CK-MB is a measurement that is at minimum 0.

   You're first thought is to fit a linear regression model of the form

$$Y_i|x_i \sim \mathcal{N}\left(\beta_0 + \beta_1 x_i, \sigma^2\right) \tag{6.89}$$

   where Y_i is a random variable that we assume generates CK-MB levels for patient i and x_i is the reported number of packs smoked per day.

   (a) What is the support of a Normal distribution? Is this assumption reasonable to model CK-MB?

   (b) Suppose we collect data and estimate β_0 as $\hat{\beta_0} = 200$ and $\hat{\beta_1} = 10$. How would you interpret $\hat{\beta_0}$?

3. Assume that we decide to model a dataset $\mathcal{D} = ((x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n))$ as

$$Y_i|x_i \sim \mathcal{N}\left(\beta_0 + \beta_1 x_i, \sigma^2\right) \tag{6.90}$$

   (a) Let $E = Y_i|x_i - (\beta_0 + \beta_1 x_i)$. What is the distribution of E?

   (b) Let $Z = \frac{Y_i - (\beta_0 + \beta_1 x_i)}{\sigma}$. What is the distribution of Z?

4. We assume that in linear regression

$$Y_i|x_i \sim \mathcal{N}\left(\beta_0 + \beta_1 x_i, \sigma^2\right) \tag{6.91}$$

   the variable x_i is generated randomly. Why or why not?

5. Suppose that we decide to model $\mathcal{D} = (y_1, y_2, y_3, \cdots, y_n)$ as $Y_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$. The loglikelihood function for the normal distribution is

$$\ell\ell(\mu, \sigma^2) = -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(y_i - \mu)^2\right] \tag{6.92}$$

   (a) Please compute the mle for μ by computing the derivative, treating μ as a variable and σ^2 as a constant

   (b) Please compute the mle for σ^2 by computing the derivative, treating σ^2 as a variable and μ as a constant

6. Please compute for all values of q the quantities $\mathbb{E}(R|q)$ and $\mathbb{V}(R|q)$

	q= -1	q=0	q=1
r=-2	0.10	0.90	0.25
r=-1	0.57	0.05	0.25
r= 0	0.13	0.05	0.25
r= 1	0.20	0.00	0.25

7. For the above table, why do the columns sum to the value one but the rows do not?

8. Suppose we collected the following dataset

x	y
-2	-2.83
0	-0.78
2	2.83
1	0.74
0	0.67
1	2.48
2	2.52
3	3.75
-1	-0.91
-2	-1.42

   (a) Please compute all fitted values

   (b) Please compute all residuals

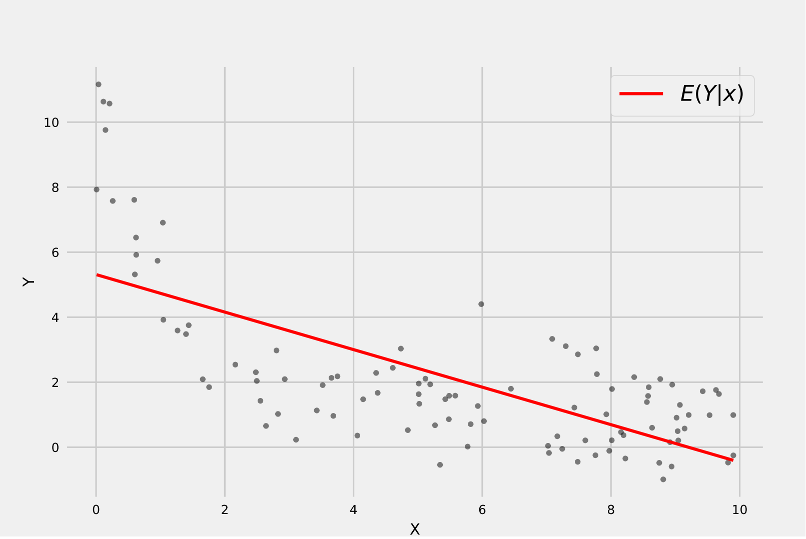
   (c) Plot residuals vs fitted values

   (d) Plot residuals vs the order that the data was collected where the top row was the first collected data point and the last row in the table was the 10th data point collected.

   (e) Does the linearity assumption appear to hold for this dataset? Why or why not?

(f) Does the equal variances assumption appear to hold? Why or why not?

9. Suppose we build a scatter plot of the relationship between observations of a variable x and variable y. We decide to fit a linear regression and add to the plot $\mathbb{E}(Y|x)$.



Please sketch what you would expect the residual vs fitted value plot to look like.

10. Suppose that you fit a linear regression model and compute the residuals. As the residuals move closer to the line $\mathbb{E}(Y|x)$ will your estimate of σ^2 be smaller or larger? Why?

11. A linear regression is fit to a dataset where we collected the number of lbs of sugar an undergraduate student consumes during a week in the Fall semester and their H1Ac level. We estimate our parameters and find that $\hat{\beta_0} = 2$ and $\hat{\beta_1} = 10$. It may (or may not) be more reasonable to express the change in H1Ac not in increments of 1lb but instead in increments of 1 ounce. Please reexpress the estimated change in H1Ac in ounces of sugar consumed.

12. Will different parameter estimates lead to different diagnostic plots? Why or why not?

13. The loglikelihood for single covariate regression is

$$\ell\ell(\beta_0, \beta_1, \sigma^2) = -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{[y_i - (\beta_0 + \beta_1 x_i)]^2\right\} \quad (6.93)$$

We will compute the maximum likelihood estimate for β_1 in a series of steps.

(a) Are only variable of interest is β_1. The paramters β_0 and σ^2 can be thought of as constants. Remove any terms in the loglikelihood that are constants. Call this new function g

(b) Compute the derivative of $g(\beta_1)$, treating β_1 as a variable.

(c) Set $g'(\beta_1) = 0$ and solve for β_1

14. We saw that we can compute estimates for β_0 and β_1 by minimizing the sums of squares or by maximum likelihood. Please describe how the sums of squares estimates and maximum likelihood estimates are related and why they are related.

15. Why can minimizing the sums of squares function $SS(\beta_0, \beta_1)$ be interpreted as finding the values $(\beta_0), \beta_1$ that minimize **the sum** of squared residuals? Why can this minimizatoin be interpreted as minimizing **the mean** of squared residuals?

16. Suppose we collect the dataset

x	y
-0.72	-3.33
-0.78	-2.15
-0.69	-1.94
-0.12	-0.68
0.70	0.20
-0.30	-0.47
-0.15	-2.20
-0.71	-1.37
-0.49	-1.16
0.92	1.77

(a) Compute the maximum likelihood estimate for β_0

(b) Compute the maximum likelihood estimate for β_1

(c) Compute the maximum likelihood estimate for σ^2

17. Suppose we collect the dataset

(a) Compute the maximum likelihood estimate for β_0

(b) Compute the maximum likelihood estimate for β_1

(c) Compute the maximum likelihood estimate for σ^2

x	y
-0.49	-2.20
-0.55	-1.01
-0.46	-0.80
0.11	0.46
0.93	1.34
-0.07	0.66
0.08	-1.07
-0.48	-0.24
-0.26	-0.03
1.15	2.90

7

Matrix calculus

thomas mcandrew

Lehigh University

CONTENTS

7.1 Introduction

Statistics is meant to impart structure to observations in nature and many times the observations that we. are interested in contain many different pieces of information. For example, when studying the relationship between patients randomized in a clinical trial we may be interested in, for each patient, their age, sex, race, body mass index, and whether or not they attain the outcome of interest. Then every observation (a patient) contains five pieces of information. We can think of this observation as living in a single, 5-dimensional space.

Mathematical objects that keep track of values in 1, 2, 3, and higher dimensional spaces are called vectors, and linear maps that take as input one vector and produce a second vector are called matrices. The rules for operating with vectors and matrices is called **matrix calculus**.

Because we wish to structure observations with many pieces of information it is natural that statistics rely on matrix calculus.

7.2 Vectors

7.2.1 Definition

Define the set of all real numbers (positive, negative, and decimal numbers) with the symbol $\mathbb{R}$, and further define the Cartesian product of the set of real numbers with itself as $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$. In general the Cartesian produce of $\mathbb{R}$ with itself N times can be denoted as $\mathbb{R}^N$.

We will informally define a vector of length N as a point in the space $\mathbb{R}^N$. A vector is typically represented as a lowercase letter and the values of the vector are enclosed in square brackets.

Example: For the space $\mathbb{R}^2$ we can define the vector

$$a = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \tag{7.1}$$

or the vector

$$c = \begin{bmatrix} 0.1 \\ -23.01 \end{bmatrix} \tag{7.2}$$

For a space $\mathbb{R}^6$ we could define a vector

$$v = \begin{bmatrix} 0 \\ 1 \\ -8 \\ 3.2 \\ -0.7 \\ 1 \end{bmatrix} \tag{7.3}$$

Vectors can be characterized by their **length**. The **length** of a vector is the number of values that are needed to define it. The vector a and c have a length of 2 and the vector v has a length of 6.

Though it is possible to discuss vectors of length one, more often we give these objects their own name and special properties. A vector of length one is called a **scalar**. Scalars do not need to be enclosed by square brackets.

Example: In any of the spaces R^N for $N > 1$ examples of scalars are 6, -0.4, 0,1, 22, etc.

7.2.2 Addition and subtraction

There are specific rules for how two vectors can interact with one another. You can add and subtract vectors, and you can multiply vectors by a scalar. Addition, subtraction; multiplication and division by a scalar all have a geometric interpretation.

To add two vectors you add the corresponding values in the first, second, third and so on position. You cannot add vectors of different lengths.

Example: Let

$$x = \begin{bmatrix} 1 \\ 0 \\ -2 \end{bmatrix} \tag{7.4}$$

and

$$y = \begin{bmatrix} 0.5 \\ 2.0 \\ -2.0 \end{bmatrix} \tag{7.5}$$

Then

$$x + y = \begin{bmatrix} 1 + 0.5 \\ 0 + 2.0 \\ -2 + (-2.0) \end{bmatrix} = \begin{bmatrix} 1.5 \\ 2.0 \\ -4.0 \end{bmatrix} \tag{7.6}$$

To subtract two vectors you subtract the corresponding values in the first, second, third and so on position. You cannot subtract vectors of different lengths.

$$x - y = \begin{bmatrix} 1 - 0.5 \\ 0 - 2.0 \\ -2 - (-2.0) \end{bmatrix} = \begin{bmatrix} 0.5 \\ -2.0 \\ 0.0 \end{bmatrix} \tag{7.7}$$

Consider $\mathbb{R}^2$, the vector $v = \begin{bmatrix} x \\ y \end{bmatrix}$ can be drawn on the Cartesian coordinate plane by starting at the origin $(0, 0)$ and then drawing a straight line from the origin to the point (x, y).

You can draw the addition of two vectors with a method nicknamed the "tip to tail" draw. Consider two vector labeled x and y. Draw the first vector x starting at the origin. Draw the second vector starting, not at the origin, but at the tip of the first vector. The vector $x + y$ is the vector starting at the origin and ending at the tip of the second vector.

7.2.3 Multiplication and division by a scalar

A new vector v can be built by multiplying or dividing a vector x by a scalar α, $v = \alpha x$. The vector v is built by multiplying each entry in x by α.

$$\begin{aligned} v &= \alpha \cdot x && (7.8) \\ &= \alpha \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} && (7.9) \\ &= \begin{bmatrix} \alpha \cdot x_1 \\ \alpha \cdot x_2 \\ \vdots \\ \alpha \cdot x_n \end{bmatrix} && (7.10) \end{aligned}$$

Geometrically, when you multiply a vector by a scalar that is greater than one the vector is stretched. When diving by that scalar (or multiplying by a scalar between 0 and one) then the vector is compressed. When you multiply a vector by negative one that vector is rotated by 180 degrees. To multiply a vector by a scalar smaller than negative one first rotate the vector 180 degrees and then stretch the vector. A vector multiplied by a scalar between negative one and zero is rotated and then compressed.

7.3 Matrices

7.3.1 Definition

A matrix is a list of vectors and can be thought of as a function that transforms one vector to another. A matrix is often denoted as a capital letter and the values of a matrix are enclosed in square brackets.

