CHAPTER 1. OVERVIEW AND DESCRIPTIVE STATISTICS

The main outcome for this course is an introduction for how to formally think and reason about uncertainty and data. Without training, people tend to be very bad at such tasks.

To illustrate this point, try the following experiment:

Step (1). Imagine the process of tossing a fair coin 100 times, each time recording the outcome. For example,

H H T H T T H . . .

Step (2). This time, toss an actual fair coin 100 times, and record the outcome each time.

Step (3). Compare the 100 recorded outcomes of the imaginary coin tosses with the 100 actual tosses of a coin.
(i) What systematic differences do you observe in these two data sets?
(ii) What do the differences in these data suggest about the way we perceive uncertainty, versus reality?

After diving into that thought experiment about randomness, next think about how you would define the notion of "data". Write down your definition, and try to be as precise and exhaustive as possible. Keep your definition with you throughout the semester, and refer back to it often to think about how your thoughts about data evolve as your understanding develops.

SECTION 1.1. POPULATIONS, SAMPLES, AND PROCESSES

Whenever you think about data, the first notion to cross your mind should be the notion of a "population". In fact, without the notion of a population data would have no context nor meaning.

Population is a broadly defined concept, but has a very precise definition in the context of a particular data set and questions of interest. A population is an entity of interest. It could be the mental health status of citizens of a certain county, for example, with respect to dementia. Or the population could be the length of time between buses at a bus stop. A population may be fully or partially observable, or not directly observable.

Alternatively, data is any information used to learn about any feature of a population of interest. For example, information collected at mental health screenings, or randomly sampled recordings of the inter-arrival times at a bus stop. Data could come in numerical form or could be formatted as written text. For example, clinician notes about a patient at a hospital represent data containing information about numerous population features that may be of interest.

Once the theoretical population and features of interest have been defined a "sample" of data is collected to make inference on the population/feature. A "sample" is any subset of the population, and is only relevant when the full population is not available.

Note that the distinction between a sample and a population is extremely important, is often very subtle, and is entirely context dependent. To illustrate, consider the following example.

EXAMPLE. Suppose I need to learn the average GPA of college students in North Carolina. Then the population consists of all college students in North Carolina, and the feature of interest is the average GPA. In theory, I could learn this number if I had access to the GPA of every college student in North Carolina. However, this may not be practical, and instead I could estimate the average GPA by sampling some subset of North Carolina college students. For example, suppose I have access to the GPA of every NCSU undergrad. In that case, the collection of GPAs of all NCSU students is the sample.

Things to think about:

(i) Is this sample representative of the population? In what ways is it not representative? For instance, are there likely systematic differences in the GPAs of NCSU, UNC, and Duke students? What about about App State students? How about Wake Tech students? Does the population of "College students in North Carolina" include community college students or only students at 4-year institutions? Maybe the population needs to be defined more precisely for the question(s) of interest.

(ii) The questions in item (i) allude to possible sampling biases. Note that a bias exists only in the context of a question of interest, relating a population to a sample.

(iii) What if alternatively, the population is defined as "College students at NCSU"? In this case, my sample contains the entire population, and I can compute exactly the feature I care about with no uncertainty.

DEFINITION. A simple random sample is a sample from a population such that any particular subset, of pre-specified size, of the sample has the same chance of being selected.

Read Section 1.1 in Devore.

SECTION 1.2. PICTORIAL AND TABULAR METHODS IN DESCRIPTIVE STATISTICS

This section begins with the introduction of stem-and-leaf displays and dot plots. It is worth reading about them in the textbook, but the main graphical representation to which these plots relate is a histogram.

DEFINITION. A variable is a quantity with values that in some way characterizes the members or objects in a population.

For example, a variable

X = Number of conflict related deaths in a given nation/time
Then $X \in \{0, 1, 2, \dots\}$

Y = Distance to commute to/from school for NCSU students.
Then $Y \in [0, \infty)$

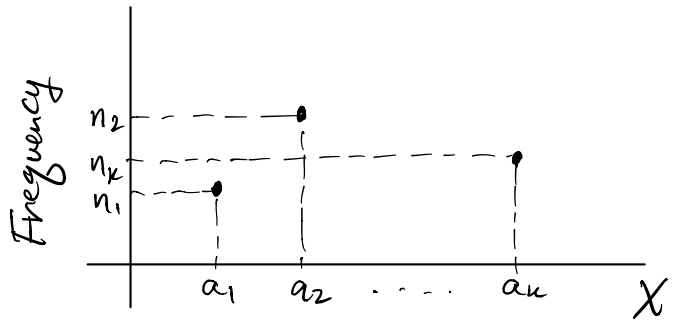
DEFINITION. A variable is discrete if it takes values only in a countable set. Alternatively, a variable is continuous if it takes values consisting of one or more entire intervals. A variable that is neither discrete nor continuous is called mixed.

Continuing with X and Y as in the example, first suppose that we have collected a sample of observations $x_1, x_2, \dots, x_n$. Assuming that these data have all come from the same population it is helpful to visualize them collectively to develop an intuition for what values of $X \in \{0, 1, 2, \dots\}$ are most common. There are many tools to visualize data to this end, but a histogram is typically among the most effective tools.

The algorithm for constructing a histogram for discrete data $x_1, \dots, x_n$ is as follows.

(1) Identify the set of unique values observed in the set $\{x_1, \dots, x_n\}$. Denote the k unique values as $a_1, \dots, a_k$.
(2) Count the number of times that each value $a_1, \dots, a_k$ occurs in the set $\{x_1, \dots, x_n\}$. Denote these frequencies by $n_1, \dots, n_k$.

(3) Plot the values $n_1, \dots, n_k$ against $a_1, \dots, a_k$



Alternatively, for step (3) you may want to plot the values $n_1/n, \dots, n_k/n$ versus $a_1, \dots, a_k$. The values n_i/n are called relative frequencies since

$$\sum_{i=1}^{k} \frac{n_i}{n} = \frac{n_1}{n} + \cdots + \frac{n_k}{n} = 1.$$

Note that it must be the case that

$$\sum_{i=1}^{k} n_i = n.$$

Note that for discrete data the histogram plot can be presented as points, lines, or bars. It is better to NOT use bars, however, to reflect that the data cannot take values continuously in an interval.

The algorithm for constructing a histogram for continuous data $y_1, \dots, y_n$ is as follows.

(1) Divide the interval $\left[\min_{1 \le j \le n}\{y_j\}, \max_{1 \le j \le n}\{y_j\}\right]$ into a partition of m bins of equal length. Denote the edges of the bins as

$$b_0 \le \min\{y_j\}, b_1, \dots, b_m \ge \max\{y_j\}$$

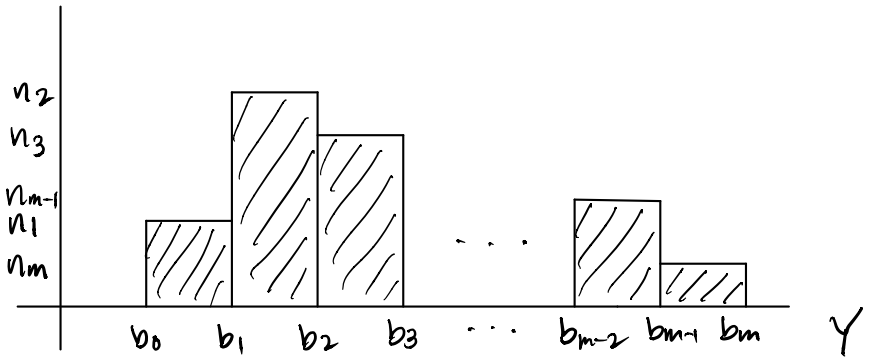
where $b_1 - b_0 = b_2 - b_1 = \cdots = b_m - b_{m-1}$.

(2) Construct the frequencies $n_1, \dots, n_m$ as

$$n_i = \left|\{j : y_j \in [b_{i-1}, b_i)\}\right|$$

for $i \in \{1, \dots, m\}$.

(3) Plot m rectangles with respective heights $n_1, \dots, n_m$ (or $n_1/n, \dots, n_m/n$) and width edges $b_0, b_1, \dots, b_m$.



Note that for continuous data the bars are touching because the variable Y could have taken any value in the intervals/bins $[b_{i-1}, b_i)$.

We also could have constructed the bins on $(b_{i-1}, b_i]$.

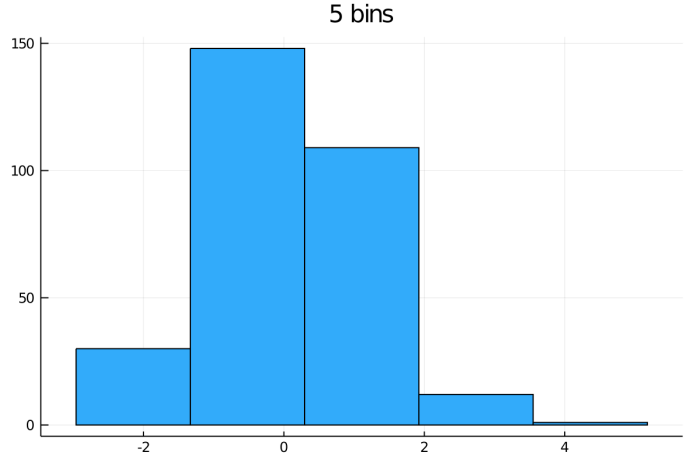
The bins all have to the same width so that the n_i values are comparable. However, what is it about the construction of the histogram for continuous data that has NOT been fully specified?

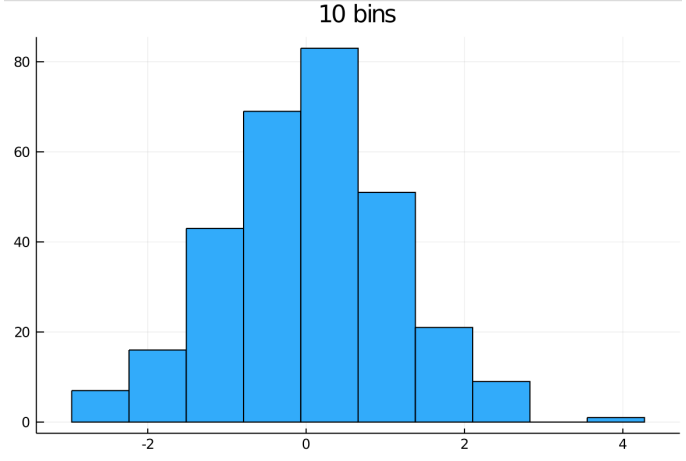
→ How to determine the correct number of bins, m?

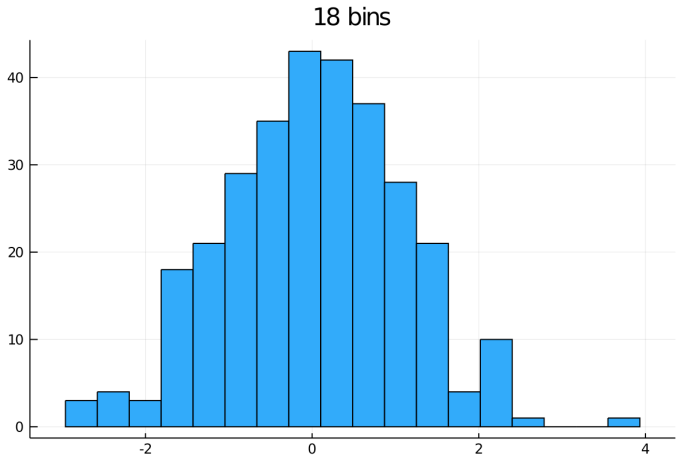
(i) There is no correct number of bins to choose
(ii) Academic papers have been written to investigate an optimal choice of number of bins.
(iii) A good choice is related to the sample size, n.
(iv) The representation of a histogram is highly sensitive to the number of bins.
(v) A reasonable choice is $\sqrt{n}$.

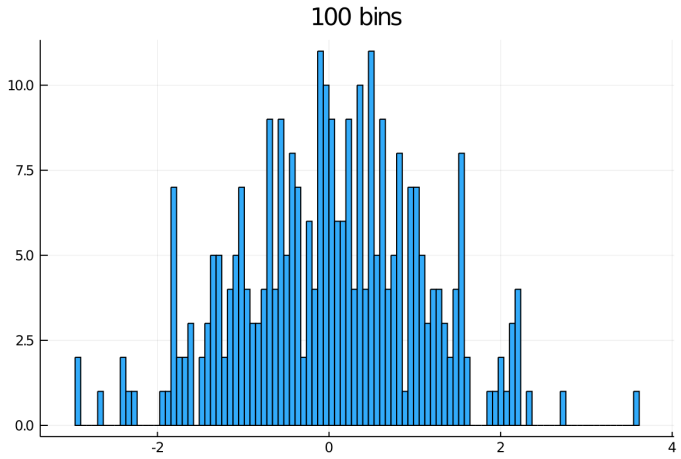
To illustrate these points, consider the following data and corresponding histograms. I have used a random number generator in the computer programming language Julia to generate $n = 300$ data points from a continuous random variable (rounded to two decimal places).

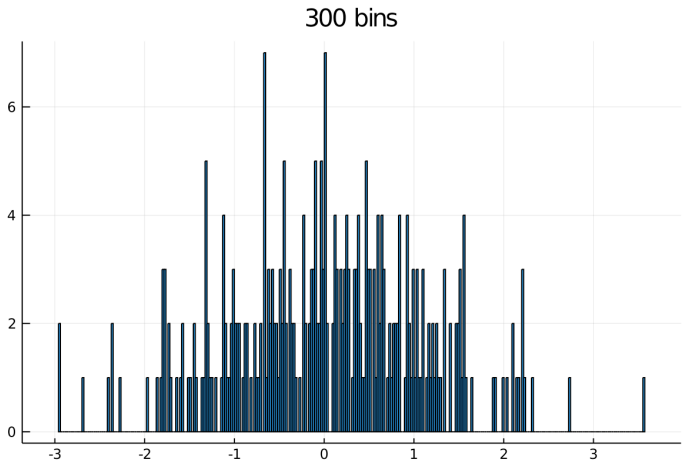
[-0.03, 3.55, 0.48, -0.88, -1.74, -1.02, 1.52, 0.24, 0.09, 0.15, 0.77, -0.22, -0.34, -1.58, -0.46, 0.03, 0.02, -0.22, 1.56, 0.51, 0.28, -1.29, -0.57, 0.15, -0.21, -2.94, -1.78, 2.22, -0.12, 1.52, -1.43, 1.21, 0.75, -0.27, -0.09, -1.08, -1.65, -0.48, 0.02, 0.47, 0.44, -0.61, -1.0, -0.45, 1.91, 0.96, 1.64, -2.96, 0.23, 1.34, 0.65, -1.14, 0.92, 0.01, -0.06, -2.28, -1.1, 0.5, -0.2, -1.82, 0.52, 0.28, 0.18, -1.31, 1.53, -1.78, 2.23, -1.73, 0.18, -0.67, 1.23, -0.03, 0.6, -0.85, -0.67, -0.94, 0.31, -2.37, -0.04, -0.47, -0.04, -1.79, -1.27, 0.51, 0.22, 0.42, 1.26, -1.87, 0.94, -0.12, -1.0, 0.2, 2.74, -1.37, -1.61, -1.03, 1.02, 0.8, -0.05, 0.04, 1.06, -1.51, -0.03, 1.15, -0.67, 0.33, 0.95, 2.1, -0.16, -0.71, 0.63, 0.58, -1.8, 0.67, 1.55, 0.74, 0.5, -0.58, 0.09, 1.11, 1.88, -1.3, -0.61, 0.36, -0.1, -0.62, 1.52, -0.31, -0.78, 0.56, 0.59, 1.49, 0.67, -0.42, 0.26, 2.14, -0.53, 2.22, 0.79, 0.13, 0.65, 0.4, -0.95, 1.46, -0.45, -0.87, -0.38, -1.34, -0.59, 0.92, -1.32, 0.36, 1.03, -0.07, 0.81, 0.83, 0.15, 1.22, -1.06, -2.4, -1.21, -0.42, -1.13, -0.85, 0.25, 0.5, -1.97, 0.33, -0.38, 0.12, -0.96, 1.34, -0.22, -0.09, 0.13, -1.78, 0.01, 1.11, -0.13, -0.73, -1.02, 0.38, 2.33, -1.31, -0.45, -0.54, 0.56, -0.22, 0.37, -0.71, 1.41, 0.7, 2.21, -1.11, 1.0, -0.09, 1.5, 0.65, 0.82, -0.44, -0.65, 0.02, 0.99, 0.66, 0.26, -0.01, 1.47, 0.34, 0.48, 1.29, -0.49, 0.6, 0.61, -0.08, 1.07, -1.32, -0.35, -1.02, 1.55, -1.5, -0.78, 0.55, -0.14, -0.12, 1.11, 0.12, 0.73, 1.19, -1.1, 0.39, -0.96, -0.75, -0.15, 0.83, -0.1, 0.9, -0.91, -1.13, -0.48, -0.51, 0.83, -2.36, -0.33, -1.24, -1.04, 0.41, -0.56, 0.19, -1.71, 1.0, 2.16, 0.93, -0.67, 0.01, -0.58, 0.93, -1.44, 0.28, 0.26, -0.66, -1.58, 0.65, 0.78, 0.84, 1.58, -0.01, -0.82, 1.16, -0.67, -0.02, 2.03, -0.35, 1.28, 1.99, -1.79, 0.38, 2.11, 1.41, 0.38, -0.37, -0.44, -1.12, 1.35, -0.17, 1.25, 1.56, 0.63, 0.01, -1.44, 0.2, 1.03, 0.48, 0.47, 1.03, -0.59, -2.68, -0.67, -1.32, -0.55, 1.16]











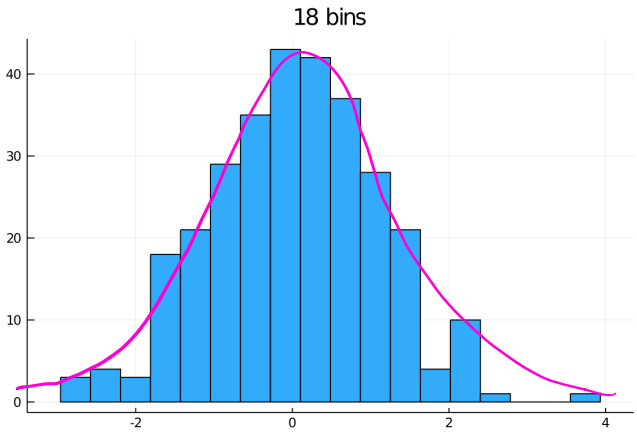
The figures show five different histograms of the same data, each using a different number of bins. Note that

$$\sqrt{n} = \sqrt{300} \approx 18,$$

and observe how different the picture of the data looks for different numbers of bins.