Example: We can build a matrix

$$A = \begin{bmatrix} 2 & -1 & 0 \\ 0.5 & 0.1 & 10 \end{bmatrix} \tag{7.11}$$

Entries in a matrix A can be referred to by row number and column number. For example the $(2,1)$ entry in the matrix A is value 0.5. We can also refer to the (2,1) entry in the matrix A as $A_{(2,1)}$

The **shape** of a matrix is a tuple that describes the number of rows of the matrix and the number of columns. The shape of the matrix A above is $(2,3)$.

An important operation—called the **transpose**—interchanges rows and columns of a matrix. Define a matrix

$$B = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 4 & 5 \\ 5 & 1 \end{bmatrix} \tag{7.12}$$

with shape $(4,2)$. Then the transpose of B, labeled B' or sometimes B^t, is the matrix with shape $(2,4)$ and where the first row of B is the first column of B', the second row of B is the second column of B', and so on.

$$B' = \begin{bmatrix} 1 & 2 & 4 & 5 \\ 2 & 3 & 5 & 1 \end{bmatrix} \tag{7.13}$$

If the matrix B has shape $(2,4)$ then the matrix B' has shape $(4,2)$.

We can think of vectors as matrix where the number of rows equals the length of the vector and the number of columns equals one. This allows us to consider both a vector

$$r = \begin{bmatrix} 3 \\ 2 \\ 1 \\ 0 \end{bmatrix} \tag{7.14}$$

and a vector transpose $r = [3,2,1,0]'$. A vector with several rows and one column is often called a **column vector** and a vector with several columns and one row is often called a **row vector**. The vector r is an example of a column vector.

Like vectors, matrices can interact with other matrices and with vectors.

7.3.2 Matrix times vector

If the number of columns of a matrix M is equal to the length of a vector v then M and v can be multiplied together to produce another vector.

$$M \cdot v = q \tag{7.15}$$

If the shape of the matrix M is (n, p) and the vector v is of length p then the new vector q will be of length n.

The values of each entry in q are combinations of individual rows in the matrix M and the vector v.

Let

$$M = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \tag{7.16}$$

and

$$v = \begin{bmatrix} 0 \\ -3 \\ -2 \end{bmatrix} \tag{7.17}$$

To compute the first entry in q we work with the first row of M and the vector v.

$$\begin{align} q_1 &= M_{1,1} \cdot v_1 + M_{1,2} \cdot v_2 + M_{1,3} \cdot v_3 \tag{7.18} \\ &= 1 \cdot 0 + 2 \cdot (-3) + 3 \cdot (-2) \tag{7.19} \\ &= 0 - 6 - 6 = -12 \tag{7.20} \end{align}$$

where q_1 denotes the first entry in q, v_i denotes the i^{th} entry in v and $M_{i,j}$ denotes the (i, j) entry in the matrix M.

To compute the second entry in q we work with the second row of M and the vector v, and so on.

$$\begin{align} q_2 &= M_{2,1} \cdot v_1 + M_{2,2} \cdot v_2 + M_{2,3} \cdot v_3 \tag{7.21} \\ &= 4 \cdot 0 + 5 \cdot (-3) + 6 \cdot (-2) \tag{7.22} \\ &= 0 - 13 - 12 = -25 \tag{7.23} \end{align}$$

The new vector q generated by multiplying M and v is the vector

$$q = \begin{bmatrix} -12 \\ -25 \end{bmatrix} \tag{7.24}$$

We could rewrite, if desired, the above vector q as a scalar times a vector

$$q = -1 \begin{bmatrix} 12 \\ 25 \end{bmatrix} \tag{7.25}$$

In general, given a matrix A with shape (N, P) and a vector b with length P, multiplying Ab produces a new vector c where

$$c_i = \sum_k A_{i,k}\, b_k \tag{7.26}$$

Because a matrix times a vector produces a new vector we can think of a matrix as a function. The input to our function is a vector (such as the vector v) and the output is another vector (such as the vector q above)

Matrices and vectors have special properties when interacting. A matrix can be "distributed" across two vectors Given a matrix M and the vectors q and r, the matrix product

$$M(q + r) = Mq + Mr \tag{7.27}$$

Scalars can be "factored" out of a matrix, vector multiplications. Let α be a scalar, then

$$M(\alpha q) = \alpha\,(Mq) \tag{7.28}$$

7.3.3 Matrix-matrix operations

7.3.3.1 Addition and subtraction

Matrices of the same shape can be added and subtracted by adding (subtracting) corresponding entries.

Given the matrix

$$A = \begin{bmatrix} -1 & 2 & -3 \\ 0 & -1 & 2 \end{bmatrix} \tag{7.29}$$

and the matrix

$$B = \begin{bmatrix} 9 & 8 & 7 \\ 10 & 2 & 0.5 \end{bmatrix} \tag{7.30}$$

then the (i, j) entry of a new matrix $C = A + B$ is the sum of the (i, j) entry in A and (i, j) entry in B or

$$C_{ij} = A_{ij} + B_{ij} \tag{7.31}$$

The (i, j) entry of a matrix $D = A - B$ is the difference of the (i, j) entry in A and (i, j) entry in B or

$$D_{ij} = A_{ij} - B_{ij} \tag{7.32}$$

7.3.3.2 Multiplication

A matrix A can be multiplied by a matrix B to produce a new matrix C with the following formula

$$C_{ij} = \sum_{k} A_{i,k} B_{k,j} \tag{7.33}$$

Example: Given a matrix

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 2 \end{bmatrix} \tag{7.34}$$

and a matrix

$$B = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix} \tag{7.35}$$

The entry in the new matrix C, in the (1,1) position is

$$\begin{aligned} C_{1,1} &= A_{1,1}B_{1,1} + A_{1,2}B_{2,1} && (7.36) \\ &= 0(1) + 1(3) = 3 && (7.37) \end{aligned}$$

$$\begin{aligned} C_{1,2} &= A_{1,1}B_{1,2} + A_{1,2}B_{2,2} && (7.38) \\ &= 0(4) + 1(2) = 2 && (7.39) \end{aligned}$$

$$\begin{aligned} C_{2,1} &= A_{2,1}B_{1,1} + A_{2,2}B_{2,1} && (7.40) \\ &= -1(1) + 2(3) = 5 && (7.41) \end{aligned}$$

$$\begin{aligned} C_{2,2} &= A_{2,1}B_{1,2} + A_{2,2}B_{2,2} && (7.42) \\ &= -1(4) + 2(2) = 0 && (7.43) \end{aligned}$$

$$C = \begin{bmatrix} 3 & 2 \\ 5 & 0 \end{bmatrix} \tag{7.44}$$

The above operation for computing an entry in C occurs frequently in matrix calculus. This operation is called an **inner product**. The inner product of the vector v of length n and the vector q of length n is equal to

$$v'q = \sum_{k=1}^{n} v_k q_k \tag{7.45}$$

We can think of the matrix A as consisting of 2 row vectors, the vectors $A_{1,:} = [0, 1]$ and $A_{2,:} = [-1, 2]$, and the matrix B as consisting of two column vectors

$$B_{:,1} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \tag{7.46}$$

and

$$B_{:,2} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \tag{7.47}$$

Then the matrix $C_{i,j} = A_{i,:}B_{:,j}$ or

$$C = \begin{bmatrix} A'_{1,:}B_{:,1} & A'_{1,:}B_{:,2} \\ A'_{2,:}B_{:,1} & A'_{2,:}B_{:,2} \end{bmatrix} \tag{7.48}$$

Important: The matrix product AB is not necessarily equal to the matrix product BA. For matrix multiplication, the order of multiplication matters.

7.3.4 The matrix inverse

There exists a special matrix called **identity matrix**. The identity matrix I of size n is a matrix with n rows and n columns with ones along the diagonal entries (the entries $I_{k,k}$ for k from 1 to n) and then zeroes for all other entries.

$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{7.49}$$

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{7.50}$$

A matrix A has an inverse B if (i) A has the same number of rows as columns and (ii)

$$AB = I \tag{7.51}$$

$$BA = I \tag{7.52}$$

A matrix B that satisfies the above properties is called the inverse of A and is denoted A^{-1}.

A matrix inverse is similar to an inverse in one dimensional algebra. In 1d algebra an inverse b for the variable a satisfies

$$ab = 1 \tag{7.53}$$

$$ba = 1 \tag{7.54}$$

and the variable b is denotes a^{-1} or $\frac{1}{a}$. Matrix with a different number of rows and columns will never have an inverse.

7.4 Multivariate Linear regression

Multivariate regression models are ubiquitous in science because of their ability to relate two or more measurable quantities to a target variable.

Multivariate linear regression supposes that many different covariates $(x_1, x_2, \cdots, x_p)$ may be able to describe the distribution of a random variable Y.

Multivariate linear regression extends single covariate regression, and we will see below that adding more covariates makes estimating, as well as communicating, parameters more difficult.

7.4.1 Data and model setup

Multivariate linear regression supposes that an experiment generates a single data point that contains one measurement of our target of interest and two or more measurements of covariates $(x_1, x_2, \cdots)$ that may help explain our target.

A single observation takes the form

$$d = (y_1, x_1, x_2, x_3, \cdots, x_p) \tag{7.55}$$

where y_1 is our target and $x_1, x_2, \cdots, x_p$ are p covariates that are measured at the same time as our target.

A dataset containing n observation then can be written as

$$\mathcal{D} = ((y_1, x_{1,1}, x_{1,2}, x_{1,3}, \cdots, x_{1,p}), \tag{7.56}$$
$$(y_2, x_{2,1}, x_{2,2}, x_{2,3}, \cdots, x_{2,p}), \cdots, (y_n, x_{n,1}, x_{n,2}, x_{n,3}, \cdots, x_{n,p}) \tag{7.57}$$

The model for multivariate linear regression assumes

$$Y_i | x_{i,1:p} \sim \mathcal{N}\left(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots, \beta_p x_{i,p}, \sigma^2\right) \tag{7.58}$$

where we assume that the observation y_i was generated from a corresponding random variable Y_i with the above Normal distribution conditioned on all x covariate values.

The above model can also be expressed as

$$Y_i | x_{i,1:p} = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots, \beta_p x_{i,p} + \epsilon_i \tag{7.59}$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{7.60}$$

Compared to single covariate linear regression, multivariate linear regression begins with $\beta_0 + \beta_1 x_{i,1}$ as in single covariate linear regression and then further adjusts the mean by multiplying each additional covariate $x_{i,k}$ with a corresponding β_k and summing all of these quantities $\beta_k x_{i,k}$.

7.4.2 A single categorical variable

A natural experiment where one will need to use multivariate regression is when associating a continuous target variable y to a categorical explanatory variable x that can take one of C possible categories.

Suppose that we wish to relate y, a continuous variable, to x, a variable that can take one of the three values 'A', 'B', 'C' then linear regression does not—at first—make much sense.

Given the dataset

$$\mathcal{D} = \begin{bmatrix} y & x \\ 0.2 & ''A'' \\ 0.4 & ''A'' \\ -0.5 & ''B'' \\ 0.6 & ''C'' \\ -0.5 & ''B'' \end{bmatrix} \tag{7.61}$$

The naive model would assume that

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{7.62}$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{7.63}$$

However, we cannot multiply the value β_1 by the letters A, B, C.

One solution is to convert this single categorical variable into many numerical variables. We will need to (i) choose a **reference** value for x that we will compare all other y values corresponding to other categories to and (ii) create two additional columns—one per category excluding the reference—that contain the value one when that category value is present in x and the value zero otherwise.

For example, suppose for the above dataset we choose a reference of "B". Then we will create one column that identifies values of "A" and a second column that identifies values of "C".

Our new dataset will look like

$$\mathcal{D} = \begin{bmatrix} y & x & x_A & x_c \\ 0.2 & ''A'' & 1 & 0 \\ 0.4 & ''A'' & 1 & 0 \\ -0.5 & ''B'' & 0 & 0 \\ 0.6 & ''C'' & 0 & 1 \\ -0.5 & ''B'' & 0 & 0 \end{bmatrix} \tag{7.64}$$

We can update our naive single covariate linear regression with the following multivariate linear regression.

$$Y_i = \beta_0 + \beta_1 x_{i,A} + \beta_2 x_{i,C} + \epsilon_i \tag{7.65}$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{7.66}$$

To interpret the above model, lets compute the expected value of Y when we observe a "A", "B", and a "C".