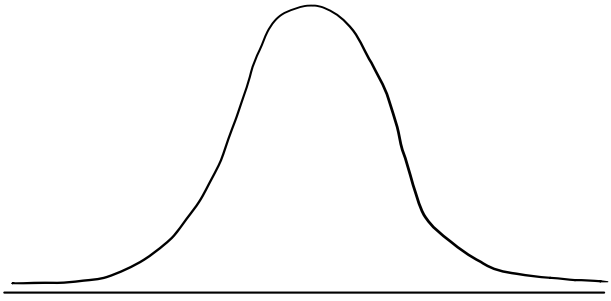
Aside from this topic of simple visual representations of univariate data, histograms are an approximation to a deeper and intrinsic feature of a random variable, the probability density (or mass) function. These functions will be an important topic that we will focus on in coming chapters.

The pink line is a representation of the (unnormalized) probability density function for these data, as estimated by the histogram. The larger the sample size, n, the better

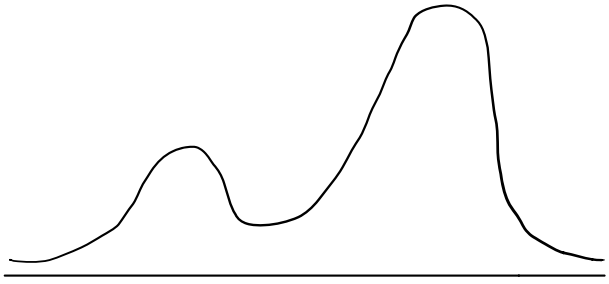


approximation to the true (or theorized) density function, from the histogram.

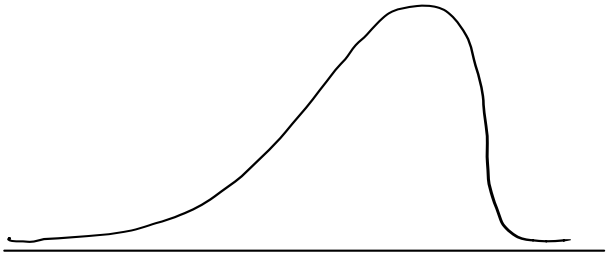
There are a variety of characteristics used to describe a given histogram. Most basically we describe the modality and the skew of the histogram.



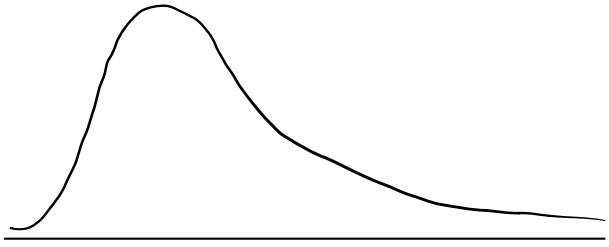
"Unimodal"
"symmetric"



"Bimodal" or "Multimodal"



"Negative Skew"



"Positive skew"

Read Section 1.2 in Devore.

SECTION 1.3. MEASURES OF LOCATION

In this section we develop more precise numerical summaries of data to describe the central tendency of a distribution of data. Namely, this includes the mean and the median.

Most people educated in the United States are familiar with the notion of a mean as an average of a collection of numbers. This concept is taught in primary school, and this is where we will begin our discussion.

DEFINITION. Suppose $x_1, \ldots, x_n$ is a collection of numbers. The arithmetic mean is defined as $\bar{x}_n := \frac{1}{n}\sum_{i=1}^{n} x_i$.

Note that $\bar{x}_n$ is called the arithmetic mean as opposed to the geometric mean or the harmonic mean. Further, we often call it the sample mean. Why?

→ This is where you should think a bit deeper. What is the point of computing or even defining $\bar{x}_n$?

→ We are trying to measure something about the broader population from which these data were generated/sampled.

→ This is why the notion of a population (and sample) are very important.

→ When you interpret an average you should ask whether the average is being used to reflect the unknown population mean, or is simply being used to describe a static collection of numbers.

We also have the notation μ used to refer directly to the population mean. Greek letters are commonly used in the field of statistics to denote population features. Much more will be said about population features in future chapters.

The next thing to consider is why the sample mean is thought of as a natural measure of location. In fact, this idea was understood as very strange a few hundred years ago.

→ First, for using a sample mean to describe the center of a set of data, $\bar{x}_n$ may not even represent an observed data value. For example, suppose $n=2$, $x_1 = 4$, $x_2 = 5$. Then

$$\bar{x}_2 = 4.5 \notin \{x_1, x_2\}$$

→ The sample mean is a feature of a set of data at the group level, not at the individual level.

→ Mathematical results for large sample theory shows that under reasonable conditions

$$\bar{X}_n \longrightarrow \mu \quad \text{as} \quad n \longrightarrow \infty$$

Before discussing a few other properties of the sample mean, I will introduce another measure of location. This will give helpful context for making better sense of the properties.

DEFINITION. For a sample of observed data $x_1, \ldots, x_n$ the sample median is any number M_n that satisfies

(a) $\sum_{i=1}^{n} 1\{x_i < M_n\} = \sum_{i=1}^{n} 1\{x_i > M_n\} = \frac{n}{2}$, if n is even.

(b) $M_n = x_j$ such that $\sum_{i=1}^{n} 1\{x_i < x_j\} = \sum_{i=1}^{n} 1\{x_i > x_j\} = \frac{n-1}{2}$, if n is odd.

where $1\{\cdot\}$ is the indicator function (1 if $\{\cdot\}$ true, 0 else).

As was the case with the mean, the sample median is the sample analogue of the population median. We will likely discuss the notion of the population median in coming chapters.

Do not confuse the definition of the sample median with the algorithm that you have surely learned for finding the sample median, in previous courses. The basic algorithm is as follows.

Step 1. Order the observed data $x_1, x_2, \ldots, x_n$ in increasing order

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)} \quad \text{— these are called order statistics.}$$

Step 2a. If n is odd, then $M_n = x_{\left(\frac{n+1}{2}\right)}$

Step 2b. If n is even, then any number $M_n \in \left(x_{\left(\frac{n}{2}\right)}, x_{\left(\frac{n}{2}+1\right)}\right)$ suffices.

Note that in the textbook it is recommended to choose the average value,

$$M_n = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2},$$

however, there is no particularly strong reason to do so. For example, suppose

$$\{x_1, \ldots, x_6\} = \{61, 45, 39, 18, 14, 33\}$$

Then $$\{x_{(1)}, \ldots, x_{(6)}\} = \{14, 18, 33, 39, 45, 61\}$$

So the median $M_n \in (33, 39)$. We could call $M_n = \frac{33+39}{2} = 36$, but is 36 any more representative of the "middle" than any other number in the interval $(33, 39)$, based on the definition of the median?

Suppose we observe another data point, x_7. Then

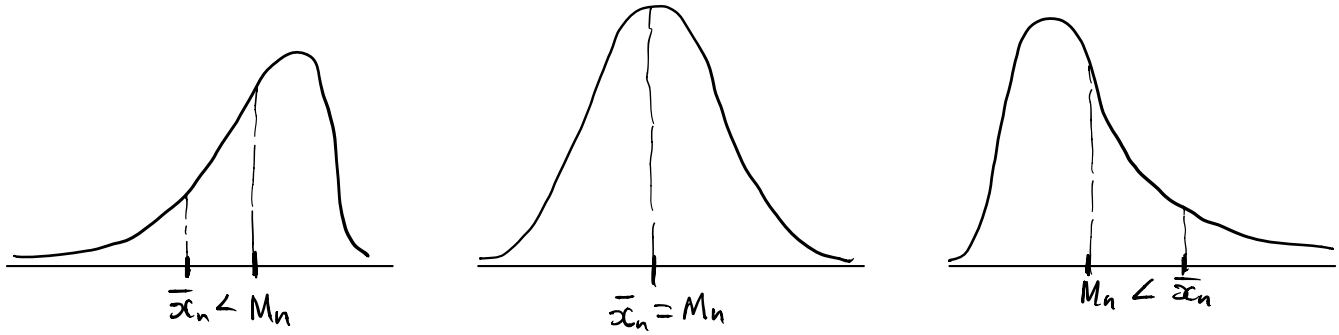
(a) If $x_7 < 33$, $M_n = 33$
(b) If $x_7 > 39$, $M_n = 39$
(c) If $x_7 \in [33, 39]$, $M_n = x_7$

None of these cases relate to 36, unless it happens that $x_7 = 36$.

Now consider a few distinguishing properties of the mean and median.

(1) The median is said to be "robust" to extreme values or outliers, while the mean is not.
EXAMPLE. Consider the set of numbers $\{1, 2, 3\}$. Then $\bar{x}_3 = 2 = M_3$. Observe that if the next observation is -100, then the data set $\{-100, 1, 2, 3\}$ has $M_4 \in (1, 2)$ which is not much different from $M_3 = 2$, but $\bar{x}_4 = -23.5 \ll 2 = \bar{x}_3$.

(2) The relative positions of the mean and median allude to the symmetry or skewness of the data. That is,



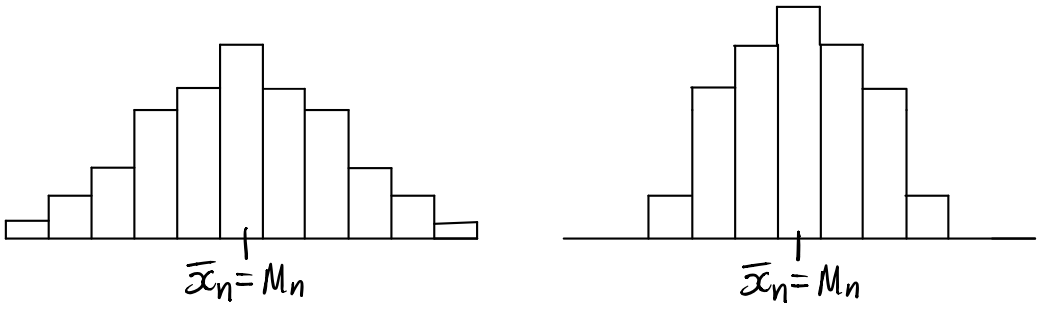
Negative skew | Symmetric | Positive skew

Other measures of location such as a trimmed mean are briefly discussed in Section 1.3.

Read Section 1.3 in Devore.

SECTION 1.4. MEASURES OF VARIABILITY

Measures of location are a useful first metric to consider when describing a sample of data, but the certainly do not give a full picture. For example,



These are two histograms representing data that has the same mean and median (i.e., location), but different variability.

Perhaps the simplest measure of variability is the range, $x_{(n)} - x_{(1)}$, and is useful for giving a sense of the magnitude and scale of the observed values.

More commonly, it is useful to consider deviations in the observed values x_i from some measure of location. This allows us to interpret the x_i relative to a meaningful and common point of reference. But how to measure deviations?

Start by considering $\bar{x}_n$ as the location point of reference. Then we can compute the deviations

$$x_1 - \bar{x}_n, \quad x_2 - \bar{x}_n, \quad \ldots, \quad x_n - \bar{x}_n$$

Reducing these deviations to a single value by summing them together gives

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x}_n) &= \left(\sum_{i=1}^{n} x_i\right) - n\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\ &= \left(\sum_{i=1}^{n} x_i\right) - \left(\sum_{i=1}^{n} x_i\right) \\ &= 0\end{aligned}$$

This is because positive deviations cancel with negative deviations. Instead, consider first squaring the deviations so that they are all positive. This leads to what is referred to as the sample variance.

DEFINITION. The sample variance, denoted by S_n^2, is given by

$$S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2.$$

The sample standard deviation is defined as $S_n = \sqrt{S_n^2}$.

Note that the sample variance is a natural analogue to the sample mean. The standard deviation is more interpretable because it is on the same scale as the data, whereas the variance is units squared.

The reason for using $\frac{1}{n-1}$ instead of $\frac{1}{n}$ in the expression for S_n^2 is so that it is an "unbiased estimator" of the population variance, σ^2. For a population of finitely many members, say N,

$$\sigma^2 := \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

A useful formula for computing S_n^2 is the following.

$$\begin{aligned}(n-1)\,S_n^2 &= \sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \\ &= \sum_{1}^{n}(x_i^2 - 2x_i\bar{x}_n + \bar{x}_n^2) \\ &= \left(\sum_{1}^{n} x_i^2\right) - 2\bar{x}_n\sum_{1}^{n} x_i + n\bar{x}_n^2 \\ &= \left(\sum_{1}^{n} x_i^2\right) - 2n\bar{x}_n^2 + n\bar{x}_n^2 \\ &= \left(\sum_{1}^{n} x_i^2\right) - n\bar{x}_n^2 \\ &= \left(\sum_{1}^{n} x_i^2\right) - \frac{1}{n}\left(\sum_{1}^{n} x_i\right)^2\end{aligned}$$

Next, consider two properties of the sample variance.

(1) S_n^2 is location invariant. That is, the sample of data $x_1, \dots, x_n$ has the same sample variance as the transformed values

$$y_1 := x_1 + c \;,\; y_2 := x_2 + c \;,\; \dots,\; y_n := x_n + c,$$

for any constant c.

Proof
$$\begin{aligned}\sum_1^n (y_i - \bar{y}_n)^2 &= \sum_1^n \left(x_i + c - \left[\tfrac{1}{n}\sum_1^n (x_j + c)\right]\right)^2 \\ &= \sum_1^n \left(x_i + c - \left(\tfrac{1}{n}\sum_1^n x_j\right) - \tfrac{1}{n}\sum_1^n c\right)^2 \\ &= \sum_1^n (x_i + c - \bar{x}_n - c)^2 \\ &= \sum_1^n (x_i - \bar{x})^2\end{aligned}$$

Hence, $\frac{1}{n-1}\sum_1^n (y_i - \bar{y}_n)^2 = \frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2$ #

(2) If $y_1 = cx_1$, $y_2 = cx_2$, $\dots$, $y_n = cx_n$ for any constant c, then the sample variance of the values $y_1, \dots, y_n$ is the sample variance of the $x_1, \dots, x_n$ scaled by c^2.

Proof.
$$\begin{aligned}\sum_1^n (y_i - \bar{y}_n)^2 &= \sum_1^n \left(cx_i - \left[\tfrac{1}{n}\sum_1^n cx_j\right]\right)^2 \\ &= \sum_1^n c^2 \left(x_i - \tfrac{1}{n}\sum_1^n x_j\right)^2 \\ &= c^2 \sum_1^n (x_i - \bar{x}_n)^2\end{aligned}$$

So $\frac{1}{n-1}\sum_1^n (y_i - \bar{y}_n)^2 = c^2 \cdot \frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2$

And $\sqrt{\frac{1}{n-1}\sum_1^n (y_i - \bar{y}_n)^2} = |c| \sqrt{\frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2}$ #

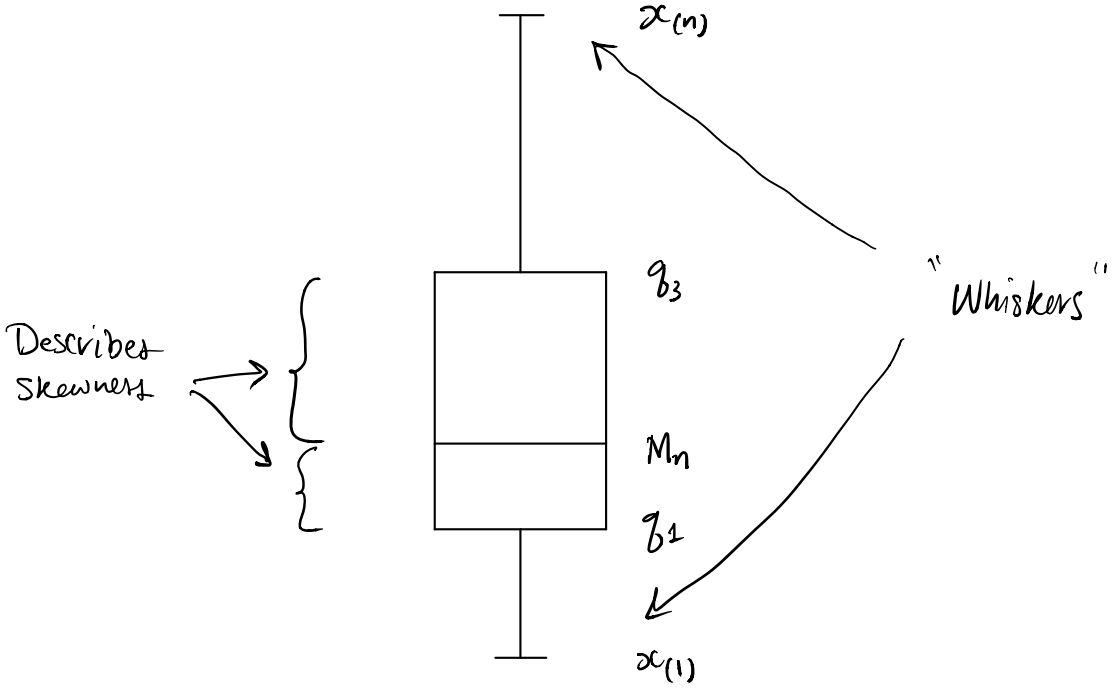
The last topic of chapter 1 is boxplots. These are super useful for presenting a summary of data, and are inherently robust to extreme values in the data. A boxplot is most commonly constructed from the five number summary,

$$x_{(1)} \;,\; q_1 \;,\; M_n \;,\; q_3 \;,\; x_{(n)}$$

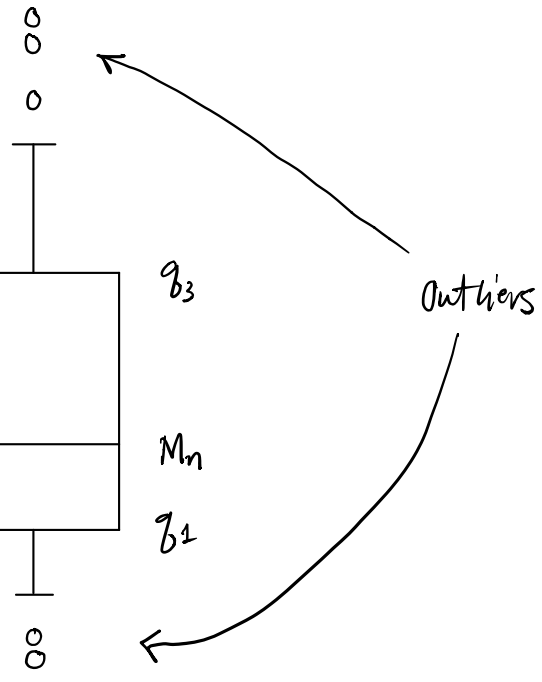
where q_1 and q_3 are the 25th and 75th quantiles, respectively. Precisely, q_1 is the median of the lower half of the ordered set of data values, where the lower half are all values less than the median, M_n. Similarly, q_3 is the median of the upper half of the observed data values.