When we observe an "A" then the covariate x_A is the value one and the covariate x_c is the value zero. So then

$$\mathbb{E}(Y|x = A) = \beta_0 + \beta_1(1) + \beta_2(0) \tag{7.67}$$
$$= \beta_0 + \beta_1 \tag{7.68}$$

When we observe a "B" then the covariate x_A is the value zero and the covariate x_c is the value zero. So then

$$\mathbb{E}(Y|x = B) = \beta_0 + \beta_1(0) + \beta_2(0) \tag{7.69}$$
$$= \beta_0 \tag{7.70}$$

When we observe a "C" then the covariate x_A is the value zero and the covariate x_c is the value one. So then

$$\mathbb{E}(Y|x = C) = \beta_0 + \beta_1(0) + \beta_2(1) \tag{7.71}$$
$$= \beta_0 + \beta_2 \tag{7.72}$$

The intercept term (β_0) is the expected value of Y when we observe a x value that is equal to the reference. All other categories include this reference β_0. The parameter β_1 is the change in the expected value of Y when an observation includes $x = A$ **compared to the expected value when x=B**. The parameter β_2 is the change in the expected value of Y when an observation includes $x = C$ **compared to the expected value when x=B**.

7.4.3 Two or more continuous covariates and adjustment for confounding

We can include several continuous valued x covariates without the need to first transform them as we must do when we include a x covariate that is categorical. For example, suppose that we wish to associate the annual cost of household health insurance to annual family income, number of dependents that are claimed on previous taxes by the individual who pays the insurance, and age of the oldest resident of the household.

Then we can generate a model

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \epsilon_i \tag{7.73}$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{7.74}$$

and estimate the parameters $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2)$. Standard software can produce point estimates, confidence intervals, and hypothesis tests at a desired false positive rate, but software can not interpret the hypothesized impact of these parameters on Y.

One method to understand whether a covariate x is associated with Y while adjusting for all other covariates that are considered is to build an **added variable plot** between Y and x.

Suppose our experiment generates three covariates (x_1, x_2, x_3) and a target y. An added variable plot for the covariate x_3 follows these steps:

- Fit a linear regression to $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1 + \epsilon_{Y|x3}$
- Fit a linear regression to $X_3 = \beta_0^* + \beta_1^* x_1 + \beta_2^* x_1 + \epsilon_{X_3}$
- Plot the pairs ($\epsilon_{Y|x3}, \epsilon_{X_3}$)
- Fit the regression $\epsilon_{Y|x3} = \beta_3 \epsilon_{X_3}$ and plot the expected value.

The residuals $\epsilon_{Y|x3}$ include any relationships that are not captured by associating Y, x_1, and x_2 linearly. The residuals ϵ_{X_3} are values that linearly relating x_1 and x_2 could not capture, and so when we relate $\epsilon_{Y|x3}$ to ϵ_{X_3} we are asking about the association between patterns in Y that could not be captured by x_1 and x_2, and patterns in x_3 that could not be explained by x_1 and x_2.

7.4.4 Estimating β and σ

\nobreak

7.5 Homework

1. Please compute the following
   (a) $v + x$
   (b) $v - x$
   (c) $2v + x/2$ where

$$v = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \quad x = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}$$

2. Let

$$A = \begin{bmatrix} 1/2 & -2 & 0 \\ -10 & -5 & -1 \end{bmatrix} \; B = \begin{bmatrix} 1 & 0 \\ 0 & -1 \\ 1 & 1 \end{bmatrix} \; x = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}$$

If possible, please compute the below. If it is not possible to compute one of the below quantities then please write "Cannot compute" and explain why.

   (a) AB

   (b) BA

   (c) Ax

   (d) Bx

   (e) BAx

3. Suppose that we collect data on the total number of confirmed influenza hospitalizations over the 2022/2023 influenza season and hypothesize that the total number of cases may be related to population density of the state and the percent of the state population that is vaccinated before October 1, 2022.

   Please describe a multivariate linear regression that may be used to model this data. Please include: (i) the model, a description of each variable in the dataset, a summary of what the vector y and matrix X might look like.

4. Let $\beta' x$, where

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \; ; x = \begin{bmatrix} x_3 \\ x_2 \\ x_1 \\ x_0 \end{bmatrix}$$

   Please expand $\beta' x$ as a series of sums and products.

8

Logistic regression

thomas mcandrew

Lehigh University

CONTENTS

8.1 Introduction

Logistic regression is used to model a target that reports the presence (often mapped to the value one) or absence (often mapped to the value zero) of a phenomena.

8.2 Data setup

Compared to multivariate linear regression, we assume an identical setup for the data except for one change. Assume that we collect observations with p pieces of information and a target of interest y. Then we may describe our dataset with n observations as

$$\mathcal{D} = \big((y_1, x_{1,1}, x_{1,2}, \cdots, x_{1,p}), (y_2, x_{2,1}, x_{2,2}, \cdots, x_{2,p}), \cdots, (y_n, x_{n,1}, x_{n,2}, \cdots, x_{n,p})\big) \tag{8.1}$$

The key difference between the data for linear regression and the data for logistic regression is that for linear regression the target variable y is assumed to take any negative or positive values. However, for logistic regression we assume the only two options for y are 0 or 1.

8.2.1 Modeling the target

Because the target takes the value 0 or 1 a natural choice to model this target is the Bernoulli distribution. That is, we start by assuming that the random variables Y_i that generate observations y_i follow a Bernoulli distribution.

$$Y_i \| x_{i,1}, x_{i,2}, \cdots, x_{ip} \sim \text{Bernoulli}(\theta) \tag{8.2}$$

8.2.2 A function to constrain the parameter space

A first choice for the function above may be the same as linear regression

$$\theta(x_i) = \beta_0 + \sum_{k=1}^{p} \beta_k x_{i,k} \tag{8.3}$$

However, because $\theta(x)$ is the **probability** assigned to $Y = 1$ the function $\theta(x)$ needs to be confined in the interval $[0, 1]$. The function above is not limited to the interval [0,1] and so we need to find a new function that guarantees that $\theta(x)$ is between 0 and 1.

8.2.3 The full model

\nobreak

8.3 One unit change

\nobreak

8.4 Log odds and odds ratio

\nobreak

8.5 Decision boundaries

\nobreak

8.6 Homework

1. Consider the following logistic regression model

$$Y_i \sim \text{Bern}[\theta(x_i)] \tag{8.4}$$

$$\theta(x) = e^{\beta_0+\beta_1 x_i} \Big/ (1 + e^{\beta_0+\beta_1 x_i}) \tag{8.5}$$

   Please derive the change in the log odds for a k unit change in x,

2. Suppose you find estimates $\beta_0 = 0.5$ and $\beta_1 = -1$ for the following logistic regression model

$$Y_i \sim \text{Bern}[\theta(x_i)] \tag{8.6}$$

$$\theta(x) = e^{\beta_0+\beta_1 x_i} \Big/ (1 + e^{\beta_0+\beta_1 x_i}) \tag{8.7}$$

   (a) What does the model assume is the probability of obtaining $y = 1$ when $x = 0$ (i.e. $P(Y = 1|x = 0)$)?

   (b) What does the model assume is the probability of obtaining $y = 1$ when $x = 2$ (i.e. $P(Y = 1|x = 2)$)?

   (c) What does the model assume is the probability of obtaining $y = 1$ when $x = 3$ (i.e. $P(Y = 1|x = 3)$)?

   (d) What does the model assume is the probability of obtaining $y = 1$ when $x = 4$ (i.e. $P(Y = 1|x = 4)$)?

   (e) Why is the change in the probability of a one from $x = 2$ to $x = 3$ (i.e. $P(Y = 1|x = 3) - P(Y = 1|x = 2)$) different compared to the change in the probability of a one from $x = 3$ to $x = 4$ ($P(Y = 1|x = 4) - P(Y = 1|x = 3)$)?

3. Suppose the following model

$$Y_i \sim \text{Bern}[\theta(x_i)] \tag{8.8}$$

$$\theta(x) = e^{\beta_0+\beta_1 x_{i,1}+\beta_2 x_{i,2}} \Big/ (1+e^{\beta_0+\beta_1 x_i}) \tag{8.9}$$

   (a) Please write down the conditional expected value for Y, $\mathbb{E}(Y|x_1, x_2)$

   (b) Please write down the conditional variance for Y, $\mathbb{V}(Y|x_1, x_2)$

4. You are asked to build a model to summarize the association between diabetes type II status (the target) and three covariates: age, in years, sex (Male/Female), and fasting glucose level (mg/dL) for patients at a local hospital.

   The dataset contains observations on 1000 patients. You build a model

$$\log\left[\frac{\theta(x)}{\theta(1-x)}\right] = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} \tag{8.10}$$

   where $x_{i,1}$ is age, $x_{i,2}$ is sex with a reference level male, and $x_{i,3}$ is fasting glucose level.

   When you estimate the parameters you find that

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
-0.5	1.2	-0.9	-0.75

   (a) Convert the above estimates to odds ratios

   (b) Please interpret a one unit change in age, sex, and fasting glucose level.

   (c) Write down the equation that describes a decision boundary if we assign a 1 when the probability of a 1 is greater than 0.50 and a 0 otherwise.

Part III

Algorithms lab

9

Laboratory 01

thomas mcandrew, david braun

Lehigh University

CONTENTS

9.1 Jupyter Notebooks

All of our lab work will take place in Jupyter Notebooks. Jupyter Notebooks are a tool for organizing textual descriptions of work and computer programs. The goal is to produce one document to communicate a set of scientific ideas and allow another to understand exactly how you arrived at yoru conclusions.

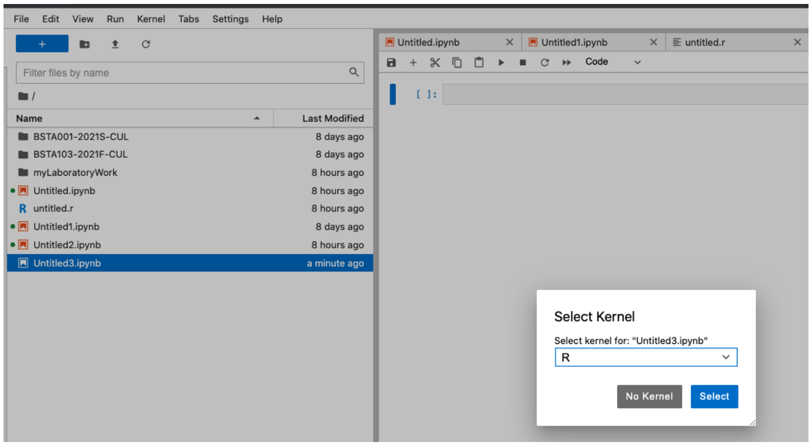
Jupyter has some important buttons.

9.1.1 File

\nobreak

9.1.1.1 A new notebook

Under file->New->Notebook you can create a new notebook. When asked to "Select Kernel" click on the drop down menu and select "R"



Name	Last Modified
BSTA001-2021S-CUL	8 days ago
BSTA103-2021F-CUL	8 days ago
myLaboratoryWork	8 hours ago
Untitled.ipynb	8 hours ago
untitled.r	8 hours ago
Untitled1.ipynb	8 days ago
Untitled2.ipynb	8 hours ago
Untitled3.ipynb	a minute ago

Select Kernel

Select kernel for: "Untitled3.ipynb"

R

FIGURE 9.1
kernelselect.png

9.1.1.2 The notebook

A notebook is a collection of cells. A **cell** is a container that can hold text or computer code. Cells in Jupyter looks like gray rectangles. There are three cell types in Jupyter: (i) Code, (ii) Markdown, and (iii) Raw. The two that we will focus on are Code and Markdown.

The “Code” cell holds computer code that the R kernel (see below about a kernel) can use to compute. We may want to import data, run a statistical analysis, and output results. This is for the “Code” cell.

“Markdown” is itself a special language that a Jupyter Notebook interprets as text. The “Markdown” cell is most useful for write ups, descriptions of a Code cell above or below, or scientific conclusions, comments, and thoughts. When you need to write, think Markdown.