Commonly, an outlier is defined as any data value that exceeds $1.5(q_3 - q_1)$ distance from q_1 and q_3. The quantity $q_3 - q_1$ is commonly called the interquartile range.

Box plots are often presented in the form



Alternatively, the whiskers can be shortened to extend only to the smallest and largest observed values that are not considered outliers. That may look like



Read Section 1.4 in Devore.

CHAPTER 2. PROBABILITY

This chapter begins our rigorous treatment of studying and quantifying uncertainty.

SECTION 2.1. SAMPLE SPACES AND EVENTS.

DEFINITION. An experiment is any activity or process whose outcome is subject to uncertainty.

We will use the term "experiment" loosely throughout this semester. It can refer to a controlled scientific investigation such as clinical trials, or to a process as simple as tossing a coin.

DEFINITION. The sample space, denoted by S, is the set of all possible outcomes of an experiment.

For example, tossing a coin has two possible outcomes, H or T. So

$$S = \{H, T\}$$

If instead the experiment consisted of tossing two coins, then

$$S = \{HH, HT, TH, TT\}.$$

DEFINITION. An event is any subset of outcomes contained in the sample space S.

For example, if the experiment consists of tossing two coins, then $A = \{HH, TH\}$ is an event since $A \subseteq S$.

In probability theory, events are precise mathematical formulations of outcomes of interest. The basic idea is that we can quantify the uncertainty about outcomes that can be expressed within a certain class of events. We cannot quantify the uncertainty about outcomes that cannot be expressed in such a way.

The use of the word "event" is virtually synonomous with the word "set," in the context of this course.

So what system of logic applies when working with sets?
→ Basic results from set theory.

DEFINITION. The union of two events A and B, denoted by $A \cup B$, is the event containing all outcomes in either A or B.

EXAMPLE. $A = \{HH\}$, $B = \{TH\}$, then $A \cup B = \{HH, TH\}$

$C = \{HH, TT\}$, then $A \cup C = \{HH, TT\} = C$

Note that "or" in mathematics is inclusive.

DEFINITION. The intersection of two events A and B, denoted by $A \cap B$, is the event consisting of all outcomes in both A and B.

EXAMPLE. Take A, B, C as in the previous example. Then

$$A \cap B = \emptyset$$

$$A \cap C = \{HH\} = A$$

Let $D = \{HT, TT\}$. Then $C \cap D = \{TT\}$

DEFINITION. The complement of an event A is the event containing all of the outcomes in $\mathcal{S}$ that are not in A. The complement of A is denoted as A^c.

EXAMPLE. $A^c = \{TH, HT, TT\}$, $B^c = \{HH, HT, TT\}$

$C^c = \{HT, TH\}$, $D^c = \{TH, HH\}$

$$(A \cup B)^c = \{HT, TT\} = A^c \cap B^c$$

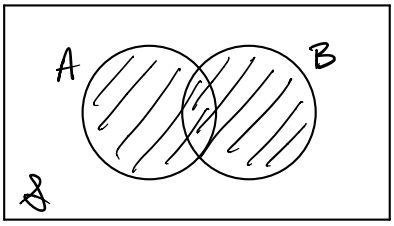
$$(A \cap B)^c = \emptyset^c = \mathcal{S} = A^c \cup B^c$$

DEFINITION. The set difference of two events A and B, denoted by $A \setminus B$, is the event containing all outcomes included in A that are not also included in B.

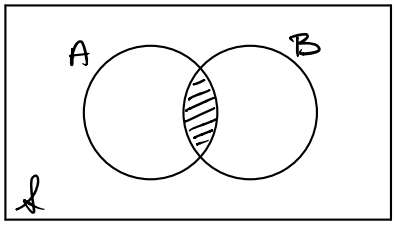
EXAMPLE. $C \setminus A = \{TT\}$, $A \setminus C = \emptyset$

Observe that for any event $E \subseteq \mathcal{S}$, $\mathcal{S} \setminus E = E^c$.

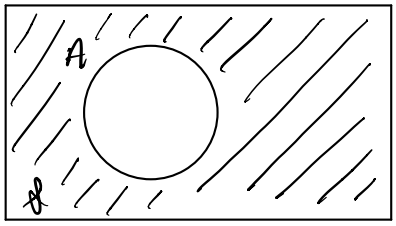
The pictures to have in mind when dealing with set operations are the following.



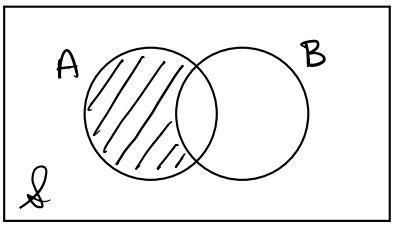
$A \cup B$



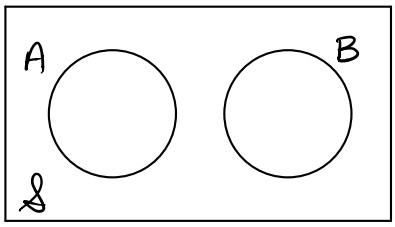
$A \cap B$



A^c



$A \setminus B$



$A \cap B = \emptyset$

Mutually exclusive events A and B.

DEFINITION. Two events A and B are said to be mutually exclusive or disjoint if $A \cap B = \emptyset$.

Read Section 2.1 in Devore.

SECTION 2.2. AXIOMS, INTERPRETATIONS, AND PROPERTIES OF PROBABILITY.

As I described at the beginning of section 2.1, the purpose for defining the notion of an event is to construct a class of events for which we can quantify their likelihood of occurring. We quantify these likelihoods by constructing an event or set function, taking values in the interval $[0,1]$, called a probability measure.

To have a relative likelihood interpretation for quantifying the uncertainty about events $A \subseteq S$, a probability measure must satisfy the following axioms. Denote a probability measure by $P : S \longrightarrow [0,1]$.

(1) For any event $A \subseteq S$, $P(A) \geq 0$

(2) $P(S) = 1$

(3) For any sequence $A_1, A_2, \ldots$ of disjoint events contained in S,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

EXAMPLE. Let $S = \{HH, HT, TH, TT\}$, and

Probability	1/4	1/4	1/4	1/4
Outcome	HH	HT	TH	TT

Then $P(S) = 1/4 + 1/4 + 1/4 + 1/4 = 1$.

Let $A = \{HH, HT\}$ and $B = \{TT\}$, so that $A \cap B = \emptyset$

Then $P(A \cup B) = P(\{HH, HT, TT\}) = 3/4$

$P(A) + P(B) = \frac{2}{4} + \frac{1}{4} = 3/4 = P(A \cup B)$

From Axioms 1, 2, and 3 we can derive a variety of other properties of probability measures.

PROPOSITION. $P(\emptyset) = 0$.

Proof. Define the sequence of events $A_1 := \emptyset$, $A_2 := \emptyset$, $A_3 := \emptyset$, ... Then

$$A_i \cap A_j = \emptyset \cap \emptyset = \emptyset \quad \forall i, j.$$

Next, observe that $\bigcup_{i=1}^{\infty} A_i = \emptyset$

Hence, by axiom 3,

$$P(\emptyset) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) = P(\emptyset) + P(\emptyset) + \cdots,$$

and so

$$0 = \sum_{i=2}^{\infty} P(\emptyset)$$

Since $P(\emptyset) \geq 0$ by axiom 1, it must be the case that $P(\emptyset) = 0$. #

PROPOSITION. For any finite collection of disjoint events $A_1, A_2, \ldots, A_K$,

$$P\left(\bigcup_{i=1}^{K} A_i\right) = \sum_{i=1}^{K} P(A_i).$$

Proof. Define the sequence $A_{K+1} := \emptyset$, $A_{K+2} := \emptyset$, ... Then the sequence $A_1, A_2, \ldots, A_K, A_{K+1}, A_{K+2}, \ldots$ is a disjoint sequence of events. As such, axiom 3 along with the previous proposition gives

$$\begin{aligned} P\left(\bigcup_{i=1}^{K} A_i\right) &= P\left(\bigcup_{i=1}^{\infty} A_i\right) \\ &= \sum_{i=1}^{\infty} P(A_i) \\ &= \sum_{i=1}^{K} P(A_i) + \sum_{i=K+1}^{\infty} P(\emptyset) \\ &= \sum_{i=1}^{K} P(A_i) \end{aligned}$$

#

PROPOSITION. For any event A, $P(A) + P(A^c) = 1$.
Equivalently, $P(A) = 1 - P(A^c)$ or $P(A^c) = 1 - P(A)$.

Proof. Observe that, by definition, $A \cap A^c = \emptyset$. Then let $B_1 := A$ and $B_2 := A^c$. Accordingly, the previous proposition gives

$$\begin{aligned} P(A) + P(A^c) &= \sum_{i=1}^{2} P(B_i) \\ &= P\left(\bigcup_{i=1}^{2} B_i\right) \\ &= P(\Omega) \\ &= 1 \quad , \text{ by axiom 2.} \end{aligned}$$

#

PROPOSITION. For any event A, $P(A) \leq 1$.

Proof. By axiom 1, $P(A^c) \geq 0$. Then

$$-P(A^c) \leq 0$$

and so

$$P(A) - P(A^c) \leq P(A)$$

which yields

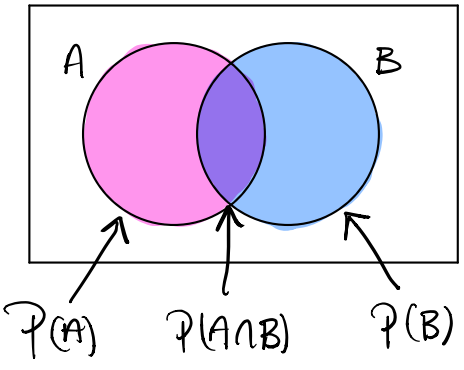
$$P(A) \leq P(A) + P(A^c) = 1.$$

#

PROPOSITION. For any two events A and B,

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

Proof. First think about what the picture of this relation looks like.



Observe that $P(A) + P(B)$ "double counts" the region given by $A \cap B$. So the intuition is that we need to subtract off one copy.

To formally prove the proposition, the strategy is to express the arbitrary sets A and B as a disjoint union. That is,

$$A \cup B = (A \setminus B) \cup B$$

Then

$$\begin{aligned} P(A \cup B) &= P((A \setminus B) \cup B) \\ &= P(A \setminus B) + P(B) \quad \text{by axiom 3.} \end{aligned}$$

But how to compute $P(A \setminus B)$?

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

$= A \setminus B$

disjoint union

Thus,

$$P(A) = P((A \cap B) \cup (A \setminus B))$$
$$= P(A \cap B) + P(A \setminus B) \quad \text{by axiom 3.}$$

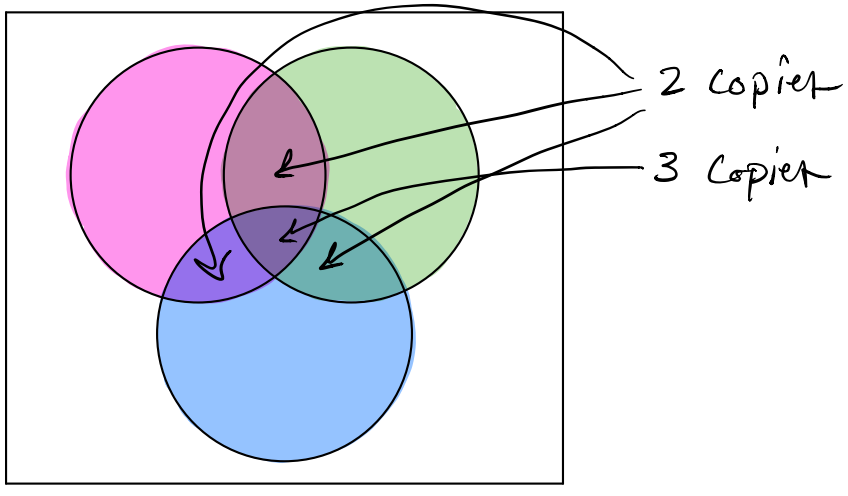
From here we get that $P(A \setminus B) = P(A) - P(A \cap B)$. And so

$$P(A \cup B) = P(A \setminus B) + P(B)$$
$$= P(A) - P(A \cap B) + P(B). \qquad \#$$

PROPOSITION. For any three events A, B, C,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$

The intuition is the same as in the case of the arbitrary union of two events. When we add the probabilities $P(A) + P(B) + P(C)$, we "double count" the regions $A \cap B$, $A \cap C$, $B \cap C$, and so we must subtract off. However, in doing so in this case we remove all three copies of the region $A \cap B \cap C$, and so we must add it back.



How to prove this proposition, though?

Proof. First let's parse the expression in the form of simpler problems that we have already solved. Accordingly, if we denote

$$D := B \cup C,$$

then as we have now shown,

$$P(A \cup B \cup C) = P(A \cup D)$$
$$= P(A) + P(D) - P(A \cap D)$$
$$= P(A) + P(B \cup C) - P(A \cap (B \cup C))$$

Further, observe that $P(B \cup C) = P(B) + P(C) - P(B \cap C)$. So

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(B \cap C) - P(A \cap (B \cup C)).$$

Next, let's argue that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

$\rightarrow$ If $x \in A \cap (B \cup C)$, then $x \in A$ and $x \in (B \cup C)$

So $\underbrace{x \in A \text{ and } x \in B}_{x \in A \cap B}$ or $\underbrace{x \in A \text{ and } x \in C}_{x \in A \cap C}$

$$\underbrace{\qquad\qquad\qquad\qquad}_{x \in (A \cap B) \cup (A \cap C)}$$

Thus, $A \cap (B \cup C) \subseteq (A \cap B) \cup (A \cap C)$.

For the other direction, if $x \in (A \cap B) \cup (A \cap C)$.

then $x \in A \cap B$ or $x \in A \cap C$, so that

$x \in A$ and $x \in B$, or $x \in A$ and $x \in C$

In any case, $x \in A$ and $\underbrace{x \in B \text{ or } x \in C}_{x \in B \cup C}$

$$\underbrace{\qquad\qquad\qquad\qquad}_{x \in A \cap (B \cup C)}.$$

Hence, $A \cap (B \cup C) \supseteq (A \cap B) \cup (A \cap C)$.

Which proves that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Finally, where we left off,

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

which concludes the argument since

$$P((A \cap B) \cap (A \cap C)) = P(A \cap B \cap C). \qquad \#$$

Read Section 2.2 in Devore.

SECTION 2.3. COUNTING TECHNIQUES

Counting may sound simple enough, but consider this section!

The motivation for developing strategies for counting within the context of this course is that the situation will arise often in which there are many outcomes, all having equal probability, that occur, and so the uncertainty quantifications amounts to a counting problem.

For example, suppose I toss two six-sided, fair dice. Then there are 36 outcomes, each having the same probability. If I want to know the probability of observing the sum of the two tosses being 7, then the strategy is to count how many outcomes are contained in this event and scale by the total possible number of outcomes. Denoting the event by A,

$$A = \{ \underbrace{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)}_{\text{6 outcomes sum to 7, each having equal probability}} \}$$

So $P(A) = \frac{6}{36} = \frac{1}{6}$ Since there are $6 \cdot 6 = 36$ outcomes total

1	2	3	4	5	6
1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6

The basic idea is that if an experiment can be expressed in terms of a sample space S of outcomes that are all equally likely, then for any event A,

$$P(A) = \frac{N(A)}{N},$$

where $N(A)$ is the number of outcomes contained in the event A, and N is the number of elements in S.

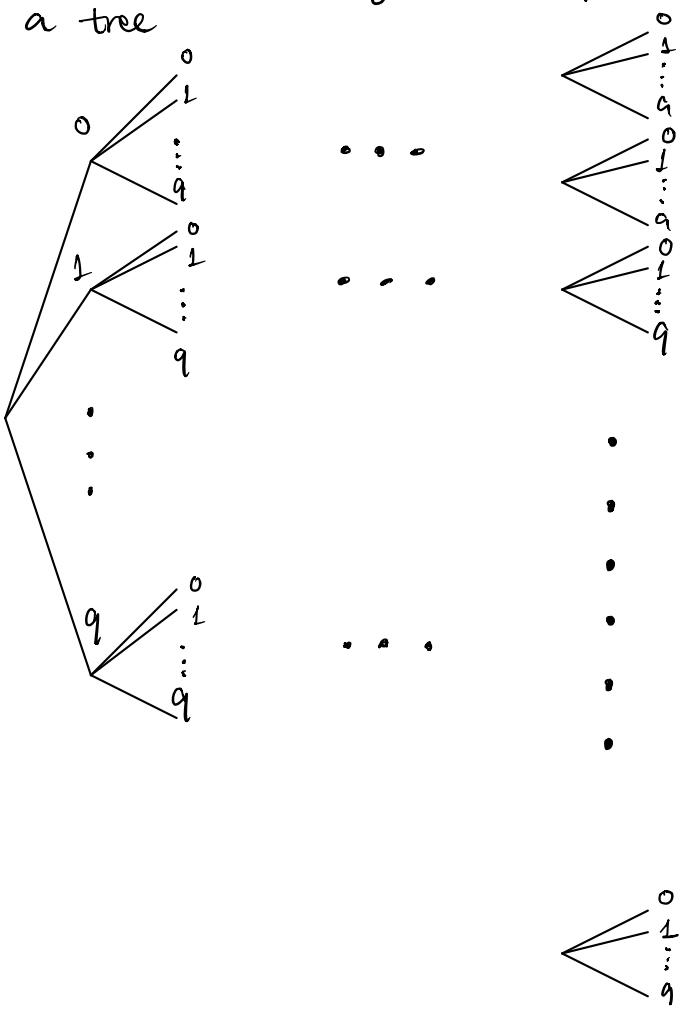
The counting problems we will discuss can generally be described as falling

into one of two categories.

(1) Number of distinct k-tuples described by some event.

For example, how many ten digit phone numbers can be constructed from the ten numbers $\{0, 1, 2, \ldots, 9\}$?