9.1.1.3 Save your work

You can always save your work, and should do so often, by clicking File -> Save Notebook.

9.1.1.4 Export for submission

In class, we will ask that you submit your work on Coursesite as a **PDF**. Work in another format will not be accepted. To export your notebook as a PDF, choose File->Save and Export Notebook As->PDF

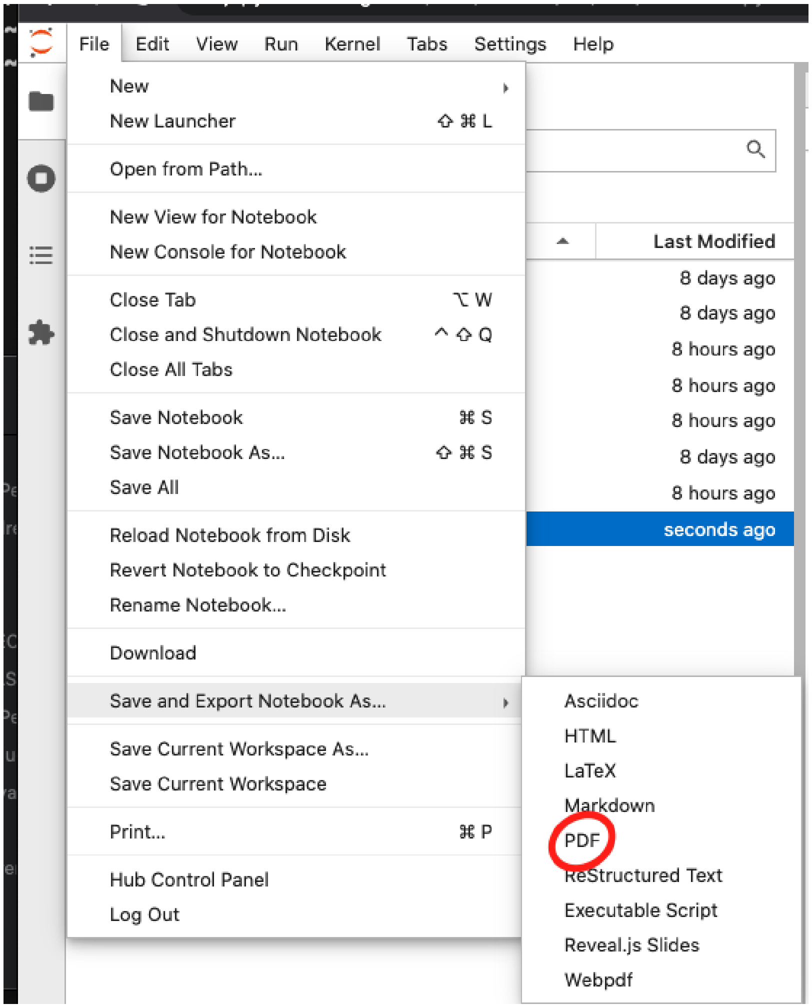


FIGURE 9.2
Caption

After you click PDF, a PDF file will be created and saved in a "Downloads" folder on your local machine. Make sure the PDF file contains (1) Your first and surname, the date, and a descriptive title.

9.1.2 Kernel

The kernel is the component that executes code inside your notebook. No kernel, no running code.

Over the course, you may find that your notebook has disconnected or otherwise will no longer execute the code you wrote. Most often, the kernel has stopped. To restart you kernel select Kernel->Restart Kernel.

9.2 Programming and R

The R programming language, while not explictly written for statistics, has a long history as a tool for data analysis, statistics, machine learning, and data science. R supports all of the main paradigms in computing and you will be able to transfer what you learn in R to other programming languages without much difficulty.

Programming is difficult. Like any skill, programming take time to master. Error messages will be commonplace, you will find it difficult to ask the computer to calculate what you want. You will be frustrate and that is ok. Over time you will learn to read the error messages, code will flow more easily. The most important part of programming is daily practice.

When we **execute** code, we ask the computer to translate what we wrote into binary and return a set of results that may or may not be stored in memory. In the Jupyter environment we execute code by pressing "Run" or by using the shortcut "Shift+Enter".

9.3 Arithmetic

R supports all standard artithmetic calculations. Lets "Run" our first computation.

R can interpret addition

```
[100]: 2+2
```

4

Subtraction

```
[101]: 9-3
```

6

Division

```
[102]: 3/4
```

```
0.75
```

multiplication

```
[103]: 4*4
```

```
16
```

and exponentiation

```
[104]: 3^9
```

```
19683
```

As expected, we can compute more difficult arithemetic expressions.

```
[105]: (2^4)+3/2 - 1
```

```
16.5
```

9.4 Vectors

The **vector** is the fundemental object in R.

A mathematical vector is an ordered list of numbers. They are denoted by a sequence of numbers surrounded by square brackets.

$$v = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \tag{9.1}$$

Above, the vector **v** is a vector of length 3 and contains, in order, the values 1, 2, and 3.

In R, vectors are goven a name and stored in the computer in one of two ways: (i) using the **c()** operator or (ii) using the assign function.

9.4.1 Assignment

\nobreak

9.4.1.1 c()

We can store a vector named v with the values 1,2,3 in R as follows

```
[106]: v = c(1,2,3)
```

9.4.1.2 assign

We can also use the assign function to store a vector, named q, with the values 3,2,1 as follows

```
[107]: assign("q",c(3,2,1))
```

9.4.1.3 equals

The equals sign **does not** represent two objects are equal to one another. The equals sign in compiuter programming stands for "assign".

When we write `v = c(1,2,3)`, this is understood as "we assign the variable v to the vector (1,2,3). As an example, lets create a vector `(4,5,6)` names `x` and then assign the variable `y` to be the same as `x`

```
[108]: x = c(4,5,6)
       y = x
```

The last line above does not ask whether or not `x` is the same as `y`. Instead, this line assigns the variable `y` to be the same vecor as `x`.

9.5 Print

When we created the vectors **v** and **q** "nothing happened". Though the vector v and q were created and stored in the computer, R does not display these on your screen by default. One way to view any object in R is to print it.

You can print an object, `x`, R by writing `print(x)`

```
[109]: print(v)
```

```
[1] 1 2 3
```

```
[110]: print(q)
```

```
[1] 3 2 1
```

[111]:

```
print(x)
print(y)
```

```
[1] 4 5 6
[1] 4 5 6
```

You do not need to print any object, ever. Printing is not necessary. You should use print to explore whether you programmed something write or to communicate scientific results.

9.6 Combining vectors

We can append one vector to another in R by using the c() operator. Suppose we wish to combine the two vectors

$$x = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \; ; \; z = \begin{bmatrix} -1 \\ 0.2 \\ 90 \end{bmatrix} \tag{9.2}$$

into one vector

$$r = \begin{bmatrix} 1 \\ 2 \\ 3 \\ -1 \\ 0.2 \\ 90 \end{bmatrix} \tag{9.3}$$

Lets first create the vectors `x` and `z`

[112]:

```
x = c(1,2,3)
z = c(-1,0.2,90)
```

Now we can create the vector `r`

[113]:

```
r = c(x,z)
```

If we want to check our work, we can print out `r`.

[114]:

```
print(r)
```

```
[1]  1.0  2.0  3.0 -1.0  0.2 90.0
```

9.7 Indexing and access

Vectors are useful for storing several different numbers. We can access single elements, or several elements inside a vector by (i) naming the vector we want to access, (ii) typing square brackets “[]”.

9.7.1 Numeric indexing

If we want to access the 4th element in `r`, we can type

```
[115]: r[4]
```

-1

If we want to access, the 2nd, 4th, and then the first element of `r` we can include in square brackets the vector `c(2,4,1)`

```
[116]: r[c(2,4,1)]
```

1. 2 2. -1 3. 1

We can access the 1st,2nd, and 3rd elements in `r` using the vector `c(1,2,3)`, however a shortcut is to use the **colon** operator. The colon operator takes as input two integers (a,b) separated by a colon (a:b) and expands to the vector `c(a,a+1,a+2,a+3,...,b)`.

Watch

```
[117]: z = 3:5
```

```
[118]: print(z)
```

```
[1] 3 4 5
```

The colon operator is useful for accessing items in a vector

```
[119]: r[2:5]
```

1. 2 2. 3 3. -1 4. 0.2

The above is called **numeric indexing**. Numeric indexing is the access of elements in an object (here a vector) by inputting a single number, or vector of numbers. Indices are always integers. Fractional or decimal numbers cannot be used as indices. Up until now we only used *positive* integers to access elements of a vector.

R also accepts negative integers as indices. Warning: R handles negative indices different than the majority of other progamming languages. A negative index in R standard for **exclude**.

For example, if we want to return all the elements of a vector `q = c(1,4,6,10,0.5)` except for the 2nd element, we can write `q[-2]`

[120]:
```
q = c(1,4,6,10,0.5)
q[-2]
```

1. 1 2. 6 3. 10 4. 0.5

9.7.2 Logical vectors and logical indexing

\nobreak

9.7.2.1 True and False

R, like all programming languages, understands how to operate with binary logic (aside: binary logic is not the only type. If interested, google the tetralemma). True in R is represeted as the word `TRUE` in all capitals. False in R is represented as the word `FALSE` in all capitals. The symbols `TRUE` and `FALSE` are reserved, special symbols in R. You cannot assign a variable to `TRUE` or `FALSE`.

[121]:
```
TRUE
```

TRUE

[122]:
```
FALSE
```

FALSE

9.7.2.2 Logical comparisons

R understand the following logical operators: - > "Greater than" - >= "Greater than or equal to" - < "Less than" - <= "Less than or equal to" - == "Is equal to" - != "Not equal" - `|` "OR" - `&` "AND"

9.7.2.3 Logic

Logic is a method to evaluate statements, sometimes called propositions as either True or False. The above symbols are used to evaluate statements.

When you pose a proposition to R, such as `v > -1` R will evaluate that propositon for each individual element in the vector `v`. Lets create the vector `v = c(-10,10,4)` and ask R to evaluate the proposition `v>-1`.

```
[123]: v = c(-10,10,4)
       v > -1
```

1. FALSE 2. TRUE 3. TRUE

We see that R returns a vector with the same numebr of elements as in v containing the values TRUE or FALSE. A vector that contains values TRUE/-FALSE is called a **logical vector**. Like any other vector we can store a logical vector.

```
[124]: log = v>-1
```

```
[125]: print(log)
```

```
[1] FALSE  TRUE  TRUE
```

9.7.2.4 AND, OR, and NOT

AND, OR, and NOT are **logical operators**, they allow us to combine one or more propositions. Given two propositions p_1 and p_2, the AND, OR, and NOT operator will evaluate to the following

p_1	p_2	p_1 AND p_2	p_1 OR p_2	$!p_1$
TRUE	TRUE	TRUE	TRUE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE
FALSE	TRUE	FALSE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	TRUE

TABLE 9.1
Caption

Logical operators come in handy in **logical indexing**. When we write `r[l]` where `l` is a logical vector, R will return the the values in `r` where `l` is TRUE.

```
[126]: r = c(-10,0,10,55,0.34,-0.97)
       l = c(TRUE,FALSE,TRUE,FALSE,TRUE,TRUE)

       r[l]
```

1. -10 2. 10 3. 0.34 4. -0.97

More often we will include the logical statement directly inside the square brackets

```
[127]: r[ r>0 ]
```

1. 10 2. 55 3. 0.34

9.7.2.5 Equivalence of TRUE and FALSE to 1 and 0

The symbol `TRUE` in R is understood to be the same as the value `1`, and the symbol `FALSE` in R is understood to be the same value as `0`.

[128]:
```
TRUE==1
```

TRUE

[129]:
```
FALSE==0
```

TRUE

[130]:
```
TRUE==0
```

FALSE

[131]:
```
FALSE==1
```

FALSE

9.7.3 Two functions that are useul for operating on vectors

Functions in mathematics take as input a list of objects and return a unique object. The same is true of functions in programming and so in R.

We can create our own functions in R (this will come later), but R also has a large library of built-in functions that are automatically included once you start R. Two very sueful ones are the `sum` function and the `length` function.

The `sum` function takes as input a vector and returns the sum of each element in the function.

[133]:
```
v = c(3,2,1)
sum(v)
```

6

The `length` function takes as input a vector and returns the number of elements in the vector

[134]:
```
length(v)
```

3

9.8 Assignment 01

We are recruited to track the evolution of an infectious agent for a team of public health officials (PHOs). To support future strategic planning the PHOs want to know the impact of intervention X on increases or decreases in the incidence of this infectious agent. The PHO team collected for each county in their state whether the intervention was enacted, and whether the incidence of case counts of this infectious agent increased or decreased 60 days after the intervention was in place data.

Using R, we will assign probabiltiies to the four events in our sample space { (intervention,raise),(intervention, no raise),(no intervention, raise),(no intervention, no raise) }

9.8.1 The data

In the below cell there is a few lines of code pre-programmed. Please runs this cell below.

This cell will create two vectors.

The first vector is called `intervention_rise` and contains one element for each county that has intervention X collected by the PHO team. An element is the value `1` if there was a rise in incidence for the infectious agent and `0` if there was a fall in incidence.

The second vector is called `nointervention_rise` and contains one element for each county that did not have intervention X collected by the PHO team. An element is the value `1` if there was a rise in incidence for the infectious agent and `0` if there was a fall in incidence.

9.8.2 Please complete the following

1. Use the `length` function to count the number of counties where intervention X took place
2. Use the `length` function to count the number of counties where no intervention X took place
3. Use the `sum` function to count the number of counties that observed a rise in incidence.
4. Use the `sum` function and the `not (!)` operator to count the number of counties that observed a fall in incidence.
5. Use the frequentist approach to compute the probability
   - that an intervention would take place in a county
   - that a rise in incidence was observed in a county
   - of a rise in incidence given a county implemented intervention X

  - of a rise in incidence given a county has not implemented intervention X

6. Use the multiplication rule to compute the probability
   - that an intervention and rise is observed (Hint: P(Intervention) * P(Rise | Intervention))
   - that an intervention and fall is observed
   - that no intervention and rise is observed
   - that no intervention and fall is observed
7. Compute the probability of the below events, assuming intervention and rise/fall. are independent
   - Intervention and Rise
   - Intervention and Fall
   - No Intervention and Rise

   - No Intervention and Fall
8. Do you think intervention X is effective at preventing the spread of our infectious agent?

[178]:

```
#RUN THIS CODE. DO NOT WORRY WHAT IT SAYS.
nums = runif(10^3,0,1)

intervention_rise   = c()
nointervention_rise = c()
for (i in nums){
    if (runif(1)> 0.4){
        if ( i>0.800 ){
          risefall = 1
        } else{risefall=0}
        intervention_rise = c(intervention_rise, risefall)
    }
    else {
        if ( i>0.325 ){
          risefall = 1
        } else{risefall=0}
        nointervention_rise = c(nointervention_rise, risefall)
    }
}
```

10

Laboratory 02

thomas mcandrew, david braun

Lehigh University

CONTENTS

10.0.1 Control flow

Now that we defined some fundemental objects and operations in R, we need to explore how R, and many other programmign language execute code. R executes code sequentially from top to bottom. The line of R code at the top is run first, then the second highest line, then the third highest and so on.

When code is executed line by line from top to bottom it is called **sequential control**. Sequentuntial control if the default way R executes statements.

However, we can change the order in which R executes lines of code in three ways: (i) choice (ii) repetition (iii) functions (we'll talk much more about functions next week).

10.0.1.1 Choice or Selection

An R **program** is a set of statements meant to produce one or more results. We can choose to execute all the lines in our program, or we can choose to execute some lines and not others depedent on conditions. When we execute some lines of R code to run and not others we are using a specific type of control flow called **Selection** ("we select some lines and not others").

While we often define statements as sentences that are either TRUE or

FALSE, in R we define a logical statement that evaluates to TRUE or to FALSE an **expression**.

The main ways to control which lines to execute is with the if/else, if/elsif/else, and the switch statement.

If/else

The following study at this link = https://doi.org/10.1200/jco.2005.03.0221 chose to study the effects of a novel treatment on newly diagnosed myleoma. The study investigators random;y assigned 103 patientrs to receive a control therapy and 104 patients to receieve the novel treatment. The primary outcome of interest was the rate of response among patients in control and treament where a response was defined as a 50% or greater decrease in detection of the cancer.

We could decide to define a vector v that will contain three pieces of information: a patient id that is an integer used to link a single patient to their clincial records, whether the patient was assigned to recieve a treatment or control therapy, and the percent response where 100 means complete response/reduction and 0 means no response/reduction.

[70]:

```
pt = c(1256, "TREATMENT", 87)
```

We stored a vector called pt with our three pieces of info. Now suppose we want to add an additional piece of info to our patient vector that determines if the patient had a succussful or unsuccussful response.

We need the if/else sytax. The if/else is an expression placed inside of parentheses and two blocks of code. if the expression evaluates to TRUE the first block of code will be executed and the second block is skipped. If the expression evaluates to FALSE then the second block will be executed and first block will be skippedd.

[71]:

```
if( pt[3] > 50 ){  # Our expression is pt[3] > 50
    pt = c(pt,1)
}else{
    pt = c(pt,0)
}
```

We expect that the code above will create a new vector c(1256,"TREATMENT",87,1). To be sure, lets check the variable pt by printing it.

[72]:

```
print(pt)
```

```
[1] "1256"      "TREATMENT" "87"        "1"
```

if/elseif/else and switch

The **if/else** handles one condition that either evaluates to TRUE or FALSE. However, we may need to execute a specific piece of code that depends on more than one condition. The **if/else if/else** syntax can execute a specific block of code dependent on more than one condition.

For example, suppose that we want to classify a patient's response as either above 75, between 50 and 75, not including 50 and 75, or less than or equal to 50. Let us also assume our patient is a following vector:

```
[73]: pt = c(6587,"CONTROL",24)
```

We can use the if /else if/ else syntax.

```
[74]: if( pt[3] > 75 ){
          pt = c(pt,2)
      }else if ( pt[3] >50 ) {
          pt = c(pt,1)
      }else {pt = c(pt,0) } # Our if/else if/else will run this␣
       ↪code bc the above two conditions are not met.
```

The if/else if/ else syntax begins at the top condition (`pt[3] > 75`). If the top condition is FALSE then the condition second to the top is evaluated next (`pt[3] > 50`). If the second condition is FALSE, and there are no additional `else if` conditions, then code in the `else` block is executed.

```
[75]: print(pt)
```

```
[1] "6587"    "CONTROL" "24"      "0"
```

Compund expressions

The expressions we used above evaluated a single expression as either TRUE or FALSE. At times we may need more complicated expressions that must evaluate several expressions at once. We will call these compund expressions.

For example, we may want to treatment response values different if the patient was assigned to treatment versus control. Suppose if a patient is assigned control and has a response above 50 they receieve a 1, if a patient is assigned control with a response equal to or less than 50 they recieve a 0. If a patient is assigned treatment and has a response above 50 they receieve a 2 and if their response is equal to or less than 50 they recieve a 3.

We need to evaluate two expressions based on the patient assignment and response. We are allowed to include logical comparisons (LAB01) inside expressions.

[76]:

```
pt = c(1359, "TREATMENT", 72)

if( (pt[3] > 50) && (pt[2]=="CONTROL") ){
    pt = c(pt,1)
}else if ( (pt[3] <= 50) && (pt[2]=="CONTROL") ){
    pt = c(pt,0)
}else if ( (pt[3] > 50) && (pt[2]=="TREATMENT") ){
    pt = c(pt,2)
}else if ( (pt[3] <= 50) && (pt[2]=="TREATMENT") ){
    pt = c(pt,3)
}

# check our work
print(pt)
```

```
[1] "1359"      "TREATMENT" "72"        "2"
```

The above code used the logical comparison `AND (&&)` to test whether the patient was treatment/control and if there repsonse was above or below 50.

10.0.1.2 Repitition

Often we will need to repeat a certain number of lines of code. We may need to apply the same operation to many different patients or observations or apply some set of code repeatedly until a condition it met. Loops allows us to **repeat** lines of code before continuing to execute lines of code below our loop.

While loop

The while loop repeats lines of code—called a **block**—until a specific condition is met.

[77]:

```
x = 2
while(x<1000){
    x = x^(2)
}
print(x)
```

```
[1] 65536
```

In the code above, we used the syntax `while (CONDITION) {CODE}` to execute a while loop. The code above assigned the variable `x` the value 2. The next line of code was the while loop. Because the condition evaluated to FALSE (x was not less than 1000) the code inside the parentheses—the code block—was executed once. After executing the code block the condition `x<1000` was tested again. Becasue the condition evaluated to FALSE, the code block was executed again, and again, and again, until the condition evaluated to TRUE. After the code block evaluated to TRUE, we continued to execute lines below the while loop in sequence.

The above code is equivalent to the following expanded code

[78]:

```
x=2
if (x <1000){
    x=x^2
}
if (x <1000){
    x=x^2
}
if (x <1000){
    x=x^2
}
if (x <1000){
    x=x^2
}
if (x <1000){
    x=x^2
}
if (x <1000){
    x=x^2
}
if (x <1000){
    x=x^2
}
if (x <1000){
    x=x^2
}
print(x)
```

```
[1] 65536
```

and so the while loop is a natural way to repeat a series of R code until meeting some condition.

For loop

Suppose we want to execute a code a block a *fixed* number of times. We could use a while loop to square the variable x four times using the below code:

[79]:

```
number_of_times=1
x = 2
```

```
while (number_of_times <5){
    x=x^2
    number_of_times = number_of_times+1
}
print(x)
```

```
[1] 65536
```

Because repeating code a fixed number of times is needed so often to solve a problem, a second type of loop was created—the **for loop**. The **for loop** repeats a code block a fixed number of times by iterating through a sequence.

Sequences

A sequence is a mapping, or function, from the numbers 1,2,3,4,… to some set of items $a, b, c, d, \cdots$. We often denote the items a, b, c, d, ... with a single variable name and a subscript that is called an index: $a_1, a_2, a_3, a_4, \cdots$. In the above sequence the number 1 maps to a_1, the number 2 maps to a_2 and so on.

We can create sequences of integers in R with the function `seq`. We can provide `seq` two integers, a "from" and "to" and this function will produce all integers between "from" and "to", including both the "from" value and the "to" value. We will produce all integers from -2 to 8. Watch

[80]:

```
seq(-2,8)
```

1. -2 2. -1 3. 0 4. 1 5. 2 6. 3 7. 4 8. 5 9. 6 10. 7 11. 8

This sort of request is so common in R that there is an easier way to write "from,to" sequences. We place the "from" value to the left and the "to" value to the right of a colon.

[81]:

```
-2:8
```

1. -2 2. -1 3. 0 4. 1 5. 2 6. 3 7. 4 8. 5 9. 6 10. 7 11. 8

R understand that the colon asks to produce all integers starting with the value on the left and ending with the value on the right.

Back to the for loop

Lets look how we can use sequences and the idea of the for loop to rewrite our code.

[82]:

```
x=2
for(number_of_times in 1:4){
    x=x^2
}
print(x)
```

```
[1] 65536
```

We can think of the for loop above in expanded form.

[83]:

```
x=2

x=x^2
x=x^2
x=x^2
x=x^2

print(x)
```

```
[1] 65536
```

10.0.1.3 Assignment

Lets use our new control flow skills to compute probabilities and conditional probabilities.

Run the below code to create one vector that describes whether patients were assigned to treatment or control and a second vector describing whether the patient had a succussful or unsuccesful status.

[84]:

```
draw_random_number = function(){
    return(runif(1))
}
```

Practice loop and conditional structure

After running the above code block, we can create a random number by running the following code `draw_random_number()`

1. Run the code `draw_random_number()`
2. Create and if/else statement that prints "Less than 1/2" when `draw_random_number()` is smaller than the value 0.5 and else prints "greater than or equal to 1/2" when `draw_random_number()` is 0.5 or larger.
3. Create a variable `category` and assign this variable the value -1.
4. Create an if/else if/else statement that assigns the variable category the value 1 when `draw_random_number()` is smaller than 0.25, the value 2 when `draw_random_number()` is greater than or equal to 0.25 and smaller

than 0.50, the value 3 when `draw_random_number()` is greater than or equal to 0.50 and smaller than 0.75 and the value 4 otherwise.