The enumeration of all such phone numbers can expressed as a tree



(10 ways to choose the first)
$\times$ (10 ways to choose the second)
$\vdots$
$\times$ (10 ways to choose the fourth)

$= 10 \cdot 10 \cdots 10 = 10^{10}$

In general, if there are n_1 items to choose from for the first component of the k-tuple, n_2 items for the second, $\ldots$, n_k items for the kth component, then the total number of possible k-tuples is

$$n_1 \cdot n_2 \cdots n_k = \prod_{i=1}^{k} n_i$$

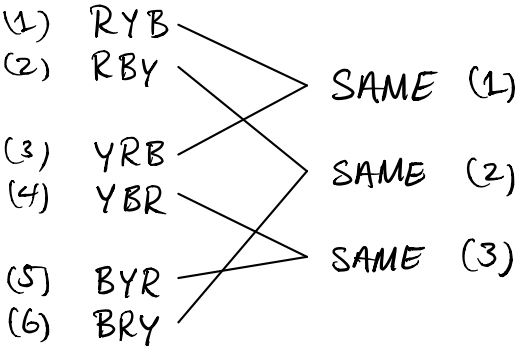
(2) Number of ways that a finite collection of objects can be ordered.
For example, consider a collection of three marbles $\{R, Y, B\}$.
How many ways are there to order these three marbles?
There should be $3 \cdot 2 \cdot 1 = 6$ ways:

$R \begin{cases} Y - B \\ B - Y \end{cases}$ (1) RYB (2) RBY

$Y \begin{cases} R - B \\ B - R \end{cases}$ (3) YRB (4) YBR

$B \begin{cases} Y - R \\ R - Y \end{cases}$ (5) BYR (6) BRY

How many ways to select two out of the three marbles?
Observe from before that with respect to the first two components,



In other words, there were $3 \cdot 2 \cdot 1 = 3!$ ways of ordering the collection $\{R, Y, B\}$, the last $1 = (3-2)!$ component does not matter, and the order of the first $2 = 2!$ components does not matter. Thus, that is,

$$\frac{3!}{(3-2)! \cdot 2!} = \frac{3 \cdot 2 \cdot 1}{1 \cdot 2 \cdot 1} = 3$$

ways to select two out of the three marbles. Alternatively, note that if the order of the first two components did matter, then there would be

$$\frac{3!}{(3-2)!} = 3 \cdot 2 = 6$$

Combinations, still.

In general, if there are n total items to choose from, then there are

$$C_{k,n} := \binom{n}{k} := \frac{n!}{(n-k)! \cdot k!} \qquad \text{"}n \text{ choose } k\text{"}$$

ways to choose k of them if the order of the k items does not matter, and

$$P_{k,n} := \frac{n!}{(n-k)!}$$

if the order does matter. An unordered subset is called a combination, while an ordered subset is called a permutation. The textbook has a nice example to further explain these expressions.

Assume that there are 7 departments within the college of engineering, and that 3 of the 7 will be selected to serve in the student council as either the chair, the vice chair, or the secretary. Then there are 7 choices for the chair, 6 remaining choices for the vice chair, and 5 choices left for the secretary. That is, there are

$$7 \cdot 6 \cdot 5 = \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{4 \cdot 3 \cdot 2 \cdot 1} = \frac{7!}{4!} = \frac{7!}{(7-3)!}$$

permutations to select from. Now, if it only matters that 3 departments are selected (i.e., the role assigned to each of the 3 is arbitrary), then since there are 3! ways to permute the chosen 3 departments, say

(1) (2) (3) (4) (5) (6)
a, b, c; a, c, b; b, a, c; b, c, a; c, a, b; c, b, a,

there are

$$\frac{7 \cdot 6 \cdot 5}{3!} = \frac{7!}{(7-3)!\,3!} = \binom{7}{3}$$

combinations.

Note that by convention $0! = 1$, so

$$\binom{n}{0} = \frac{n!}{(n-0)!\,0!} = 1$$

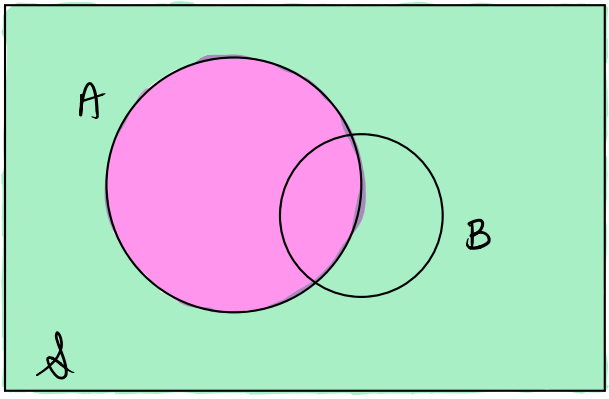
$$\binom{n}{n} = \frac{n!}{(n-n)!\,n!} = 1.$$

Read Section 2.3 in Devore.

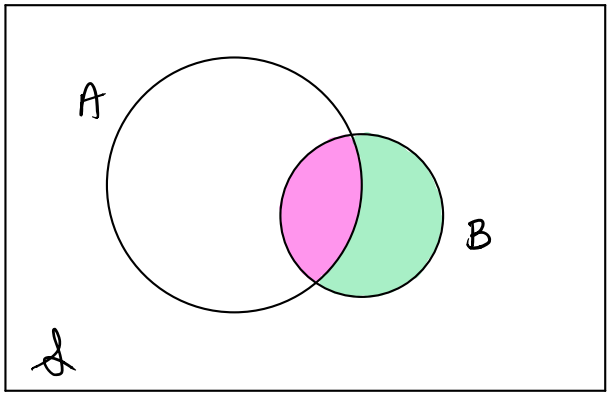
SECTION 2.4. CONDITIONAL PROBABILITY.

Up to this point we have developed ideas for how to define the probability of an event A. In this section we consider how to update our belief about the

probability that A will occur with the observation of partial information, B, relating to A. In doing so, we formalize the notion of conditional probability.



Unconditional probability of A



Conditional probability of A given B

If we know that an event B has occured, then to quantify the uncertainty about further events we must account for the new sample space given by the outcomes in B. Accordingly, the event A is reduced to the event $A \cap B$, and the probability must be renormalized so that the conditional probability of B given B is one. This is accomplished by scaling by $P(B)$ because

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

Thus, the conditional probability of the event A given B, denoted $A|B$, is

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

EXAMPLE. Consider the experiment of tossing two coins such that the sample space is given by

$$S = \{HH, HT, TH, TT\}$$

and let $A = \{HH, TT\}$ and B is the event that one of the tosses is H. Then $B = \{HH, HT, TH\}$,

$$P(A) = 2/4$$

$$P(B) = 3/4$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{HH\})}{3/4} = \frac{1/4}{3/4} = 1/3.$$

Note that for a conditional probability to be meaningful, there must be a nonzero probability that the event B occurs. For one, if there is no probability that B will occur, then there is no meaning to the event $A|B$. For two,

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

is not defined if $P(B) = 0$.

EXAMPLE. Continuing with the previous example, consider the event $C = \{TT\}$. Then

$$P(C|B) = \frac{P(C \cap B)}{P(B)} = \frac{P(\emptyset)}{P(B)} = 0$$

This example illustrates the fact that if two events are disjoint (i.e., $C \cap B = \emptyset$), then the knowledge that one occurs is knowledge that the other has not occurred. This observation will be useful later on when we consider the notion of independence for events.

From the construction of conditional probabilities yields the result, commonly referred to as the multiplication rule, that for any events $A, B \subseteq S$

$$P(A \cap B) = P(A|B) \cdot P(B)$$

or equivalently,

$$P(A \cap B) = P(B|A) \cdot P(A).$$

There are a variety of uses of this result that we will see throughout the semester. For now, it serves as a tool to possibly simplify certain probability statements. For example,

$$P(A \cap B \cap C) = P(A \mid B \cap C) \cdot P(B \cap C)$$

$$= P(A \mid B \cap C) \cdot P(B|C) \cdot P(C)$$

Now we are ready for an important result about the role that conditional probabilities play in the context of unconditional probabilities.

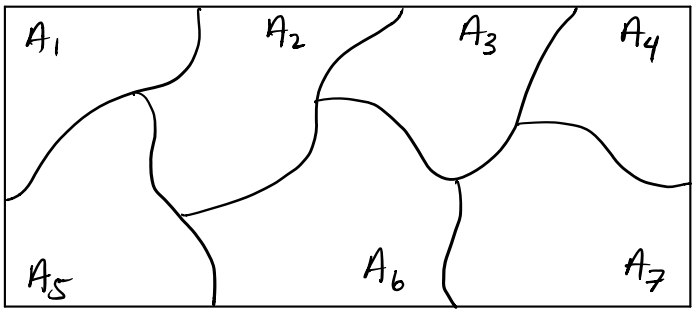
THEOREM (The law of total probability). Let $A_1, \ldots, A_k$ be a collection of mutually exclusive but collectively exhaustive events. Then for any event B,

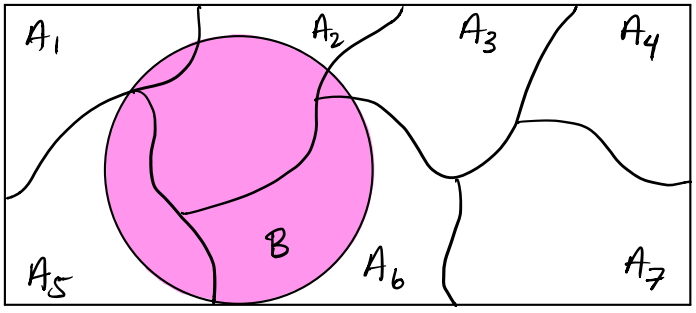
$$P(B) = \sum_{i=1}^{k} P(B|A_i) \cdot P(A_i).$$

Proof. Since the A_i are mutually exclusive, $A_i \cap A_j = \emptyset \ \forall i, j \in \{1, \ldots, k\}$ with $i \neq j$. Collectively exhaustive means that

$$\bigcup_{i=1}^{k} A_i = A_1 \cup A_2 \cup \cdots \cup A_k = S$$

Let's first parse what this statement says, using the following pictures.





The intuition gathered from this picture illustrates that for any set B,

$$\begin{aligned} B &= B \cap \Omega \\ &= B \cap (A_1 \cup A_2 \cup \cdots \cup A_k) \\ &= (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_k) \end{aligned}$$

where $\forall\, i, j \in \{1, \ldots, k\}$ with $i \neq j$,

$$(B \cap A_i) \cap (B \cap A_j) = B \cap A_i \cap A_j = \emptyset \qquad \text{since } A_i \cap A_j = \emptyset.$$

Thus, by axiom 3 of a probability measure,

$$\begin{aligned} P(B) &= P\left(\bigcup_{i=1}^{k} (B \cap A_i) \right) \\ &= \sum_{i=1}^{k} P(B \cap A_i) \\ &= \sum_{i=1}^{k} P(B \mid A_i) \cdot P(A_i). \end{aligned}$$

#

A reverse statement of this theorem is also true, commonly referred to as Bayes Theorem or inverse probability.

THEOREM (Bayes Theorem). Let $A_1, \ldots, A_k$ be a collection of mutually exclusive and collectively exhaustive events. Then for any event B with $P(B) > 0$, $\forall\, j \in \{1, \ldots, k\}$,

$$P(A_j|B) = \frac{P(B|A_j)\cdot P(A_j)}{\sum_{i=1}^{k} P(B|A_i)\cdot P(A_i)}$$

Proof. By the definition of conditional probability,

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)\cdot P(A_j)}{\sum_{i=1}^{k} P(B|A_i)\cdot P(A_i)},$$

where the second equality follows from the law of total probability. #

EXAMPLE. Let's look at example 2.31 in the textbook.

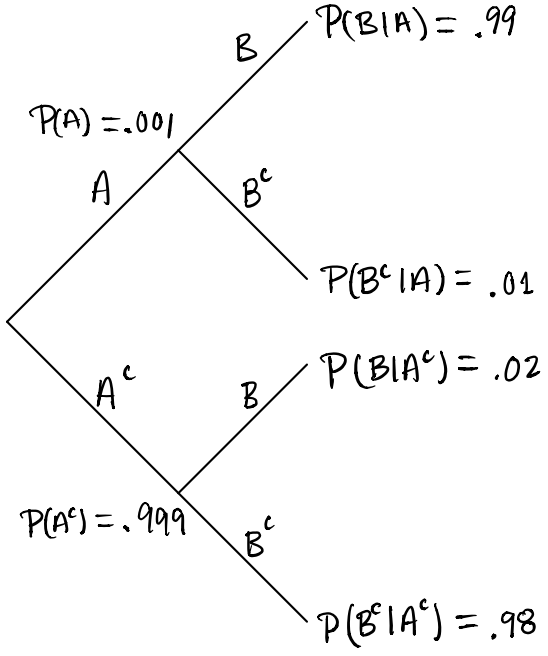
Facts:

(1) 1 in 1000 adults is afflicted with a rare disease
(2) a diagnostic test has been developed
(3) When an adult has the disease the test is positive 99% often.
(4) Without the disease the test is positive 2% often.

If a randomly selected adult tests positive, what is the probability that they have the disease?

Let $A = \{\text{the adult has the disease}\}$
$B = \{\text{positive test result}\}$

Then the question asks to evaluate $P(A|B)$. To determine this probability, consider the following tree.



So

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

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

$$= \frac{.99 \cdot .001}{.99 \cdot .001 + .02 \cdot .999}$$

$$\approx .0472$$

Read Section 2.4 in Devore.

SECTION 2.5. INDEPENDENCE.

Consider the situation where the likelihood that an event A will occur is not affected by the knowledge that some event B has occurred. In such a scenario we say that the events A and B are independent. This notion is stated more precisely in the following definition.

DEFINITION. Two events A and B are said to be independent if

$$P(A|B) = P(A).$$

Events A and B are said to be dependent if they are not independent.

Note that by the definition of conditional probability an equivalent condition for independence is that

$$P(B|A) = P(B)$$

Since

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

which gives

$$P(B) = P(B|A).$$

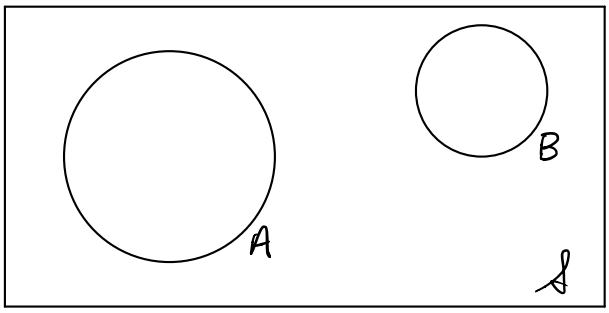
Alternatively, another equivalent condition for independence is the product rule that

$$P(A \cap B) = P(A|B) \cdot P(B) = P(A) \cdot P(B)$$

Or

$$P(A \cap B) = P(B|A) \cdot P(A) = P(B) \cdot P(A).$$

Next, consider how independence relates to mutual exclusivity.



If events A and B are mutually exclusive and we know that B happened, for instance, then we know that A could not have happened. Hence, A and B are not independent. Suppose that $P(A) > 0$ and $P(B) > 0$. Then,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\emptyset)}{P(B)} = 0 \neq P(A).$$

EXAMPLE. Consider again the process of tossing two fair coins. That is,

$$S = \{HH, HT, TH, TT\}.$$

Let A be the event that the first toss results in H, and let B be the event that the second toss results in T. Then A and B are independent since

$$\begin{aligned} P(A \cap B) &= P(\{HT\}) \\ &= \frac{1}{4} \\ &= \frac{1}{2} \cdot \frac{1}{2} \\ &= P(\{HH, HT\}) \cdot P(\{HT, TT\}) \\ &= P(A) \cdot P(B). \end{aligned}$$

Our notion of independence generalizes to more than two events with the following definition.

DEFINITION. Events $A_1, \ldots, A_n$ are mutually independent if for every $k \in \{2, \ldots, n\}$ and every collection of indices $i_1, \ldots, i_k \in \{1, \ldots, n\}$

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k}).$$

Read Section 2.5 in Devore.

CHAPTER 3. DISCRETE RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS.

SECTION 3.1. RANDOM VARIABLES.

In the context of an experiment, as we have seen many times in Chapter 2, we associate the outcome of the experiment with a numerical value. We call such an association, or mapping, a random variable, and the notion of a random variable is generalizable and important enough that we will spend the next three chapters studying it.

DEFINITION. For a given sample space $\mathcal{S}$ of some experiment, a random variable is any function $X: \mathcal{S} \longrightarrow \mathbb{R}$, where $\mathbb{R}$ is the set of all real numbers.

Important notation for random variables:

Upper case letters such as X, Y, Z are used to denote a random variable. That is, X, Y, and Z describe functions. Conversely, Lower case letters such as x, y, z are used to denote particular observations of the random variables X, Y, Z, respectively. What this means is that we have observed some outcome $\omega \in \mathcal{S}$, and

$$x := X(\omega), \quad y := Y(\omega), \quad \text{and} \quad z := Z(\omega).$$

EXAMPLE. For the experiment of tossing a coin, where $\mathcal{S} = \{H, T\}$, we can define the random variable X such that $\forall\, \omega \in \mathcal{S}$,

$$X(\omega) = \begin{cases} 1 & \text{if } \omega = H \\ 0 & \text{if } \omega = T \end{cases}.$$

Moreover, any random variable taking values 0 or 1 is called a Bernoulli random variable. Bernoulli random variables can be used to describe any binary outcome/event.

For instance, $A = \{\text{roll a dice and observe a number greater than } 3\}$
Then

$$X(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}.$$

Or

$B = \{\text{it rains tomorrow}\}$

Then

$$Y(\omega) = \begin{cases} 1 & \text{if } \omega \in B \\ 0 & \text{if } \omega \notin B \end{cases}.$$

Sometimes we refer to 1 and 0 as success and failure, respectively.

A Bernoulli random variable is a particular instance of a random variable, but more generally random variables are categorized as belonging to one of the following two categories. Recall the notions of discrete and continuous defined in Chapter 1.

DEFINITION. A random variable is called discrete if it takes values on a finite or countably infinite set of real numbers.
A random variable is called continuous if both of the following conditions are satisfied.

(i) The random variable takes values on sets consisting of one or more entire intervals of real numbers.

(ii) The probability of observing any particular value, say $c \in \mathbb{R}$, is zero. That is, if X is a continuous random variable, then

$$P(X = c) = 0 \qquad \forall c \in \mathbb{R}.$$

The remainder of Chapter 3 studies discrete random variables, while Chapter 4 studies continuous random variables.

Read Section 3.1 in Devore.

SECTION 3.2. PROBABILITY DISTRIBUTIONS FOR DISCRETE RANDOM VARIABLES.

Given a random variable X, the "probability distribution" of X is an attribute that precisely describes how the total probability of one is allocated to the X values of events. Further, a probability distribution can be expressed in a variety of different ways. For discrete random variables, often the most convenient is the probability mass function, defined next.

DEFINITION. The probability mass function (pmf) of a discrete random variable is defined for every $x \in \mathbb{R}$ as

$$p(x) := P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\}).$$

The two defining properties of a pmf are:

(i) $p(x) \geq 0 \quad \forall x \in \mathbb{R}$

(iii) $\sum_{\text{all } x} p(x) = 1.$

EXAMPLE. For the experiment of tossing a coin, where $\mathcal{S} = \{H, T\}$, and the random variable X such that $\forall \omega \in \mathcal{S}$,

$$X(\omega) = \begin{cases} 1 & \text{if } \omega = H \\ 0 & \text{if } \omega = T \end{cases},$$

The pmf of X is

x	1	0
$p(x)$	α	$1-\alpha$

,

where $\alpha := P(X=1)$ is the probability of 1 or H or success.