5. Create a vector random_values that is empty (i.e. `random_values = c()`).
6. Assign the variable `value` equal to `draw_random_number()`.
7. Create a while loop with the condition that `value` returns a value less than 0.5 - Inside the while loop, in the code block, assign the variable `value` to a new random number `draw_random_number()` - Inside the while loop, in the code block, append `value` to your vector `random_values`
8. Create a for loop that iterates the code block you create for the while loop 10 times.
9. What can you say about the vector `random_values` that would be produced from 7. versus 8.?

```
[85]: patient_status    =␣
       ↪c(0,1,1,0,0,1,1,1,0,0,0,1,1,1,0,0,0,0,1,0,0,0,1,1,1,0)
      patient_treatment =␣
       ↪c("t","c","t","c","t","c","t","t","c","t","c","t","c","t","c","t","c","t","t","c","t",␣
       ↪"c","c","t","c","c")
```

Lets create a for loop that will iterate through the positions of our two vectors `patient_status` and `patient_treatment`, performing different operations as we move to each item in the vector

1. Use the `length` function to compute the number of patients in the study and store that length in the variable `N`
2. Create a variable called `num_of_treat_assigns` and assign that variable the value 0
3. Create a variable called `num_of_contr_assigns` and assign that variable the value 0
4. Create a variable called `num_of_status_success` and assign that variable the value 0
5. Create a for loop that will run a code block starting at 1 and ending at `N`
6. Inside the code block of the for loop, use the variable `num_of_status_success` to count the number of patients with a succussful status
7. Use `N` and `num_of_status_success` to compute the probability of a success
8. Lets make our for loop more complicated. Inside our code block, create an if/else such that when patient treatment is a "t" we increment the variable `num_of_treat_assigns` by 1, else we increment the variable `num_of_contr_assigns` by 1.
9. Lets add to our if/else statement. Create two variables `num_of_trt_status_success` and `num_of_ctr_status_success` and

assign them both the value 0. Inside our loop, when a patient treatment is "t" we will implement step 8 and in addition we will increment the variable `num_of_trt_status_success` by 1. If the patient treatment is "c" then we will increment `num_of_ctr_status_success` by 1.

10. Using `num_of_trt_status_success` and `num_of_treat_assigns` compute the conditional probability of success given the patient was assigned treatment.
11. Using `num_of_ctr_status_success` and `num_of_contr_assigns` compute the conditional probability of success given the patient was assigned control.

11

Laboratory 03

thomas mcandrew, david braun

Lehigh University

CONTENTS

\nobreak

11.0.1 Functions

Functions in R are used to organize our work, clarify code for others, and allows us to apply the same finite steps to many different objects.

Example: Suppose we want to use the frequentist approach to assign probabilities to a sample space $\mathcal{G} = \{-1, 0, 1\}$ based on a dataset $\mathcal{D}$. Our dataset may look like $\mathcal{D} = (-1, 0, 0, 1, -1, 1, 1, 1, -1)$. To assign a probability to the event $\{-1\}$ we can write R code to count the number of times the outcome -1 appears in our data and divide by the number of data points in our data set. For the event $\{0\}$ we can write R code to perform the *same steps* for 0 as we performed for the event $\{-1\}$. Because we are repeating the same steps for the event $\{-1\}$ and for event $\{0\}$, a function may simplify our code.

Example: We are asked to support a clinical team that collected data on patients who are current smokers and outcomes thought to be linked to smoking. The code needs (i) to be processed, (ii) analyzed, and (iii) reported. Though we can write our code in sequence to perform all three steps, we may be able to better organize our work into three functions: one that processes the data, a function that analyses the data, and a third to report.

11.0.1.1 Anatomy of a function

A function in R has the following X parts:

- *Assignment:* We create a function by assigning a variable name to the function
- *Function* : Next we need to use the reserved word "function" so that R knows we are creating a function.
- *Arguments*: After the workd function we will include open and closed parentheses. Inside these parentheses we can include one or more arguments for the function. An argument is an input to our function.
- *Code block*: We write a sequence of steps to execute R code inside two curly brackets {}. **Important: Any variables that were created inside this code block are deleted after the function is finished.**
- *Return*: Inside the code block we can also include a return statement. This statement includes variables that were generated inside the code block that we wish to keep when the function is finished executing.

11.0.1.2 Declaring a function

When store a function in memory, we call this **declaring a function**.

Lets declare a function called `sum_two_numbers`. This function will have two arguments: one called `x` and one called `y`. Arguments are names that we use to identify specific inputs to our function—they are placeholders for variables outside of our function that we may want to use as inputs. In the code block we will write a line of code that stores the sum of `x` and `y` as the variable `z`. The second line in the code block will `return(z)`. Because we did not return `x` or `y` they will be deleted from memory after the function is finished executing.

[10]:

```
sum_two_numbers = function(x,y){
    z = x+y       # line of code in the code block. This uses␣
 ↪our x,y arguments (place holders)
    return(z)     # Return the variable z
}
```

11.0.1.3 Calling a function

When we declare a function it is stored in memory. If we want to apply our function to a set of arguments (inputs) then we **call** our function.

Lets call our function `sum_to_numbers` with the arguments 2 and 4.

```
[11]: result = sum_two_numbers(2,4)
```

We **called** our function by typing the name of the function and supplying the function with two arguments: 2 and 4. When we called the function, the following took place. 1. The variable `x` was assigned the value 2. 2. the variable `y` was assigned the value 4. 3. The first line of the code block was executed. 4. The second line of the code block was executed, returning the variable `z`. 5. The returned variable (`z`) was stored in the variable `result`.

Watch

```
[12]: print(result)
```

```
[1] 6
```

11.0.1.4 Named vs unnamed arguments

When we input our argument 2 and 4 into the function `sum_to_numbers` we did not specify which value should be assigned to the argument `x` and which should be assigned to `y`. When we do not specify the argument names we are proving a function **unnamed arguments**.

We could have called the function `sum_to_numbers` by specifying which arguments correspond to which values. When we supply a name and the value to the argument we are providing **named arguments**.

```
[13]: sum_to_numbers(y=2,x=4) # named arguments
```

6

11.0.1.5 Default arguments

When we declare a function we specify which arguments are needed for the function to execute all the lines in the coded block. We expect all these arguments to have values, but there may be a time when we do not necessarily need someone who uses our function to specify all the arguments. Instead, we can provide **default** argument values.

Lets create a function called `sumMult` that takes as input a vector that we will assign the name `v` and a logical value called `sum_or_mult`. If the value of `sum_or_mult` is TRUE then we will add all the items in the vector. If the value of `sum_or_mult` is FALSE then we will multiply all the values.

By default, we will add all the items. This means that is the user does not supply a value for the argument `sum_or_mult` then we automatically assign to `sum_or_mult` the value TRUE. To give the function `sumMult` a default value for the argument `sumMult` we will include after `sum_or_mult` an equals sign and our desired default value.

[2]:

```
sumMult = function(v,sum_or_mult=TRUE){ # This assigns the␣
 ↪sum_or_mult a default value of TRUE

    ## if sum_or_mult is TRUE, we sum. if sum_or_mult is␣
 ↪FALSE, we multiply
    if (sum_or_mult==TRUE){
        summation=0
        for (item in 1:length(v)){
            summation = summation + item
        }
        return(summation)
    }
    else{
        product = 1
        for (item in 1:length(v)){
            product = product * item
        }
        return(product)
    }
}
```

Lets create a vector called `fun_vector` and assign to it `c(4,2,-2,9,10,11,-0.3)` and lets call the function `sumMult`.

[20]:

```
fun_vector = c(4,2,-2,9,10,11,-0.3)

result1 = sumMult(fun_vector)                      # Default for␣
 ↪sum_or_mult is TRUE

result2 = sumMult(fun_vector,sum_or_mult=TRUE) # We are␣
 ↪always allowed to assign this argument a value

result3 = sumMult(fun_vector,sum_or_mult=FALSE) # and we can␣
 ↪assign this argument a different value then the default
```

[21]:

```
print(result1)
print(result2)
print(result3)
```

```
[1] 28
[1] 28
[1] 5040
```

11.0.1.6 Binding and Scope

To show that the variables summation and product are deleted after we call the function, lets try to print the variable `summation`.

[23]:
```
print(summation)
```

```
Error in print(summation): object 'summation' not found
Traceback:

1. print(summation)
```

R replies that it looked but cannot find the object summation. The function `sumMult` created a variable called `summation` operated with it inside the code block of our function and then deleted this variable.

When a variable name is assigned an object in R, the name and the object are associated with one another in the computer. The process of associating an object to a name in the computer is called **binding**.

When we call a function, the values we provide are bound to each argument name, the function is executed, and those variables are deleted. Which variables we can access during the executing of an R program is called **lexical scope**. Varables created inside a function can only be accessed and used by lines of code inside the code block. These variables are "in scope" of the function.

11.0.1.7 Assignment

Lets explore how functions can help us organize our work, clarify code for others, and repeat similar operations on objects of the same type.

Unnamed and named arguments

1. Declare a function `subtract` that takes two arguments: `x` and `y`. Inside the code block assign the variable `z` to be `x` minus `y` and return the variable `z`.
2. Call the function `subtract` on the values 2 and 6.
3. Call the function `subtract` on the values 2 and 6 but bind 2 to the argument `y` and 6 to the argument `x`.
4. Why did the results for 2. and for 3. change?

Successive subtraction

1. Declare a function `vec_subtract` that takes a vector argument that we will bind to `v`. Inside the code block compute the first item minus the second item minus the third item and so on. Store this subtarction in the variable `s` and return the variable `s`.

2. Call the function on the vector [1,2,3,4,5]

Frequentist approach to probability assignment

1. Run the code below called `random_vector`. This is a function that creates a vector of length 1000 filled with numbers between 0 and 1. The second line calls the function with no arguments and creates a variable called `rand_vec`.
2. Declare a function `freq_assign` that takes two arguments: `v` and `outcome`.
3. Inside the code block perform the following operations
4. Use the `length` function (pre-built by R) to compute the length of `v` and store this value in the variable `N`.
5. Create a variable called `outcome_of_interest` and assign to it an empty vector `c()`.
6. Create a for loop that iterates the variable `i` from the value 1 to the value `N`.
7. Inside the for loop use an if/else to identify if each item in `v` is less than or equal to `outcome`. If the item is less than or equal to `outcome` then append the value 1 to `outcome_of_interest` else append the value 0. We should expect a vector `outcome_of_interest` that is the same length as `v` which contains 1s for every value less than outcome and 0s otherwise.
8. Use the sum function in R (pre-built) to compute the sum of `outcome_of_interest` and assign this value the name `count`
9. return `count/N`
10. Call the function on `rand_vec` and record the result. Use 0.5 as the input value for `outcome`.

[8]:
```
random_vector = function(N=1000){
    return(runif(N))
}
rand_vec = random_vector()
```

12

Laboratory 04

thomas mcandrew, david braun
Lehigh University

CONTENTS

\nobreak

12.1 Matrix Algebra

Statisticians and data scientists use the language of vectors and matrices to organize data and work with models. We will take a close look at matrix algebra, taking time to relate matrix algebra to the more familar algebra that you have worked with in the past.

12.1.1 Recap on vectors and operations on vectors

\nobreak

12.1.1.1 Definition

A vector v is an ordered list of real numbers. We denote a vector with a lower case letter and enclose the values in the vector with square brackets. We can write a vector v as a *row vector*

$$v = [1, 2, 3, 4, 5] \tag{12.1}$$

or as a *column vector*

$$v = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \tag{12.2}$$

.

A vector has a **length**, defined as the number of values included in that vector. For example, the above vector is length 5.

We can create a vector in R by using the c operator.

```
[1]: v =c(1,2,3,4,5)
```

12.1.1.2 Vector times a scalar

A vector of length one is called a **scalar**. We can define the results of multiplying a vector $v = [1, 2, 3, 4]$ by a scalar α as each entry of the vector times the scalar α.

$$\alpha v = \begin{bmatrix} \alpha \times 1 \\ \alpha \times 2 \\ \alpha \times 3 \\ \alpha \times 4 \end{bmatrix} \tag{12.3}$$

For example, the vector $v = [1, 2, 3, 4, 5]$ times the scalar $\alpha = 7$ will result in a vector

$$\begin{aligned} \alpha v &= [7 \times 1, 7 \times 2, 7 \times 3, 7 \times 4] && (12.4) \\ &= [7, 14, 21, 28] && (12.5) \end{aligned}$$

R understands how to multiply vectors and scalars with no additional syntax.

```
[3]: v = c(1,2,3,4)
     alpha = 7

     alpha*v
```

1. 7 2. 14 3. 21 4. 28

12.1.1.3 Vector plus/minus a vector

A vector v plus a vector q creates a new vector that adds the individual entries of v and q

$$
\begin{aligned}
v &= [4, -1, 7] && (12.6)\\
q &= [0, 32, 9] && (12.7)\\
v + q &= [4 + 0, -1 + 32, 7 + 9] = [4, 31, 16] && (12.8)
\end{aligned}
$$

A vector v minus a vector q creates a new vector that subtracts the individual entries of q from v

$$
\begin{aligned}
v &= [4, -1, 7] && (12.9)\\
q &= [0, 32, 9] && (12.10)\\
v - q &= [4 - 0, -1 - 32, 7 - 9] = [4, -33, -2] && (12.11)\\
q - v &= [0 - 4, 32 - (-1), 9 - 7] = [-4, 33, 2] && (12.12)
\end{aligned}
$$

R also understands vector addition and subtraction with no special syntax.

```
[6]: v = c(4, -1, 7)
     q = c(0, 32, 9)

     add = v+q
     subtract = v-q

     print(add)
     print(subtract)
```

```
[1]  4 31 16
[1]   4 -33  -2
```

12.1.2 The Matrix

A **matrix** is an ordered list of vectors.

$$M = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 3 & 2 & 13 \\ 90 & 23 & 0 \end{bmatrix} \tag{12.13}$$

A matrix has a dimension which is an ordered pair of numbers: the first number indicates the number of rows of the matrix and the second number indicates the number of columns. For example, he matrix M above has dimension $(4,3)$.

To build a matrix in R we can issue the matrix command. Matrix is a function that takes a vector as an argument and can take one of many optional arguments.

If we give the matrix function only a vector, the default behavior of this function is to create a vector.

[8]:
```
A = matrix(c(1,2,3,4,5,6))
A
```

```
                                      1
                                      2
                                      3
                                      4
                                      5
A matrix: 6 × 1 of type dbl           6
```

This is because a vector can be thought of as a matrix with one column. In other words, a matrix with dimension $(N,1)$ is a vector of length N.

We can provide another arguement to the `matrix` function—ncol—to specify the number of columns we want to create.

[9]:
```
A = matrix(c(1,2,3,4,5,6), ncol=2)
A
```

```
                                    1  4
                                    2  5
A matrix: 3 × 2 of type dbl         3  6
```

Above we used the ncol argument to specify a matrix with two columns. R will automatically determine the number of rows for the matrix and fill-in the values of the matrix column by column.

If we want, we can ask R to fill in values of the matrix row by row by including an additional argument in the matrix function`byrow`

[10]:
```
v = c(1,2,3,4,5,6)
A = matrix(v, ncol=2)

B = matrix(v, ncol=2, byrow=TRUE)

print(A)
print(B)
```

```
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
     [,1] [,2]
[1,]    1    2
[2,]    3    4
[3,]    5    6
```

We see that matrix A was “column filled” and matrix B was “row filled”.

There are similar operations for matrices as there are for vectors.

12.1.2.1 Matrix plus/minus a matrix

Given a matrix A and matrix B, the sum of A and B is a new matrix C where the i,j entry of C (C_{ij}) is the sum of the corresponding entries from A and B ($C_{ij} = A_{ij} + B_{ij}$). To add two matrices they must have the same dimension.

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \tag{12.14}$$

$$B = \begin{bmatrix} 6 & 5 & 4 \\ 3 & 2 & 1 \end{bmatrix} \tag{12.15}$$

$$C = A + B = \begin{bmatrix} 1+6 & 2+5 & 3+4 \\ 4+3 & 5+2 & 6+1 \end{bmatrix} = \begin{bmatrix} 7 & 7 & 7 \\ 7 & 7 & 7 \end{bmatrix} \tag{12.16}$$

$$D = A - B = \begin{bmatrix} 1-6 & 2-5 & 3-4 \\ 4-3 & 5-2 & 6-1 \end{bmatrix} = \begin{bmatrix} -5 & -3 & -1 \\ 1 & 3 & 5 \end{bmatrix} \tag{12.17}$$

$$\tag{12.18}$$

R understands matrix addition and subtraction without any additional syntax.

[17]:
```
A = matrix(c(-1,2,4,1,3,8), ncol=3)
D = matrix(c(90,0.34,1.54,-9.43,10,0), ncol=3)

print(A)
print(D)

A+D
```

```
     [,1] [,2] [,3]
[1,]   -1    4    3
[2,]    2    1    8
      [,1]  [,2] [,3]
[1,] 90.00  1.54   10
[2,]  0.34 -9.43    0
```

A matrix: 2 × 3 of type dbl

89.00	5.54	13
2.34	-8.43	8

[14]:
```
A-D
```

A matrix: 2 × 3 of type dbl

-91.00	2.46	-7
1.66	10.43	8

12.1.2.2 Matrix times a Matrix

Two matrices A and B can be multiplied together $C = AB$ if the number of columns of A is equal to the number of rows of B. In other words, if the dimension of A is (a,b) and the dimension of B is (c,d) then b and c must be equal.

The i,j entry of this product, C_{ij} is the following sum of products:

$$\begin{aligned} C_{i,j} &= a_{i,1}b_{1,j} + a_{i,2}b_{2,j} + a_{i,3}b_{3,j} + \cdots + a_{i,N}b_{N,j} && (12.19) \\ &= \sum_{e=1}^{N} a_{i,e}b_{e,j} && (12.20) \end{aligned}$$

For example, suppose that

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \tag{12.21}$$

and that

$$B = \begin{bmatrix} 1 & 2 & -1 \\ 3 & 4 & -2 \end{bmatrix}. \tag{12.22}$$

Then the product $C = AB$ equals

$$C = AB = \begin{bmatrix} 1*1+2*3 & 1*2+2*4 & 1*-1+2*-2 \\ 3*1+4*3 & 3*2+4*4 & 3*-1+4*-2 \end{bmatrix} = \begin{bmatrix} 7 & 10 & -5 \\ 15 & 22 & -11 \end{bmatrix} \tag{12.23}$$

The above definition and computation can feel cumbersome. We can simplify the above calculations by introducing the inner product.

The inner product between two **vectors** v and q, or $v'q$, is

$$v = [1,2,3] \tag{12.24}$$
$$q = [4,5,6] \tag{12.25}$$
$$\tag{12.26}$$
$$v'q = 1 \cdot 4 + 2 \cdot 5 + 3 \cdot 6 \tag{12.27}$$
$$= 4 + 10 + 18 = 32 \tag{12.28}$$

Let us use the inner product to simplify the above matrix multiplication. First, we rewrite the matrix A as a stack of two *row* vectors

$$A = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \tag{12.29}$$

where $a_1 = [1,2]$ and $a_2 = [3,4]$

Second, we rewrite the matrix B as a stack of three *column* vectors

$$B = \begin{bmatrix} b_1 & b_2 & b_3 \end{bmatrix} \tag{12.30}$$

where $b_{1} =

$$\begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

$, $b_{2} =

$$\begin{bmatrix} 2 \\ 4 \end{bmatrix}$$

$, and $b_{3} =

$$\begin{bmatrix} -1 \\ -2 \end{bmatrix}$$

$

Then the product AB is a matrix of inner products

$$C = \begin{bmatrix} a_1'b_1 & a_1'b_2 & a_1'b_3 \\ a_2'b_1 & a_2'b_2 & a_2'b_3 \end{bmatrix} \tag{12.31}$$

Matrix multiplication is different than mulitplication between two variables that represent real numbers because matrix multiplication is **not** commutative—the product AB is not rarely equal to BA.

To compute the product of two matrices in R we need the %*% operator

[22]:
```
A = matrix(c(-1,0,1,-2,-1,3,4,5,6),ncol=3)
B = matrix(c(1,2,3,4,5,6,7,8,9),ncol=3)

print(A)
print(B)

print(A%*%B)
print(B%*%A)
```

```
     [,1] [,2] [,3]
[1,]   -1   -2    4
[2,]    0   -1    5
[3,]    1    3    6
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
     [,1] [,2] [,3]
[1,]    7   10   13
[2,]   13   25   37
[3,]   25   55   85
     [,1] [,2] [,3]
[1,]    6   15   66
[2,]    6   15   81
[3,]    6   15   96
```

Note above that we computed the product AB and BA and these products resulted in different matrices.

12.1.2.3 Matrix times a vector

A matrix A with dimension (r, c) can be multiplied by a vector v if the length of v is c. Then the product Av is a vector with the the i$^{\text{th}}$ entry of Av defined as

$$(Av)_i = \sum_{k=1}^{c} A_{i,k} v_k \tag{12.32}$$

For example, Let the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \tag{12.33}$$

and the vector

$$v = \begin{bmatrix} 4 \\ -3 \end{bmatrix} \tag{12.34}$$

then

$$Av = \begin{bmatrix} 1 * 4 + 2 * -3 \\ 3 * 4 + 4 * -3 \\ 5 * 4 + 6 * -3 \end{bmatrix} = \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix} \tag{12.35}$$

The definition above can also be simplifed by using inner products. Let

$$A = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \tag{12.36}$$

where $a_1 = [1, 2]$, $a_2 = [3, 4]$, and $a_3 = [5, 6]$. Then the product Av simplifies to

$$Av = \begin{bmatrix} a_1' v \\ a_2' v \\ a_3' v \end{bmatrix} \tag{12.37}$$

We can use the %*% operator in R to multiply together a matrix and a vector

[32]:

```r
A = matrix(c(1,2,3,4,5,6), ncol=2, byrow=TRUE)
v = c(4,-3)

print("Matrix A")
print(A)
print("Vector v")
print(v)

print("Av")
product = A%*%v
print(product)
```

```
[1] "Matrix A"
     [,1] [,2]
[1,]    1    2
[2,]    3    4
[3,]    5    6
[1] "Vector v"
[1]  4 -3
[1] "Av"
     [,1]
[1,]   -2
[2,]    0
[3,]    2
```

12.1.2.4 Matrix Transpose

Given a matrix M, the **tranpose** of M, $M^{'}$ or M^{T}, is a matrix where the first row of $M^{'}$ corresponds to the first column of M, the second row of $M^{'}$ corresponds to the second columns of M and so on. If M has dimension (r, c) than $M^{'}$ has dimensions (c, r).

For example, if

$$M = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 3 & 2 & 13 \\ 90 & 23 & 0 \end{bmatrix} \tag{12.38}$$

then the tranpose of M,

$$M^{'} = \begin{bmatrix} 1 & 4 & 3 & 90 \\ 2 & 5 & 2 & 23 \\ 3 & 6 & 13 & 0 \end{bmatrix} \tag{12.39}$$

We can use the `t()` operator to transpose a matrix in R.

```
[36]: M = matrix(c( 1,2,3,4,5,6,3,2,13,90,23,0),ncol=3,byrow=TRUE)

      print("The matrix M")
      print(M)

      print("The transpose")
      Mt = t(M)

      print(Mt)
```

```
[1] "The matrix M"
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    3    2   13
[4,]   90   23    0
[1] "The transpose"
     [,1] [,2] [,3] [,4]
[1,]    1    4    3   90
[2,]    2    5    2   23
[3,]    3    6   13    0
```

12.1.2.5 Matrix inverse

For a super fun overview of matrix multiplication and the concept of identities and inverses, click here!

The identity matrix

The **identity matrix** of dimension r, usually labled I_r, is a matrix with r rows and r columns such that the diagonal elements of the matrix, the (1,1) entry, (2,2) entry, up to (r,r) entry are the number 1 and all other entries are 0. For example,

$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{12.40}$$

This matrix is called the identity matrix because any matrix A times I returns A. To be more precise, this matrix is the *multiplicative* identity matrix. But the adjective *multiplicative* is usually dropped.