EXAMPLE. Let Y be the value of the roll of a dice. Then $Y \in \{1,2,3,4,5,6\}$ so that Y is a discrete random variable. What is the pmf of Y?

y	1	2	3	4	5	6
$p(y)$	α_1	α_2	α_3	α_4	α_5	α_6

,

where $\sum_{i=1}^{6} \alpha_i = 1$ and $\alpha_i \geq 0 \quad \forall i \in \{1, \dots, 6\}$.

Observe that in the case where the dice is fair,

$$\alpha_1 = \alpha_2 = \cdots = \alpha_6$$

so that

$$1 = \sum_{i=1}^{6} \alpha_i = \sum_{i=1}^{6} \alpha = 6\alpha$$

which gives

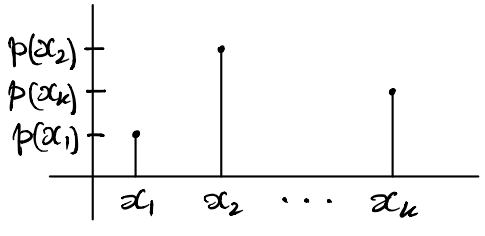
$$\alpha = \frac{1}{6}.$$

In addition to the tabular presentation of a pmf,

x	x_1	x_2	$\cdots$	x_k
$p(x)$	$p(x_1)$	$p(x_2)$	$\cdots$	$p(x_k)$

,

there are various graphical presentations such as



DEFINITION. a parameter is defined as a quantity whose value characterizes a probability distribution. The collection of all probability distributions for different values of the parameter is called a family of probability distributions.

For example, we have seen the pmf of a Bernoulli random variable,

x	1	0
$p(x)$	α	$1-\alpha$

for some given value $\alpha \in [0,1]$. In fact, for any value $p \in [0,1]$,

x	1	0
$p(x)$	p	$1-p$

defines a Bernoulli random variable with parameter p. A short hand notation is to write

$$X \sim \text{Bernoulli}(p).$$

Then $X \sim \text{Bernoulli}(\alpha)$ describes the original random variable from the example where $p = \alpha$.

Furthermore, the collection $\{p(x;\alpha) : \alpha \in [0,1]\}$ where $p(x;\alpha)$ is the Bernoulli(α) pmf is the family of Bernoulli distributions.

Note that we could also express

$$p(x;\alpha) = \alpha^x \cdot (1-\alpha)^{1-x} \cdot 1\{x \in \{0,1\}\}$$

Next, we will introduce a second defining expression of a probability distribution called the cummulative distribution function.

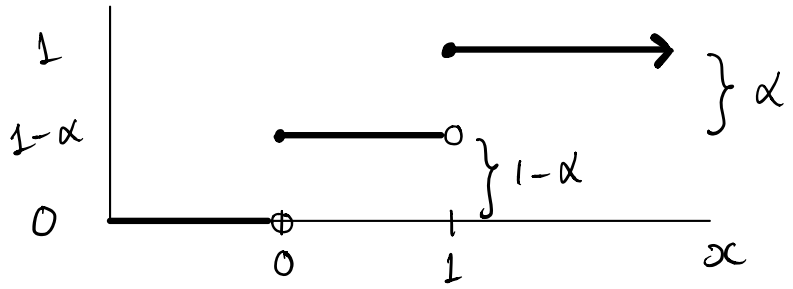
DEFINITION. The cummulative distribution function (cdf), denoted $F(x)$, of a discrete random variable X with pmf $p(x)$ is defined for every real number x by

$$F(x) := P(X \leq x) = \sum_{y : y \leq x} p(y).$$

For example, if $X \sim \text{Bernoulli}(\alpha)$, then

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1-\alpha & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

Graphically, the cdf looks like



PROPOSITION. For any two numbers a and b with $a \le b$,

$$P(a \le X \le b) = F(b) - F(a-),$$

where $a-$ denotes the largest possible value of X that is strictly less than a.

Proof. By definition,

$$F(b) = \sum_{x : x \le b} p(x)$$

and

$$F(a-) = \sum_{x : x \le a-} p(x),$$

where $p(x)$ denotes the pmf of the random variable X. Then

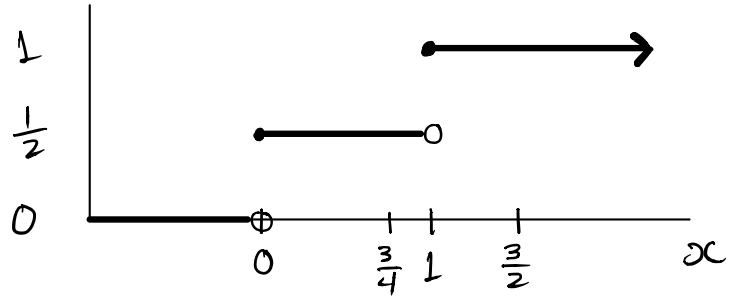
$$\begin{aligned} P(a \le X \le b) &= \sum_{x : a \le x \le b} p(x) \\ &= \sum_{x : a- < x \le b} p(x) \\ &= \underbrace{\sum_{x : x \le b} p(x)}_{= F(b)} - \underbrace{\sum_{x : x \le a-} p(x)}_{F(a-)}. \end{aligned}$$

#

EXAMPLE. Let $X \sim \text{Bernoulli}(1/2)$. Then

$$P(3/4 \le X \le 3/2) = F(3/2) - F(0) = 1 - 1/2 = 1/2,$$

since



Read Section 3.2 in Devore.

SECTION 3.3. EXPECTED VALUES

The textbook motivates this section with the following example. Let X be the number of courses for which a randomly selected student is registered, and assume that X has the frequency distribution

x	1	2	3	4	5	6	7
freq	150	450	1950	3750	5850	2550	300

Note that the sum of all of these frequencies is $n = 15{,}000$. If there are 15,000 students in attendance at the university, then this is the population frequency distribution, and the population mean number of registered courses per student, say μ, is taken to be the average of the population values;

$$\mu = \left(\frac{150}{15000}\right)\cdot 1 + \left(\frac{450}{15000}\right)\cdot 2 + \left(\frac{1950}{15000}\right)\cdot 3 + \left(\frac{3750}{15000}\right)\cdot 4 + \left(\frac{5850}{15000}\right)\cdot 5 + \left(\frac{2550}{15000}\right)\cdot 6 + \left(\frac{300}{15000}\right)\cdot 7$$

$$= 4.57.$$

Equivalently, the pmf of X is

x	1	2	3	4	5	6	7
$p(x)$	.01	.03	.13	.25	.39	.17	.02

and so

$$\mu = (.01)\cdot 1 + (.03)\cdot 2 + (.13)\cdot 3 + (.25)\cdot 4 + (.39)\cdot 5 + (.17)\cdot 6 + (.02)\cdot 7$$

$$= 4.57.$$

We also call the population mean of the distribution of X, the expected value of X.

DEFINITION. Let X be a discrete R.V. with values in some set D, and pmf $p(x)$. Then the expected value or mean value of X, denoted by $E(X)$ or μ, is

$$E(X) = \mu := \sum_{x \in D} x \cdot p(x).$$

Now lets look at the expected value for our two recurring examples.

EXAMPLE. Suppose $X \sim \text{Bernoulli}(\alpha)$ (i.e., a coin toss). Then

x	0	1
$p(x)$	$1-\alpha$	α

and so $E(X) = 0 \cdot (1-\alpha) + 1 \cdot \alpha = \alpha$

That is to say that we expect to observe a 1, α proportion of the time. If $\alpha = .5$ and 1 is H, then we expect the coin to land H half of the time.

EXAMPLE. Assume X is the outcome of rolling a fair dice. Then

x	1	2	3	4	5	6
$p(x)$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

and so

$$\begin{aligned} E(X) &= \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 \\ &= \frac{1}{6} \cdot 21 \\ &= 3.5 \end{aligned}$$

This is interpreted as the average value we would observe over many rolls of the dice.

Next, observe the fact that any function, say $h(\cdot)$, of a random variable X is also a random variable, say $Y := h(X)$.

EXAMPLE. Suppose that I bet 3 dollars that a coin lands H, $X = 1$ if the coin lands H, and that $P(X=1) = \alpha$. Then the rule for how much I make after one coin toss is

$$h(x) = \begin{cases} 3 & \text{if } x = 1 \\ -3 & \text{if } x = 0 \end{cases}$$

accordingly, the distribution of $Y := h(X)$ is

y	3	-3
$p(y)$	α	$1-\alpha$

and

$$E[h(X)] = E(Y) = \alpha \cdot 3 + (1-\alpha)(-3)$$

In general,

PROPOSITION. Let X be a discrete R.V. with values in some set D, and pmf $p(x)$. Then the expected value of any function $h(X)$ is

$$E[h(X)] = \sum_{x \in D} h(x) \cdot p(x).$$

Proof. Let $x_1, x_2, \ldots$ denote the discrete values in the set D. Then X has pmf

x	x_1	x_2	$\cdots$
$p(x)$	$p(x_1)$	$p(x_2)$	$\cdots$

which meant that $h(X)$ had pmf

$h(x)$	$h(x_1)$	$h(x_2)$	$\cdots$
$p(x)$	$p(x_1)$	$p(x_2)$	$\cdots$

and so by definition

$$E[h(X)] = \sum_{h(x) : x \in D} h(x)\, p(x) = \sum_{x \in D} h(x)\, p(x) \qquad \#$$

A commonly considered function is a linear function $h(x) = ax + b$.

PROPOSITION. Let X be a discrete R.V. Then

$$E[h(X)] = a\,E(X) + b.$$

Proof. By the previous proposition,

$$\begin{aligned} E[h(X)] &= \sum_{x \in D} h(x)\, p(x) \\ &= \sum_{x \in D} (ax + b)\, p(x) \\ &= \sum_{x \in D} \Big(ax\, p(x) + b\, p(x) \Big) \\ &= a \cdot \underbrace{\sum_{x \in D} x \cdot p(x)}_{= E(X)} + b \cdot \underbrace{\sum_{x \in D} p(x)}_{= 1} \\ &= a \cdot E(X) + b. \end{aligned}$$

$\#$

Now that we have defined the population mean for discrete random variables, and studied some of its properties, we will next define the notion of the population variance.

DEFINITION. Let X be a discrete R.V. with pmf $p(x)$ and expected value μ. Then the variance of X, denoted $\text{Var}(X)$ or σ^2, is

$$\text{Var}(X) = \sigma^2 := E[(X-\mu)^2] = \sum_{x \in D} (x-\mu)^2 \cdot p(x),$$

and the population standard deviation is

$$\sigma = \sqrt{\text{Var}(X)}.$$

An alternative expression for the variance is

$$\begin{aligned}
\text{Var}(X) &= E[(X-\mu)^2] \\
&= E[X^2 - 2X\mu + \mu^2] \\
&= E(X^2) - 2\mu E(X) + \mu^2 \\
&= E(X^2) - 2\mu^2 + \mu^2 \\
&= E(X^2) - \mu^2 \\
&= E(X^2) - [E(X)]^2.
\end{aligned}$$

EXAMPLE. Let $X \sim \text{Bernoulli}(\alpha)$, and recall that $E(X) = \alpha$. Then

$$\begin{aligned}
\text{Var}(X) &= (0-\alpha)^2 \cdot (1-\alpha) + (1-\alpha)^2 \cdot \alpha \\
&= \alpha^2 \cdot (1-\alpha) + \alpha \cdot (1-\alpha)^2 \\
&= [\alpha^2 + \alpha \cdot (1-\alpha)] \cdot (1-\alpha) \\
&= [\alpha^2 + \alpha - \alpha^2] \cdot (1-\alpha) \\
&= \alpha \cdot (1-\alpha).
\end{aligned}$$

PROPOSITION. Let X be a discrete RV with pmf $p(x)$. Then for any linear function $h(x) = ax+b$,

$$\text{Var}(aX+b) = a^2 \text{Var}(X).$$

Proof

$$\begin{aligned}
\text{Var}(aX+b) &= \sum_{x \in D} (ax+b - E(aX+b))^2 \cdot p(x) \\
&= \sum_{x \in D} (ax+b - a\mu - b)^2 \cdot p(x),
\end{aligned}$$

where $\mu := E(X)$. Then

$$\mathrm{Var}(aX+b) = \sum_{x \in D} a^2 (x-\mu)^2 \cdot p(x) = a^2 \cdot \underbrace{\sum_{x \in D} (x-\mu)^2 \cdot p(x)}_{= \mathrm{Var}(X)} . \quad \#$$

Observe the direct corollary of this theorem that

$$\sigma_{aX+b} = \sqrt{\mathrm{Var}(aX+b)} = \sqrt{a^2 \, \mathrm{Var}(X)} = |a| \sigma_X .$$

Read Section 3.3 in Devore.

SECTION 3.4. THE BINOMIAL PROBABILITY DISTRIBUTION.

At this point in the semester we have introduced most of the objects and tools that is required to ask meaningful questions in the context of quantifying uncertainty of events. What remains is to understand more fully how these objects and tools can be used.

We have considered many examples and properties of the simple coin toss random variable

$$X \sim \text{Bernoulli}(p).$$

The motivation for this section is to consider experiments in which we observe n independent Bernoulli(p) random variables. Such an experiment is called a binomial experiment and can be summarized as a binomial random variable.

DEFINITION. Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$. Then the random variable

$$Y := \sum_{i=1}^{n} X_i$$

is said to follow the binomial distribution with n trials, each having p probability of success. Notationally, we write

$$Y \sim \text{binomial}(n, p).$$

Just like the outcome of $X_i \overset{iid}{\sim} \text{Bernoulli}(p)$ is described as success (i.e., $X_i = 1$) or failure (i.e., $X_i = 0$), the outcome of

$$Y := \sum_{i=1}^{n} X_i = \text{"the number of successes in } n \text{ trials"}$$

EXAMPLE. Let $X_1, X_2, X_3 \overset{iid}{\sim} \text{Bernoulli}(p)$ represent the outcomes of three independent coin tosses, each having probability p of landing H, where H corresponds to $X_i = 1$. If Y denotes the number of H observed from X_1, X_2, X_3, then $Y \sim \text{binomial}(3, p)$ since X_1, X_2, X_3 are independent and Y can be expressed as

$$Y = X_1 + X_2 + X_3.$$

We can ask questions like, what is the probability of observing exactly 1 H? Such an event occurs if

$$\{x_1, x_2, x_3\} \in \{1, 0, 0\} \cup \{0, 1, 0\} \cup \{0, 0, 1\},$$

where each outcome in the union occurs with probability

$$p \cdot (1-p) \cdot (1-p).$$

Thus $P(\{\text{exactly one } H\}) = 3 \cdot p^1 \cdot (1-p)^2$

$$= \binom{3}{1} p^1 \cdot (1-p)^{3-1}$$

$$= P(Y = 1),$$

where $y = x_1 + x_2 + x_3$, and so $\{\text{exactly one } H\} = \{y = 1\}$. To more fully understand Y, lets construct the full pmf of Y.

$$P(Y=0) = \binom{3}{0} p^0 (1-p)^{3-0}$$

$$P(Y=1) = \binom{3}{1} p^1 (1-p)^{3-1} \quad \text{(from above)}$$

$$P(Y=2) = \binom{3}{2} p^2 (1-p)^{3-2}$$

$$P(Y=3) = \binom{3}{3} p^3 (1-p)^{3-3}$$

$$P(Y=k) = P(\emptyset) = 0 \qquad \forall\, k > 3.$$

The example illustrates that in general, if $Y \sim \text{binomial}(n, p)$, then Y has pmf

$$P(Y=k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k \in \{0, 1, 2, \ldots, n\}.$$

Furthermore, the cdf of Y is then

$$P(Y \le k) = P\left(\{Y=0\} \cup \{Y=1\} \cup \cdots \cup \{Y=k\}\right),$$

and since the union is of disjoint events,

$$\begin{aligned} P(Y \le k) &= P(Y=0) + P(Y=1) + \cdots + P(Y=k) \\ &= \binom{n}{0} p^0 (1-p)^{n-0} + \binom{n}{1} p^1 (1-p)^{n-1} + \cdots + \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j} . \end{aligned}$$

Note the role of the combination $\binom{n}{j}$ throughout our discussion of the binomial distribution. For this role, the combination is sometimes referred to as the binomial coefficient. Next, lets consider the mean and variance of a binomial random variable.

PROPOSITION. If $Y \sim \text{binomial}(n,p)$, then $E(Y) = np$ and $\text{Var}(Y) = np(1-p)$.

Proof. Recall that $Y \sim \text{binomial}(n,p)$ can be expressed as

$$Y = \sum_{i=1}^{n} X_i$$

for $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$, $E(X_i) = p$, and $\text{Var}(X_i) = p(1-p)$. Then

$$\begin{aligned} E(Y) &= E(X_1 + \cdots + X_n) \\ &= E(X_1) + \cdots + E(X_n) \\ &= np \end{aligned}$$

and

$$\begin{aligned} \text{Var}(Y) &= E(Y^2) - [E(Y)]^2 \\ &= E\left[\left(\sum_{i=1}^{n} X_i \right) \left(\sum_{j=1}^{n} X_j \right) \right] - (np)^2 \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} E(X_i X_j) - n^2 p^2 \\ &= \sum_{i=1}^{n} E(X_i^2) + \sum\sum_{i \ne j} E(X_i X_j) - n^2 p^2 \\ &= \sum_{i=1}^{n} \left(\text{Var}(X_i) + [E(X_i)]^2 \right) + \sum\sum_{i \ne j} E(X_i X_j) - n^2 p^2 \\ &= n\left(p(1-p) + p^2 \right) + \sum\sum_{i \ne j} E(X_i X_j) - n^2 p^2 \\ &= np - n^2 p^2 + \sum\sum_{i \ne j} E(X_i X_j) \end{aligned}$$

Next consider the pmf of the random variable $X_i X_j$ for $i \ne j$.

$x_i x_j$	0	1
$p(x_i x_j)$	$1-p+p(1-p)$	p^2

, $i \neq j$

Then

$$E(X_i X_j) = 0 \cdot (1-p+p(1-p)) + 1 \cdot p^2 = p^2,$$

and so

$$\sum_{i \neq j} \sum E(X_i X_j) = n(n-1) \cdot p^2,$$

which gives

$$\begin{aligned} \text{Var}(Y) &= np - n^2p^2 + n(n-1)p^2 \\ &= np - n^2p^2 + n^2p^2 - np^2 \\ &= np - np^2 \\ &= np(1-p). \end{aligned}$$

#

Read Section 3.4 in Devore.

SECTION 3.5. HYPERGEOMETRIC AND NEGATIVE BINOMIAL DISTRIBUTIONS.

Recall that the binomial distribution describes the number of successes, X, in n trials of some experiment. Suppose, instead, that I am interested in the number of trials, N, needed to observe r successes. If the probability of success is denoted by p, then we say,

$$Y := N - r \sim \text{negative-binomial}(r, p).$$

To construct the pmf of Y, consider the event that exactly y failures in n trials are needed to observe r successes. Then the r-th success has to be observed on the n-th trial, and there are

$$\binom{n-1}{r-1} = \binom{y+r-1}{r-1}$$

such outcomes that satisfy this condition, each occurring with probability $p^r(1-p)^y$. Hence, the pmf of Y is

$$p(y) = \binom{y+r-1}{r-1} p^r (1-p)^y, \quad \text{for } y \in \{0, 1, 2, \dots\}.$$

Note that we could have similarly defined the negative binomial distribution in terms of the number of trials, N.

The negative binomial experiment could be used to describe how many drugs we need to test in order to find r promising candidates. In the special case when $r=1$ (i.e., how many failures until the first success),

$$p(y) = \binom{y}{0} p^1 (1-p)^y = p(1-p)^y,$$

and we denote $Y \sim \text{geometric}(p)$.

PROPOSITION. If $Y \sim \text{negative-binomial}(r, p)$, then

$$E(Y) = \frac{r(1-p)}{p} \quad \text{and} \quad \text{Var}(Y) = \frac{r(1-p)}{p^2}.$$

The proof of this proposition is a bit tedious, but a simple argument follows for the case when $r=1$ (i.e., $Y \sim \text{geometric}(p)$), for $E(Y)$.

By definition,

$$\begin{aligned} E(Y) &= \sum_{y=0}^{\infty} y\, p(1-p)^y \\ &= (1-p)p \sum_{y=0}^{\infty} y(1-p)^{y-1} \\ &= (1-p)p \sum_{y=0}^{\infty} \frac{d}{dp}(1-p)^y \cdot (-1) \\ &= -(1-p)p \frac{d}{dp} \sum_{y=0}^{\infty} (1-p)^y \\ &= -(1-p)p \cdot \frac{d}{dp}\left(\frac{1}{1-(1-p)}\right) \end{aligned}$$

Since a geometric series has the expression $\sum_{x=0}^{\infty} a^x = \frac{1}{1-a}$ if $|a|<1$. Then

$$\begin{aligned} E(Y) &= -(1-p)p \frac{d}{dp}\left(\frac{1}{p}\right) \\ &= -(1-p)p(-1)p^{-2} \\ &= \frac{1-p}{p}. \end{aligned}$$

Note that the negative binomial distribution is also well-defined as a distribution of values of y being non-negative real numbers.

For completeness I will briefly introduce the hypergeometric distribution, but its usefulness in statistical inference is considerably less than the other distributions we have/will study.

Let X be the number of successes in a simple random sample of size n, drawn from a population consisting of M successes and $N-M$ failures. Then the random variable X is said to follow the hypergeometric distribution, having pmf,

$$p(x) = P(X=x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}},$$

for any integer x satisfying $\max\{0, n-N+M\} \le x \le \min\{n, M\}$.

EXAMPLE. Suppose that scientists are studying a certain species that is in danger of becoming extinct. Further, in a particular region of the world it is believed that only 25 animals from the endangered species are alive. To investigate this belief the scientists capture and tag 5 of the animals, and then release them back into the population. After some time, the scientists capture a random sample of 10 animals from the population. The idea is that it is increasingly unlikely to observe the same 5 tagged animals, the larger the population is. Letting X denote the number of tagged animals observed from the sample of 10, and assuming that there are in fact 25 animals in the population, there are

$$\binom{M}{x} = \binom{5}{x} \quad \text{ways to observe } x \text{ of the 5 tagged,}$$

$$\binom{N-M}{n-x} = \binom{25-5}{10-x} \quad \text{ways to observe } n-x \text{ of the untagged.}$$

This gives $\binom{5}{x}\cdot\binom{20}{10-x}$ outcomes that characterize the event. Moreover, there exist $\binom{N}{n} = \binom{25}{10}$ ways to observe a sample of 10. Hence,

$$p(x) = P(X=x) = \frac{\binom{5}{x}\cdot\binom{20}{10-x}}{\binom{25}{10}},$$

for $x \in \{0, 1, 2, \ldots, 5\}$.

The mean and variance of a hypergeometric random variable X are

$$E(X) = n\cdot\frac{M}{N} \quad \text{and} \quad Var(X) = \left(\frac{N-n}{N-1}\right)\cdot n\cdot\frac{M}{N}\cdot\left(1-\frac{M}{N}\right),$$

where $\frac{M}{N}$ is the proportion of successes in the population. For large N, $\frac{M}{N} \approx p$, the probability of success in a binomial experiment of n trials. Then $E(X) \approx n\cdot p$ and $Var(X) \approx \left(\frac{N-n}{N-1}\right)\cdot n\cdot p(1-p)$, both of which match that of the binomial distribution, up to the scaling factor of

$\left(\frac{N-n}{N-1}\right)$ for the variance. This scaling factor is referred to as the finite population correcting factor. Since $N-n < N-1$, it follows that a hypergeometric random variable has a smaller variance than a binomial random variable. This is an important consideration for using the binomial distribution as an approximation to the hypergeometric distribution.

Read Section 3.5 in Devore.

SECTION 3.6. THE POISSON PROBABILITY DISTRIBUTION.

In this section we move beyond probability distributions that can be derived from simple experiments consisting of draws from a population or tosses of a coin. Nonetheless, the Poisson distribution is still related in a mathematical sense to the binomial distribution.

DEFINITION. A discrete random variable X is said to have a Poisson distribution with parameter $\mu > 0$ if the pmf of X is

$$p(x) = \frac{\mu^x \cdot e^{-\mu}}{x!}, \text{ for } x \in \{0, 1, 2, \dots\}.$$

Now that we are transitioning to study distributions with mass functions having concise mathematical expressions it becomes less transparent whether the candidate pmf satisfies the defining properties of a pmf (non-negative and sums to one). So let's consider the Poisson pmf.

Non-negative: Since $\mu > 0$, $\mu^x > 0 \quad \forall x \in \{0, 1, 2, \dots\}$

Similarly, $e \approx 2.718 < 3$

So that $e^{\mu} < 3^{\mu}$ and $e^{-\mu} > \left(\frac{1}{3}\right)^{\mu} > 0$

Lastly, since a factorial is always non-negative,

$$p(x) = \frac{\mu^x \cdot e^{-\mu}}{x!} > 0 \quad \forall x \in \{0, 1, 2, \dots\}$$

Sums to one: Recall the Taylor series for the exponential function about the point zero,

$$e^y = 1 + y + \frac{y^2}{2!} + \frac{y^3}{3!} + \cdots = \sum_{x=0}^{\infty} \frac{y^x}{x!}$$

Evaluating at the point $y = \mu$ gives

$$e^{\mu} = \sum_{x=0}^{\infty} \frac{\mu^x}{x!} \quad \text{and so} \quad 1 = \sum_{x=0}^{\infty} \frac{e^{-\mu}\mu^x}{x!}.$$

The Poisson random variable arises in contexts for which we are describing the uncertainty for observing x occurrences of some rare event. For example, suppose that you are late for work in the morning on days for which there is a traffic accident, but on any given day there is only a .004 probability of an accident. If X is the number of days in a year that you are late for work, and you work $\approx 52 \cdot 5 = 260$ days per year then a reasonable model is

$$X \sim \text{Poisson}(\mu = .004 \cdot 260 = 1.04).$$

We can compute, for instance,

$$P(X=5) = \frac{(1.04)^5 e^{-1.04}}{5!} \approx .00358$$

Note that the binomial $(260, .004)$ distribution is not appropriate because 260 is only an approximation to the number of days you typically work in a year (e.g., you might be sick or on holiday some days). Still, the binomial distribution gives reasonable approximation to Poisson probabilities for large n and small p. In this case, if $Y \sim \text{binomial}(260, .004)$, then

$$P(Y=5) = \binom{260}{5} (.004)^5 (1-.004)^{260-5} \approx .00351$$

Alternatively, if you live in a city and the probability of a traffic accident is, say .2 on any given day, then

$$P(X=5) = \frac{(52)^5 e^{-52}}{5!} \approx 8.27 \cdot 10^{-17}$$

and

$$P(Y=5) = \binom{260}{5} (.2)^5 (1-.2)^{260-5} \approx 5.92 \cdot 10^{-19}.$$

More formally, we can precisely describe the relationship between the Poisson and binomial distributions as follows.

PROPOSITION. Given a value of $\mu > 0$, and a sequence $p_1, p_2, \ldots$ with $p_n \in [0,1]$ for every $n \in \{1, 2, \ldots\}$, if $n \cdot p_n \longrightarrow \mu$ as $n \longrightarrow \infty$, then

$$\binom{n}{x} p_n^x (1-p_n)^{n-x} \longrightarrow \frac{\mu^x e^{-\mu}}{x!}$$

as $n \to \infty$, for any $x \in \{0, 1, 2, \ldots\}$.

Proof. I will give a proof of the argument here. The argument mostly relies on the fact that for any sequence of numbers $\{a_n\}$ such that $a_n \longrightarrow a$ as $n \longrightarrow \infty$, for some fixed number a,

$$\lim_{n \to \infty} \left(1 + \frac{a_n}{n}\right)^n = e^a.$$

Observe the following relations for any fixed $x \in \{0, 1, 2, \ldots, n\}$. For large n, with $np_n \approx \mu$,

$$\binom{n}{x} p_n^x (1-p_n)^{n-x} = \frac{n!}{(n-x)!\,x!} \cdot \frac{(np_n)^x}{n^x} \cdot \left(1 - \frac{np_n}{n}\right)^{n\left(1-\frac{x}{n}\right)}$$

$$\approx \frac{n\cdot(n-1)\cdots(n-(x-1))(n-x)!}{n^x \cdot x! \cdot (n-x)!} \cdot \mu^x \left(1 + \frac{(-np_n)}{n}\right)^n$$

$$\approx \frac{n\cdot(n-1)\cdots(n-(x-1))}{n\cdot n \cdots n} \cdot \frac{\mu^x}{x!} \cdot e^{-\mu}$$

$$\approx 1\cdot\left(1-\frac{1}{n}\right)\cdots\left(1-\left(\frac{x-1}{n}\right)\right) \cdot \frac{\mu^x}{x!} \cdot e^{-\mu}$$

$$\approx 1\cdot 1\cdots 1 \frac{\mu^x}{x!} \cdot e^{-\mu}$$

#

Since a pmf completely characterizes the distribution of a random variable, this approximation should also suggest expressions for the mean and variance of a Poisson(μ) random variable X as

$$E(X) \approx \lim_{n\to\infty} n\cdot p_n = \mu$$

and

$$\mathrm{Var}(X) \approx \lim_{n\to\infty} np_n(1-p_n) = \mu.$$

PROPOSITION. If $X \sim \text{Poisson}(\mu)$, then $E(X) = \mu$ and $\mathrm{Var}(X) = \mu$.

Proof. By definition,

$$E(X) = \sum_{x=0}^{\infty} x \cdot \frac{\mu^x}{x!} \cdot e^{-\mu}$$

$$= \sum_{x=1}^{\infty} x \cdot \frac{\mu^x}{x!} \cdot e^{-\mu}$$

$$= \mu \cdot \sum_{x=1}^{\infty} \mu^{-1} \frac{\mu^x}{(x-1)!} \cdot e^{-\mu}$$

$$= \mu \cdot \sum_{x=1}^{\infty} \frac{\mu^{x-1}}{(x-1)!} \cdot e^{-\mu}$$

$$= \mu \cdot \underbrace{\sum_{k=0}^{\infty} \frac{\mu^k}{k!} \cdot e^{-\mu}}_{=1}$$

$$= \mu.$$

Similarly,

$$E(X^2) = \sum_{x=0}^{\infty} x^2 \cdot \frac{\mu^x}{x!} \cdot e^{-\mu} \qquad k = x-1, \quad x = k+1$$

$$= \mu \cdot \sum_{x=1}^{\infty} x \cdot \frac{\mu^{x-1}}{(x-1)!} \cdot e^{-\mu}$$

$$= \mu \cdot \sum_{k=0}^{\infty} (k+1) \frac{\mu^k}{k!} e^{-\mu}$$

$$= \mu \left[\underbrace{\sum_{k=0}^{\infty} k \frac{\mu^k}{k!} e^{-\mu}}_{= E(X) = \mu} + \underbrace{\sum_{k=0}^{\infty} \frac{\mu^k}{k!} e^{-\mu}}_{= 1} \right]$$

$$= \mu(\mu+1)$$

So that

$$\mathrm{Var}(X) = E(X^2) - E^2(X) = \mu(\mu+1) - \mu^2 = \mu. \qquad \#$$

Read Section 3.6 in Devore.

CHAPTER 4. CONTINUOUS RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

SECTION 4.1. PROBABILITY DENSITY FUNCTIONS

We move to this chapter to study random variables that take values on a set of real numbers that is neither finite nor countably infinite. Recall the distinction between discrete and continuous random variables, given at the beginning of Chapter 3,

DEFINITION. A random variable is called discrete if it takes values on a finite or countably infinite set of real numbers.
A random variable is called continuous if both of the following conditions are satisfied.

(i) The random variable takes values on sets consisting of one or more entire intervals of real numbers.

(ii) The probability of observing any particular value, say $c \in \mathbb{R}$, is zero. That is, if X is a continuous random variable, then

$$P(X = c) = 0 \qquad \forall c \in \mathbb{R}.$$

For example, the amount of time spent in the waiting room by patients in the emergency department at a hospital is an important continuous random variable.

Another important distinction between discrete and continuous random variables is the discretization of a continuous random variable. We do this naturally for various measurements. For instance, the age of a person is a continuous real valued number, but if you were to ask me for my age I would provide the discretized measurement, 30 years old.

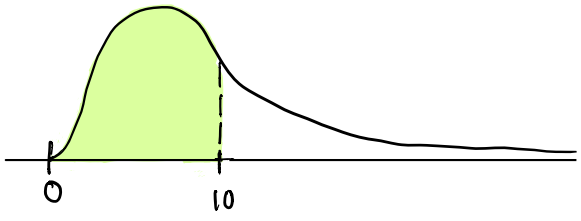
Let's consider some attributes for continuous random variables. Just as there was a pmf for discrete random variables, there is a probability density function for continuous random variables.

DEFINITION. Let X be a continuous random variable. Then a probability density function (pdf) of X is a function $f(\cdot)$ such that for any two numbers, a and b with $a \leq b$,

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

This definition means that probabilities of events relating to continuous random variables are given by the area under the curve (i.e., an integral) corresponding to some pdf. For example, if $f(\cdot)$ is the density of the emergency department waiting time, X (in minutes), then

$$P(0 \leq X \leq 10) = \int_0^{10} f(x)\,dx.$$



Similar to a pmf, a pdf has two defining properties:

(1) $f(x) \geq 0$ for every $x \in \mathbb{R}$

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

Note that for a continuous random variable, X, with pdf $f(\cdot)$,

$$P(X = 0) = \int_0^0 f(x)\,dx = 0$$

That is, the area under the curve of a continuous function is zero.

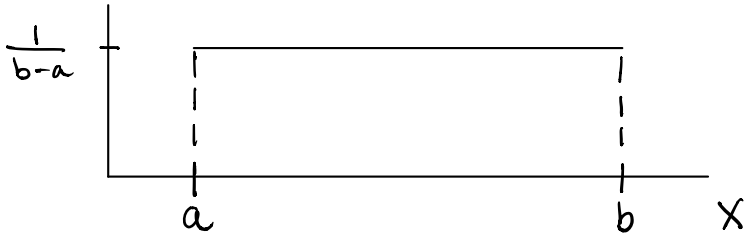
Perhaps the simplest and most fundamental continous random variable is the uniform random variable on an interval $[a, b]$.

DEFINITION. A continuous random variable is said to follow the uniform distribution on the interval $[a,b]$, with $a<b$, if X has pdf,

$$f(x) = \frac{1}{b-a} \cdot 1\{a \le x \le b\}.$$

In this case, we denote $X \sim \text{uniform}(a,b)$.

Graphically,



To verify that $f(x) = \frac{1}{b-a} \cdot 1\{a \le x \le b\}$ is a pdf, observe that

(1) $f(x) \ge 0 \quad \forall x \in \mathbb{R}$ since $b > a$

(2) $\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{\infty} \frac{1}{b-a} \cdot 1\{a \le x \le b\}\,dx = \int_a^b \frac{1}{b-a}\,dx = \frac{b-a}{b-a} = 1.$

The special case when $X \sim \text{uniform}(0,1)$ is referred to as a "standard" uniform random variable. By properties that we will learn in subsequent sections, if $X \sim \text{uniform}(0,1)$, then

$$Y = a + X(b-a) \sim \text{uniform}(a,b).$$

The uniform distribution characterizes experiments (mapped to an interval $[a,b]$) such that the likelihood of any interval $[c,d]$ is proportional to the length of the interval $[c,d] \cap [a,b]$, where the constant of proportionality is always $\frac{1}{b-a}$.

Basically, the uniform distribution is the continuous analogue of the discrete distribution of drawing marbles out of a bag of n marbles. In Chapters 2 and 3 we derived probabilities of events arising from such experiments, explicitly, by using counting techniques.

Moreover, if $X \sim \text{uniform}(0,1)$, then $Y := 1\{X \le p\} \sim \text{Bernoulli}(p)$.

Proof. By construction, $Y \in \{0,1\}$, so simply observe that

$$P(Y=1) = P(X \le p) = \int_0^p \frac{1}{1-0}\,dx = p,$$

and

$$P(Y=0) = 1 - P(Y=1) = 1-p.$$

#

Read Section 4.1 in Devore.

SECTION 4.2. CUMMULATIVE DISTRIBUTION FUNCTIONS AND EXPECTED VALUES.

Just like we introduced and studied cdfs for discrete random variables, we will do the same for continuous random variables. In fact, the cdf is defined the same as it was for discrete random variables.