The inverse matrix

The **inverse** of a matrix A—A^{-1}—is a matrix such that, if A has the same number of rows and columns (called square) then

$$AA^{-1} = A^{-1}A = I \tag{12.41}$$

The idea of a matrix inverse is the same as in algebra. If we have a variable a

then the inverse of a, called a^{-1} is the unique number such that

$$a \cdot a^{-1} = a^{-1} \cdot a = 1.$$

In matrix algebra, the identity matrix takes the place of the “1” in algebra.

We can compute the inverse of a matrix A in R using the `solve` function.

```
[41]: A = matrix(c(3,4,5,6),ncol=2)

      print("A")
      print(A)

      print("A inverse")
      Ainverse = solve(A)
      print(Ainverse)
```

```
[1] "A"
     [,1] [,2]
[1,]    3    5
[2,]    4    6
[1] "A inverse"
     [,1] [,2]
[1,]   -3  2.5
[2,]    2 -1.5
```

12.1.3 Assignment

1. Define the vector v of length 3 with the following numbers: -1,0,1
2. Define the matrix A with dimension (2,3) by filling, column wise, the values: 1,2,3,4,5,6
3. Multiply A by v (Av)
4. Multiply v by A (vA). Why does R return an error?
5. Write a function that takes two argument which are both matrices and returns the product of these matrices.

13

Laboratory 05

thomas mcandrew, david braun

Lehigh University

CONTENTS

\nobreak

13.0.1 Data Frame

We learned in lecture that a dataframe is a specific way to organize data points sampled from a sample space $\mathcal{G}$. A dataframe supposes a dataset $\mathcal{D} = (d_1, d_2, \cdots, d_N)$ should be organized so that each row is a single outcome d_i and each column is a position in the tuple $d_i = (x, y, z, ...)$

The dataframe is a common way computational scientists think about data.

13.0.1.1 How to read a CSV file into a data frame

The function `read.csv` takes as an argument a string that indicates a file on your local computer OR a URL online. Below the `read.csv` function is used to import into memory a **dataframe** related to bad drivers from the statsitcal news outlet called *FiveThirtyEight*.

The news article is here https://fivethirtyeight.com/features/which-state-has-the-worst-drivers/

[23]:

```
d = read.csv("https://raw.githubusercontent.com/
 ↪fivethirtyeight/data/master/bad-drivers/bad-drivers.csv")
d # Jupyter will by default print any variable
```

State	Number.of.drivers.involved.in.fatal.collisions.per.billion.miles
<chr>	<dbl>
Alabama	18.8
Alaska	18.1
Arizona	18.6
Arkansas	22.4
California	12.0
Colorado	13.6
Connecticut	10.8
Delaware	16.2
District of Columbia	5.9
Florida	17.9
Georgia	15.6
Hawaii	17.5
Idaho	15.3
Illinois	12.8
Indiana	14.5
Iowa	15.7
Kansas	17.8
Kentucky	21.4
Louisiana	20.5
Maine	15.1
Maryland	12.5
Massachusetts	8.2
Michigan	14.1
Minnesota	9.6
Mississippi	17.6
Missouri	16.1
Montana	21.4
Nebraska	14.9
Nevada	14.7
New Hampshire	11.6
New Jersey	11.2
New Mexico	18.4
New York	12.3
North Carolina	16.8
North Dakota	23.9
Ohio	14.1
Oklahoma	19.9
Oregon	12.8
Pennsylvania	18.2
Rhode Island	11.1
South Carolina	23.9
South Dakota	19.4
Tennessee	19.5
Texas	19.4
Utah	11.3
Vermont	13.6
Virginia	12.7
Washington	10.6
West Virginia	23.8
Wisconsin	13.8
Wyoming	17.4

A data.frame: 51 × 8

13.0.1.2 Selecting a column and the $ operator

With a dataframe you can select rows by asking R to return the number column where the first column in the dataframe is 1, the second column 2, and so on. You can also ask R to return a column of your dataframe by column name.

The 3rd column of the dataframe that we read records, by state, the percentage of fatal collision that involved a driver who was impaired by alcohol.

We can ask R for the 3rd column by using square brackets in the same way that we used square brackets to select rows and columns of matrices.

[24]:
```
d[,3]
```

```
1. 39 2. 41 3. 35 4. 18 5. 35 6. 37 7. 46 8. 38 9. 34 10. 21 11. 19 12. 54 13. 36 14. 36
15. 25 16. 17 17. 27 18. 19 19. 35 20. 38 21. 34 22. 23 23. 24 24. 23 25. 15 26. 43
27. 39 28. 13 29. 37 30. 35 31. 16 32. 19 33. 32 34. 39 35. 23 36. 28 37. 32 38. 33
39. 50 40. 34 41. 38 42. 31 43. 21 44. 40 45. 43 46. 30 47. 19 48. 42 49. 34 50. 36
51. 42
```

We can request a column from a dataframe by asking for 1. The dataframe 2. square brackets Inside the square brackets you divide the rows you want selected and the columns you want selected by a comma.

For example, if we wanted to look at the 10th row and 3rd column we can write

[25]:
```
d[10,3]
```

21

If instead we wanted to view rows 10,11,12 and columns 3 and 4 we could write

[26]:
```
d[ 10:12, c(3,4)  ]
```

A data.frame: 3 × 2

	Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Speeding <int>	Percenta <int>
10	21	29
11	19	25
12	54	41

The above examples select columns and rows by number. One advantage of a dataframe is that we can select columns **by name**.

For example, we may be interested in rows 10-12 of the column "Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Speeding"

[27]:

```
d[10:12,"Percentage.Of.Drivers.Involved.In.Fatal.Collisions.
 ↪Who.Were.Speeding"]
```

1. 21 2. 19 3. 54

We can ask for **all** rows that correspond to a column by leaving the entry to the left of the comma blank.

[28]:

```
d[,"Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.
 ↪Were.Speeding"]
```

1. 39 2. 41 3. 35 4. 18 5. 35 6. 37 7. 46 8. 38 9. 34 10. 21 11. 19 12. 54 13. 36 14. 36 15. 25 16. 17 17. 27 18. 19 19. 35 20. 38 21. 34 22. 23 23. 24 24. 23 25. 15 26. 43 27. 39 28. 13 29. 37 30. 35 31. 16 32. 19 33. 32 34. 39 35. 23 36. 28 37. 32 38. 33 39. 50 40. 34 41. 38 42. 31 43. 21 44. 40 45. 43 46. 30 47. 19 48. 42 49. 34 50. 36 51. 42

Finally, there is a shorthand for the code `d[,"Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Speeding"]` using the dollarsign operator. The dollar sign operator can be thought of as a function that takes a column name as input and returns the column of data inside your data frame corresponding to that column.

[41]:

```
d$Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.
 ↪Speeding
```

```
NULL
```

13.0.1.3 Logical indexing

We are allowed to use logical indexing like we learned when selecting items in vectors and matrices to select rows and columns of a data frame. For example, suppose we are interested in **all** columns of our data frame where the Percentage Of Drivers Involved In Fatal Collisions Who Were Speeding is above 40%.

[30]:

```
d[ d$ "Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.
 ↪Were.Speeding" > 40, ]
```

A data.frame: 8 × 8

	State	Number.of.drivers.involved.in.fatal.collisions.per.billion.miles	P
	<chr>	<dbl>	<
2	Alaska	18.1	41
7	Connecticut	10.8	46
12	Hawaii	17.5	54
26	Missouri	16.1	43
39	Pennsylvania	18.2	50
45	Utah	11.3	43
48	Washington	10.6	42
51	Wyoming	17.4	42

or maybe we are interested in not all columns, but just the states where this percentage is above 40%. We can subset to a specific column by including in square brackets the name of the column

[34]:
```
d[ d$ "Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.
↪Were.Speeding" > 40, "State" ]
```

1. 'Alaska' 2. 'Connecticut' 3. 'Hawaii' 4. 'Missouri' 5. 'Pennsylvania' 6. 'Utah' 7. 'Washington' 8. 'Wyoming'

If we wanted to include more than one column then we can include a vector that contains each column we want to select.

[36]:
```
d[ d$"Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.
↪Were.Speeding" > 40, c("State","Car.Insurance.Premiums....
↪") ]
```

A data.frame: 8 × 2

	State	Car.Insurance.Premiums....
	<chr>	<dbl>
2	Alaska	1053.48
7	Connecticut	1068.73
12	Hawaii	861.18
26	Missouri	790.32
39	Pennsylvania	905.99
45	Utah	809.38
48	Washington	890.03
51	Wyoming	791.14

13.0.1.4 Functions for data frames

There are several useful functions that take as an argument a data frame. The `nrow` function takes a data frame as an argument and returns the number of rows (i.e. the number of data points) in the data frame. The `ncol` function takes a data frame as an argument and returns the number of columns in the data frame.

[31]:

```
number_of_rows = nrow(d)
number_of_columns = ncol(d)

print(number_of_rows)
print(number_of_columns)
```

```
[1] 51
[1] 8
```

The `colnames` function takes as an argument a data frame and returns a vector that contains the name of each column in the data frame.

[32]:

```
column_names = colnames(d)
print(column_names)
```

```
[1] "State"
[2] "Number.of.drivers.involved.in.fatal.collisions.per.billion.
 ↪miles"
[3] "Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.
 ↪Were.Speeding"
[4]
"Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.
 ↪Alcohol.Impaired"
[5] "Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.
 ↪Were.Not.Distracted"
[6] "Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Had.
 ↪Not.Been.Involve
d.In.Any.Previous.Accidents"
[7] "Car.Insurance.Premiums..."
[8]
"Losses.incurred.by.insurance.companies.for.collisions.per.
 ↪insured.driver..."
```

The `summary` function is a function that takes as an argument a data frame and returns, for each column in the data frame, the minimum value, maximum value, mean (average), median, 25th percentile (called the 1st quartile), and the 75th percentile (called the 3rd quartile).

[33]:

```
summary(d)
```

```
    State
 Length:51
 Class :character
 Mode  :character
```

```
Number.of.drivers.involved.in.fatal.collisions.per.billion.
↪miles
Min.   : 5.90
1st Qu.:12.75
Median :15.60
Mean   :15.79
3rd Qu.:18.50
Max.   :23.90
Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.
↪Speeding
Min.   :13.00
1st Qu.:23.00
Median :34.00
Mean   :31.73
3rd Qu.:38.00
Max.   :54.00
Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.
↪Alcohol.Impaired
Min.   :16.00
1st Qu.:28.00
Median :30.00
Mean   :30.69
3rd Qu.:33.00
Max.   :44.00
Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.
↪Not.Distracted
Min.   : 10.00
1st Qu.: 83.00
Median : 88.00
Mean   : 85.92
3rd Qu.: 95.00
Max.   :100.00
Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Had.Not.
↪Been.Involved.In.Any.Previous.Accidents
Min.   : 76.00
1st Qu.: 83.50
Median : 88.00
Mean   : 88.73
3rd Qu.: 95.00
Max.   :100.00
Car.Insurance.Premiums...
Min.   : 642.0
1st Qu.: 768.4
Median : 859.0
```

```
Mean   : 887.0
3rd Qu.:1007.9
Max.   :1301.5
Losses.incurred.by.insurance.companies.for.collisions.per.
↪insured.driver...
Min.   : 82.75
1st Qu.:114.64
Median :136.05
Mean   :134.49
3rd Qu.:151.87
Max.   :194.78
```

13.0.2 Assignment

We are going to look at a dataset that was collected on University “Fight Songs” that are played during sporting events. The article about the data set is here= https://projects.fivethirtyeight.com/college-fight-song-lyrics/

1. The data is at the URL (https://raw.githubusercontent.com/fivethirtyeight/data/master/fight-songs/fight-songs.csv). Please read in this raw CSV as a data.frame
2. Use the nrow and ncol function to describe the number of fight songs sampled and the number of different characteristics collected about each song.
3. How many songs were sampled with a BPM (beats per minute) above 150?
4. Select the rows where BPM is greater than 150 and select the column “nonsense” (Whether or not the song uses nonsense syllables (e.g. “Whoo-Rah” or “Hooperay”))
5. Summarize the data frame. Why do you think the summary function does not produce useful information for many of the columns?

Bibliography