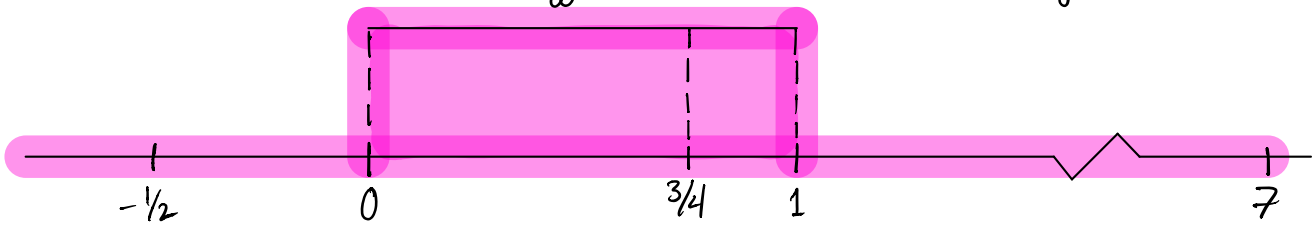
DEFINITION. The cdf, $F(\cdot)$, for a continuous random variable X is defined for every $x \in \mathbb{R}$ by

$$F(x) := P(X \leq x) = \int_{-\infty}^{x} f(y)\, dy,$$

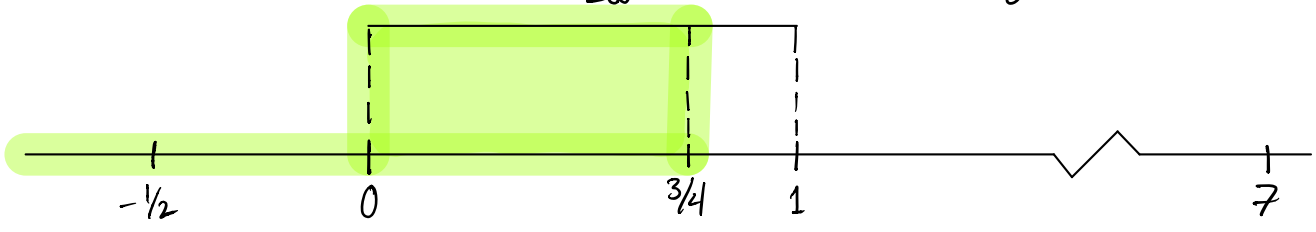
where $f(\cdot)$ is the pdf of X.

EXAMPLE. Suppose that $X \sim \text{uniform}(0,1)$, and let $F(\cdot)$ denote the cdf of X. Then

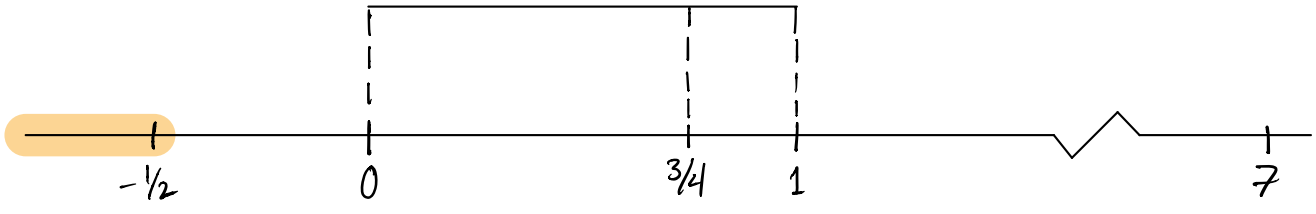
$$F(7) = P(X \leq 7) = \int_{-\infty}^{7} \frac{1}{1-0} 1\{0 \leq y \leq 1\}\, dy = \int_{0}^{1} dy = 1$$



$$F(3/4) = P(X \leq 3/4) = \int_{-\infty}^{3/4} 1\{0 \leq y \leq 1\}\, dy = \int_{0}^{3/4} dy = 3/4$$



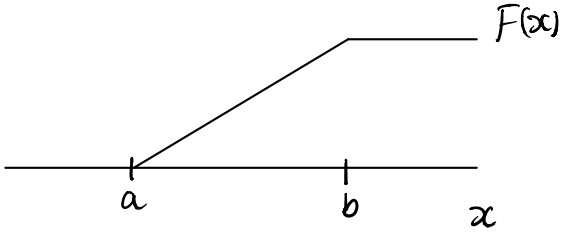
$$F(-1/2) = P(X \leq -1/2) = \int_{-\infty}^{-1/2} 1\{0 \leq y \leq 1\}\, dy = 0.$$



Further, observe that if $X \sim \text{uniform}(a,b)$, then for any $x \in \mathbb{R}$,

$$F(x) = \int_{-\infty}^{x} \frac{1}{b-a} 1\{a \le y \le b\}\, dy$$
$$= \left[\int_{a}^{\min\{x,b\}} \frac{1}{b-a}\, dy\right] \cdot 1\{x > a\}$$
$$= \left[\frac{\min\{x,b\} - a}{b-a}\right] \cdot 1\{x > a\}$$

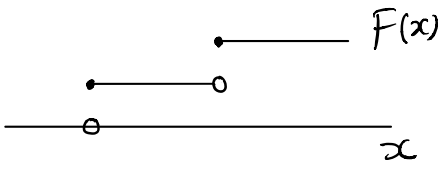
Equivalently,
$$F(x) = \begin{cases} 1 & \text{if } b \le x \\ \frac{x-a}{b-a} & \text{if } a < x < b \\ 0 & \text{if } x \le a \end{cases}.$$



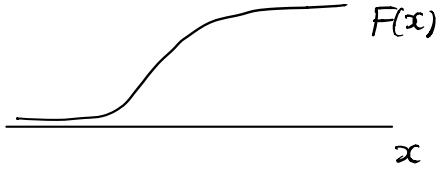
In general, whether a random variable is continuous or discrete, a cdf, $F(\cdot)$, has the following three defining properties.

(1) $F(\cdot)$ is non-decreasing. That is, if $x \le y$, then $F(x) \le F(y)$.

(2) $F(\cdot)$ is right-continuous. That is, $\forall x \in \mathbb{R}$, $\lim_{y \to x^-} F(y) = F(x)$.



discrete



Continuous

(3) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

PROPOSITION. Let X be a continuous random variable with pdf $f(\cdot)$ and cdf $F(\cdot)$. Then for any $a \in \mathbb{R}$,

$$P(X > a) = 1 - F(a),$$

and for any $b \in \mathbb{R}$ with $b > a$,

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

Proof. For the first statement,

$$P(X > a) = 1 - P(\{X > a\}^c)$$
$$= 1 - P(X \leq a)$$
$$= 1 - F(a),$$

by definition. For the second statement, note that

$$P(a \leq X \leq b) = P(\{X = a\} \cup \{a < X < b\} \cup \{X = b\})$$
$$= \underbrace{P(X = a)}_{=0} + P(a < X < b) + \underbrace{P(X = b)}_{=0}$$
$$= P(a < X < b).$$

The same argument shows that

$$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X < b) = P(a < X \leq b).$$

Then

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$
$$= \int_{-\infty}^b f(x)\,dx - \int_{-\infty}^a f(x)\,dx$$
$$= P(X \leq b) - P(X \leq a)$$
$$= F(b) - F(a).$$

#

Note that it similarly follows that

$$P(X \geq a) = 1 - F(a).$$

Recall that for a discrete random variable Y, with cdf $F_Y(\cdot)$, we can obtain the pmf of Y as

$$p(y) = P(Y = y) = P(Y \leq y) - P(Y \leq y-) = F_Y(y) - F_Y(y-),$$

where $y-$ is the largest value Y can take that is less than y. If Y is strictly integer-valued, then

$$p(y) = P(Y = y) = F_Y(y) - F_Y(y-1).$$

What is the analogue for constructing the pdf of a continuous random

variable X from its cdf?

PROPOSITION. If X is a continuous random variable with pdf $f(\cdot)$ and cdf $F(\cdot)$, then for every x at which the derivative $F'(x)$ exists,

$$f(x) = F'(x).$$

Proof. By definition,

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\,dt.$$

Then by the fundamental theorem of calculus, $F'(x) = f(x)$. #

EXAMPLE. Let $X \sim \text{uniform}(a, b)$. Then

$$\frac{d}{dx} F(x) = \frac{d}{dx} \begin{cases} 1 & \text{if } b \le x \\ \frac{x-a}{b-a} & \text{if } a < x < b \\ 0 & \text{if } x \le a \end{cases}$$

$$= \begin{cases} 0 & \text{if } b \le x \\ \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{if } x \le a \end{cases}$$

$$= f(x).$$

Next, recall the notion of percentiles from Chapter 1. This notion translates nicely for continuous random variable. For example, suppose that X represents the height of an individual in a population. Assuming that we do not discretize our measurements of height, X is a continuous random variable with some cdf $F(\cdot)$. Then an individual is in the 70th percentile of population height if the individual's height x satisfies $F(x) = .7$. That is,

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\,dt = .7.$$

Similarly, you would say that your height x is in the top 30th percentile if

$$.3 = \int_{x}^{\infty} f(t)\,dt = 1 - \int_{-\infty}^{x} f(t)\,dt = 1 - F(x).$$

Note, the function $1 - F(\cdot)$ is often referred to as the survival function of X. Now we generalize this percentile notion for any percentile.

DEFINITION. For $p \in [0,1]$, the $(100 \cdot p)$-th percentile of the distribution of a continuous random variable X, denoted by $\eta(p)$, is defined by

$$p = F(\eta(p)) = \int_{-\infty}^{\eta(p)} f(x)\, dx.$$

Recall from chapter 1, our notion of a sample median is a value such that half of the sample is larger and half of the sample is smaller, in value. Accordingly, our notion of a population median if a value such that half of the population values are larger and half are smaller. That is, $\tilde{\mu}$ is the population median of a random variable X, with cdf $F(\cdot)$, if

$$F(\tilde{\mu}) = P(X < \tilde{\mu}) = P(X > \tilde{\mu}) = 1 - F(\tilde{\mu})$$

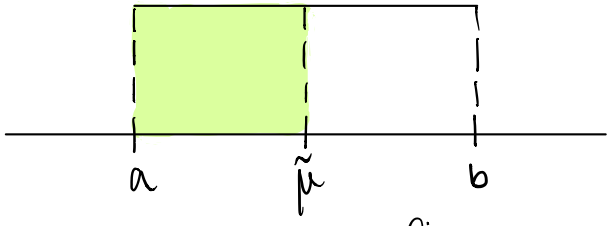
which gives,

$$2F(\tilde{\mu}) = 1,$$

and so

$$F(\tilde{\mu}) = .5 \quad \left(\text{i.e.,} \quad \int_{-\infty}^{\tilde{\mu}} f(x)\, dx = .5 \right)$$

For example, if $X \sim \text{uniform}(a, b)$, then



From the picture it follows that the median is the average of a and b,

$$\tilde{\mu} = \frac{a+b}{2}$$

To verify this,

$$F\left(\frac{a+b}{2}\right) = \int_{-\infty}^{\frac{a+b}{2}} \frac{1}{b-a} \cdot 1\{a \le x \le b\}\, dx$$

$$= \int_{a}^{\frac{a+b}{2}} \frac{1}{b-a}\, dx$$

$$= \frac{1}{b-a} \cdot \left(\frac{a+b}{2} - a\right)$$

$$= \frac{1}{2}.$$

To continue our development of continuous random variables, we need to define the expected value and variance in the continuous setting. Recall that for a discrete random variable Y with values in some finite or countably infinite set D,

$$E(Y) = \sum_{y \in D} y \cdot p(y) \quad \text{and} \quad \text{Var}(Y) = \sum_{y \in D} (y - E(Y))^2 \cdot p(y).$$

In the continuous setting, the set D is uncountably infinite, and so the sums become integrals. Recall from calculus that an integral is the limit of Riemann sums, as the partition element widths goes to zero.

DEFINITION. The expected or mean value of a continuous random variable X with pdf $f(\cdot)$ is

$$\mu := E(X) := \int_{-\infty}^{\infty} x f(x) \, dx.$$

EXAMPLE. If $X \sim \text{uniform}(a, b)$, then

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x \frac{1}{b-a} \cdot 1\{a \le x \le b\} \, dx \\ &= \frac{1}{b-a} \int_a^b x \, dx \\ &= \frac{1}{b-a} \cdot \left(\frac{x^2}{2} \Big|_a^b \right) \\ &= \frac{b^2 - a^2}{2(b-a)} \\ &= \frac{b+a}{2}. \end{aligned}$$

PROPOSITION. If X is a continuous random variable with pdf $f(\cdot)$ and $h(X)$ is any function of X, then

$$E[h(X)] = \int_{-\infty}^{\infty} h(x) \cdot f(x) \, dx.$$

Proof. a bit beyond the scope of ST 371. #

Using this result, we can again (as in the discrete case) understand the variance as an expected value.

DEFINITION. The variance of a continuous random variable X with pdf $f(\cdot)$ and mean μ is

$$\sigma^2 := \mathrm{Var}(X) := E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\,dx.$$

The standard deviation of X is $\sigma := \sqrt{\mathrm{Var}(X)}$.

EXAMPLE. If $X \sim \mathrm{uniform}(a,b)$, then

$$\begin{aligned}
\mathrm{Var}(X) &= \int_{-\infty}^{\infty} \left(x - \tfrac{a+b}{2}\right)^2 \frac{1}{b-a} \cdot 1\{a \le x \le b\}\,dx \\
&= \int_a^b \left[x^2 - x(a+b) + \tfrac{1}{4}(a+b)^2\right] \cdot \frac{1}{b-a}\,dx \\
&= \left[\frac{x^3}{3} - \frac{x^2}{2}(a+b) + \frac{x}{4}(a+b)^2\right]\Big|_a^b \cdot \frac{1}{b-a} \\
&= \left[4(b^3-a^3) - 6(b^2-a^2)(a+b) + 3(b-a)(a+b)^2\right] \cdot \frac{1}{12(b-a)} \\
&= \left[4b^3 - 4a^3 - 6b^2a + 6a^3 - 6b^3 + 6a^2b + 3(b-a)(a+b)^2\right] \cdot \frac{1}{12(b-a)} \\
&= \left[2a^3 - 2b^3 - 6b^2a + 6a^2b + 3ba^2 - 3a^3 + 6ab^2 - 6a^2b + 3b^3 - 3ab^2\right] \cdot \frac{1}{12(b-a)} \\
&= \left[b^3 - a^3 + 3a^2b - 3ab^2\right] \cdot \frac{1}{12(b-a)} \\
&= \left[(b^2+a^2)(b-a) + 2a^2b - 2ab^2\right] \cdot \frac{1}{12(b-a)} \\
&= \left[(b^2+a^2)(b-a) - 2ab(b-a)\right] \cdot \frac{1}{12(b-a)} \\
&= \left[b^2 - 2ab + a^2\right] \frac{1}{12} \\
&= \frac{(b-a)^2}{12}
\end{aligned}$$

PROPOSITION. Let X be a continuous random variable. Then

$$\mathrm{Var}(X) = E(X^2) - E^2(X).$$

Proof. By definition,

$$\begin{aligned}
\mathrm{Var}(X) &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx \\
&= \int_{-\infty}^{\infty} (x^2 - 2x\mu + \mu^2) f(x)\,dx
\end{aligned}$$

$$= \underbrace{\int_{-\infty}^{\infty} x^2 f(x)\,dx}_{=E(X^2)} - 2\mu \underbrace{\int_{-\infty}^{\infty} x f(x)\,dx}_{=E(X)} + \mu^2 \underbrace{\int_{-\infty}^{\infty} f(x)\,dx}_{=1}$$

$$= E(X^2) - 2\mu^2 + \mu^2$$

$$= E(X^2) - \mu^2. \qquad \#$$

Read Section 4.2 in Devore.

SECTION 4.3. THE NORMAL DISTRIBUTION.

The normal or Gaussian distribution is perhaps the most widely applied of all distributions. It does a nice job at describing measurement errors, and a very large class of random variables have the property that sums and averages of iid copies are approximately Gaussian. This fact is supported by mathematical theorems called central limit theorems, to be described in the next chapter.

DEFINITION. A continuous random variable X is said to follow a normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ if the pdf of X is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \text{for } -\infty < x < \infty$$

We denote $X \sim N(\mu, \sigma^2)$.

It is a medium-hard calculus problem to show that

$$\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sqrt{2\pi\sigma^2}$$

So that

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1.$$

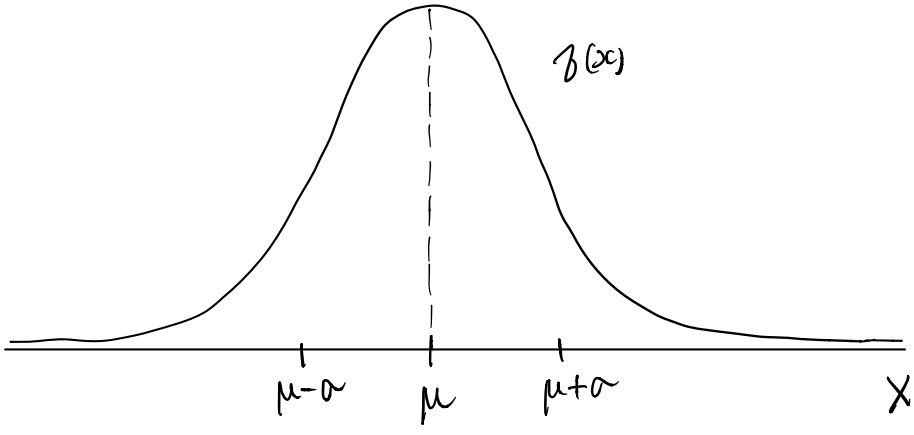
Conveniently, it can also be shown that if $X \sim N(\mu, \sigma^2)$, then

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu,$$

and

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sigma^2.$$

Graphically, the Gaussian density has the form,



(1) It is always unimodal and symmetric about the point $x = \mu$.

(2) The mean μ is also the median (i.e., $P(X \leq \mu) = .5$).

(3) μ is often referred to as the location, and σ the scale.

The special case with $\mu = 0$ and $\sigma = 1$ is called standard normal distribution. That is, $Z \sim N(0,1)$ if the density of Z is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad \text{for} \quad -\infty < z < \infty.$$

Unfortunately, there does not exist a nice closed-form expression for the cdf of Z, but it is referred to often enough that it is commonly denoted

$$\Phi(z) := P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy.$$

For example, Table A.3 in the appendix gives the value of $\Phi(z)$ for z in the range -3.49 to 3.49, by increments of .01. Alternatively, most computer software such as Excel, R, Python, Julia, etc. can be used to compute $\Phi(z)$ as well as probabilities of intervals for any $N(\mu, \sigma^2)$ random variable.

The following proposition allows us to relate any $N(\mu, \sigma^2)$ random variable to the standard normal distribution.

PROPOSITION. If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$.

Proof. $\Phi(z) = P(Z \leq z) = P\left(\frac{X-\mu}{\sigma} \leq z\right) = P(X \leq \sigma z + \mu) = F(\sigma z + \mu)$

where $F(\cdot)$ is the cdf of X. Taking derivatives of both sides gives the density of Z,

$$\begin{aligned}\phi(z) &:= \Phi'(z) \\ &= F'(\sigma z + \mu) \\ &= \sigma f(\sigma z + \mu) \\ &= \sigma \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(\sigma z + \mu - \mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}} .\end{aligned}$$

#

It can similarly be shown that the converse to this statement is also true. This result is useful because it allows us to express

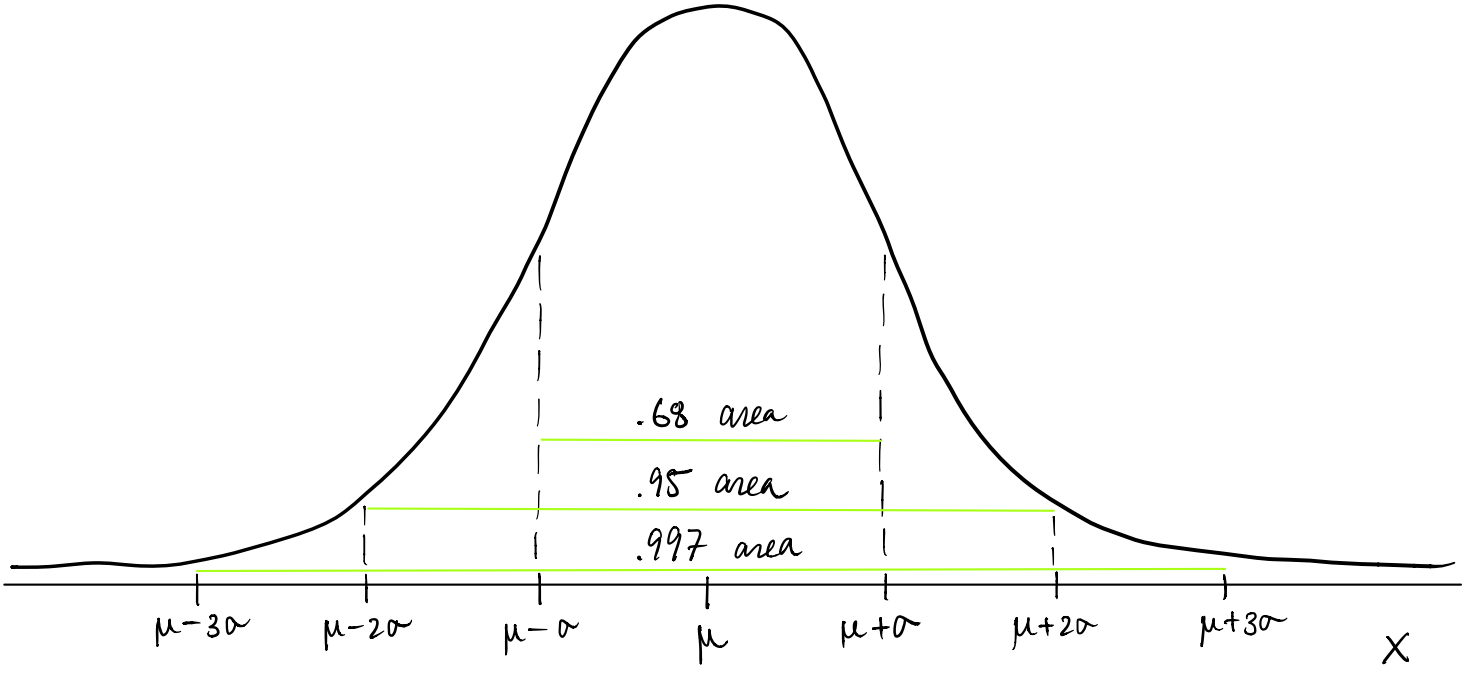
$$\begin{aligned}P(a \le X \le b) &= P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).\end{aligned}$$

Note that for any random variable, Y,

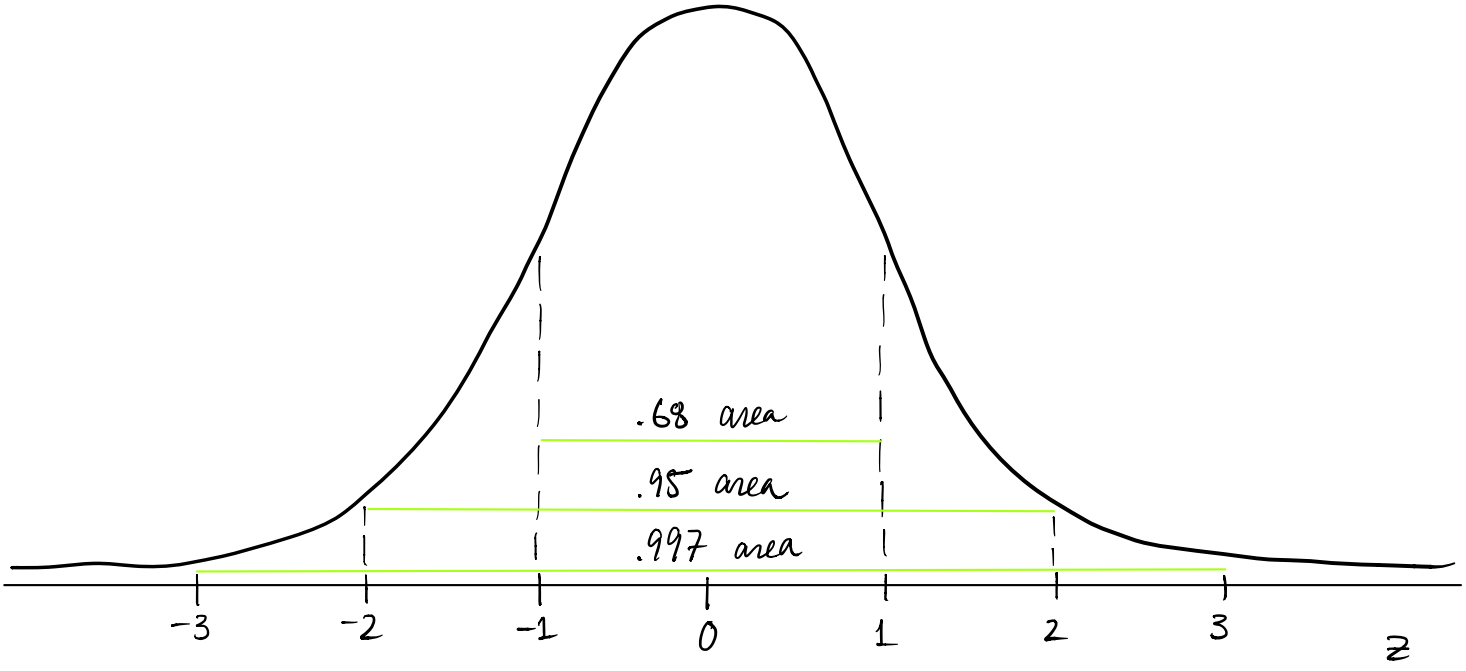
$$\frac{Y - E(Y)}{\sqrt{\mathrm{Var}(Y)}}$$

is called a standardized random variable. The idea of standardization will come up later when we discuss approximations using the normal distribution.

The following picture gives an important set of general rules for Gaussian probabilities



And, in particular, for the standard normal distribution,



Next we will consider a result that demonstrates that the Gaussian distribution also serves as a reasonable approximation for certain discrete random variables.

PROPOSITION. Let $X \sim \text{binomial}(n,p)$. Then for fixed p and x not too close to 0 or n,

$$\binom{n}{x} p^x (1-p)^{n-x} \longrightarrow \frac{1}{\sqrt{2\pi\, np(1-p)}} \cdot e^{-\frac{(x-np)^2}{2np(1-p)}}$$

as $n \longrightarrow \infty$.

Proof. The proof is beyond the scope of ST371, but the basic elements come from the fact that we can express

$$X = \sum_{i=1}^{n} Y_i$$

where $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Bernoulli}(p)$. Then apply the central limit theorem, which we will cover in the next chapter.

#

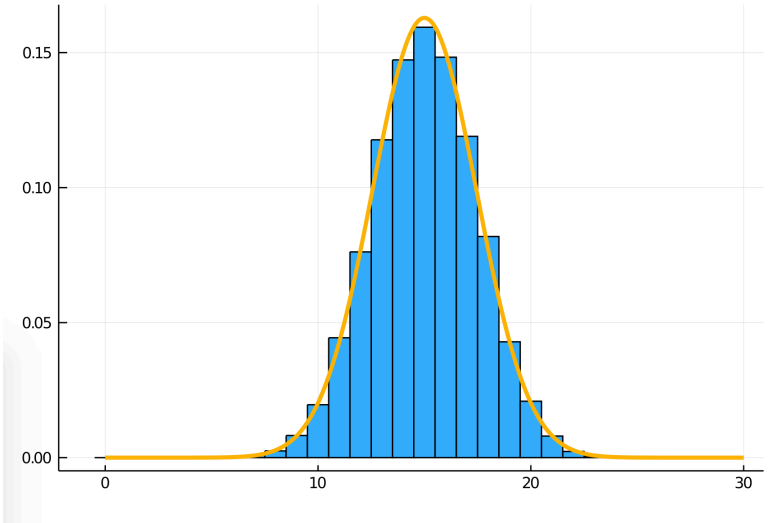
For instance, if $X \sim \text{binomial}(n,p)$. Then for x not too close to 0 or n,

$$P(X \le x) = P\left(\frac{X-np}{\sqrt{np(1-p)}} \le \frac{x-np}{\sqrt{np(1-p)}}\right)$$

$$\approx \Phi\left(\frac{x-np}{\sqrt{np(1-p)}}\right).$$

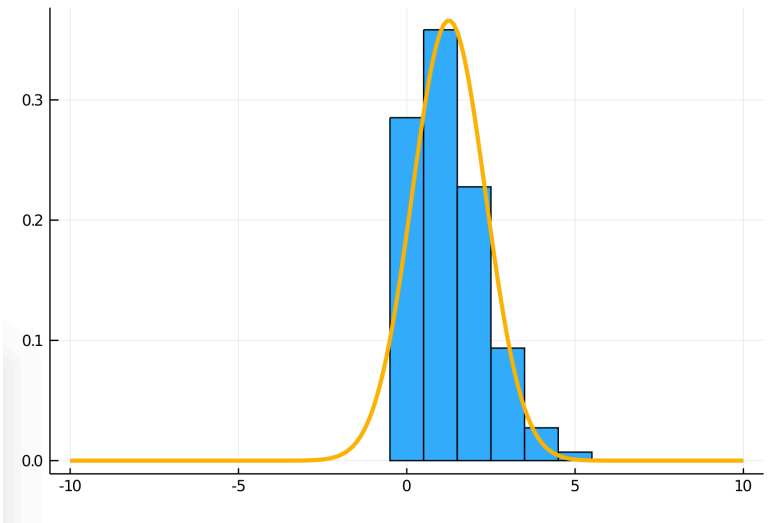
← standardization

The reason for the assumption that x is not too close to 0 or n is because the tails of the binomial and Gaussian distributions are intrinsically different; the support of the binomial pmf is between 0 and n, whereas the Gaussian pdf has support on the entire real line.



histogram of $X \sim \text{binomial}(25, .6)$

density of $Y \sim N(25 \cdot .6, 25 \cdot .6 \cdot .4)$



histogram of $X \sim \text{binomial}(25, .05)$

density of $Y \sim N(25 \cdot .05, 25 \cdot .05 \cdot .95)$

The textbook gives the rule that the normal approximation to the binomial is "adequate" if both $n \cdot p \geq 10$ and $n(1-p) \geq 10$. The two figures above illustrate why. However, with a computer, binomial probabilities are just as easy to compute as Gaussian probabilities, so just compute the exact probability. In modern day data analytics, such approximations are important for more advanced inference/models.

Read Section 4.3 in Devore.

SECTION 4.4. THE EXPONENTIAL AND GAMMA DISTRIBUTIONS.

The exponential and gamma distributions are widely used to describe the time to or between events.

DEFINITION. The continuous random variable X is said to follow the exponential distribution with rate parameter $\lambda > 0$ if the pdf of X is

$$f(x) = \lambda e^{-\lambda x} \cdot 1\{x \geq 0\}.$$

Denote $X \sim \text{exponential}(\lambda)$.

The exponential distribution is simple enough that we can work out the first two moments and the cdf, by hand.

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$$

$$= \int_{-\infty}^{\infty} x\, \lambda e^{-\lambda x} \cdot 1\{x \geq 0\}\, dx$$

$$= \int_{0}^{\infty} x\, \lambda e^{-\lambda x}\, dx$$

Integration by parts: $u = x\lambda \qquad v = -\frac{1}{\lambda} e^{-\lambda x}$

$$du = \lambda\, dx \qquad dv = e^{-\lambda x}\, dx$$

$$E(X) = u \cdot v \Big|_0^{\infty} - \int_0^{\infty} v\, du$$

$$= x\lambda\left(-\frac{1}{\lambda} e^{-\lambda x}\right)\Big|_0^{\infty} - \int_0^{\infty} -\frac{1}{\lambda} e^{-\lambda x}\, \lambda\, dx$$

$$= -\left(\frac{\infty}{e^{\infty}} - \frac{0}{1}\right) + \; -\frac{1}{\lambda} e^{-\lambda x}\Big|_0^{\infty}$$

$$= 0 - \frac{1}{\lambda}(0 - 1)$$

$$E(X) = \frac{1}{\lambda}$$

Next work out the second moment.

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \lambda e^{-\lambda x} \cdot 1\{x \geq 0\}\, dx$$

$$= \int_0^{\infty} x^2 \lambda e^{-\lambda x}\, dx$$

Integration by parts: $u = x^2\lambda \qquad v = -\frac{1}{\lambda} e^{-\lambda x}$

$du = 2\lambda x\, dx \qquad dv = e^{-\lambda x}\, dx$

$$E(X^2) = u \cdot v \Big|_0^{\infty} - \int_0^{\infty} v\, du$$

$$= x^2\lambda\left(-\frac{1}{\lambda} e^{-\lambda x}\right)\Big|_0^{\infty} - \int_0^{\infty} -\frac{1}{\lambda} e^{-\lambda x}\, 2\lambda x\, dx$$

$$= -\left(\frac{\infty^2}{e^{\infty}} - \frac{0}{1}\right) + \frac{2}{\lambda} \underbrace{\int_0^{\infty} x \cdot \lambda e^{-\lambda x}\, dx}_{= E(X)}$$

$$= 0 + \frac{2}{\lambda} \cdot \frac{1}{\lambda}$$

$$= \frac{2}{\lambda^2}$$

Thus,

$$\mathrm{Var}(X) = E(X^2) - E^2(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

The cdf of $X \sim$ exponential(λ) is

$$F(x) = P(X \leq x)$$

$$= \int_{-\infty}^{x} \lambda e^{-\lambda t} \cdot 1\{t \geq 0\}\, dt$$

$$= \left[\int_0^x \lambda e^{-\lambda t}\, dt\right] \cdot 1\{x \geq 0\}$$

$$= \left[-e^{-\lambda t}\Big|_0^x\right] \cdot 1\{x \geq 0\}$$

$$= \left[-(e^{-\lambda x} - 1)\right] \cdot 1\{x \geq 0\}$$

$$F(x) = (1 - e^{-\lambda x}) \cdot 1\{x \geq 0\}.$$

The random variable $X \sim$ exponential(λ) could represent the amount of time in minutes that you have to wait on hold when you call your internet

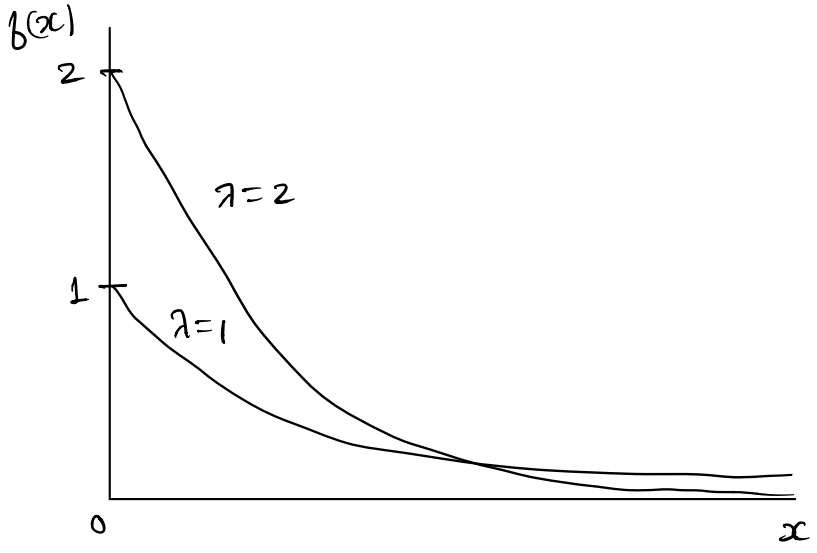
provider to tell them that you are switching to another provider now that your promotional period has ended! If people typically are on hold, on average, for 30 minutes, then

$$X \sim \text{exponential}\left(\lambda = \frac{1}{30}\right)$$

The probability that your wait-time exceeds 30 minutes is then

$$\begin{aligned} P(X > 30) &= 1 - F(30) \\ &= 1 - \left(1 - e^{-\frac{1}{30} \cdot 30}\right) \\ &= e^{-1} \\ &\approx .368 \end{aligned}$$

Observe that $P(X > E(X)) \neq .5$, like was the case with the $N(\mu, \sigma^2)$ distribution. In particular, the exponential distribution is not symmetric.



To continue with the call center waiting times, the number of callers, Y_t, in a given length of time, t, is also a random variable with mean, say $\alpha \cdot t$, for some fixed $\alpha > 0$ (the number of callers depends on the elapsed time). A reasonable model is then Poisson(αt), and the exponential distribution describes the inter-arrival times of the callers.

PROPOSITION. Suppose that $Y_t \sim$ Poisson(αt), and that the number of (Poisson) events occurring in nonoverlapping intervals of time are independent. Then the elapsed time between two successive events follows the exponential(α) distribution.

Proof. The proof is beyond the scope of ST 371. #

Although the proof is too complicated for this course, it is not difficult to verify for the time to first event, say X_1. That is, for $t>0$,

$$\begin{aligned} P(X_1 < t) &= 1 - P(X_1 \geq t) \\ &= 1 - P(\text{"no events before time } t\text{"}) \\ &= 1 - P(Y_t = 0) \\ &= 1 - \frac{(\alpha t)^0 e^{-\alpha t}}{0!} \\ &= 1 - e^{-\alpha t} \end{aligned}$$

which is the cdf of the exponential(α) distribution.

Another important property of the exponential distribution is often called a memoryless property. Suppose that X is the time to some event, and that it is characterized by $X \sim$ exponential(λ). Next, assume that the event has not occurred by some time t_0. Then what is the probability that it does not occur by time $t+t_0$, for any $t>0$? Does the knowledge that $X > t_0$ alter the probability that the event will occur after any amount of time into the future?

$$\begin{aligned} P(X > t+t_0 \mid X > t_0) &= \frac{P(\{X > t+t_0\} \cap \{X > t_0\})}{P(X > t_0)} \\ &= \frac{P(X > t+t_0)}{P(X > t_0)} \\ &= \frac{1 - F(t+t_0)}{1 - F(t_0)} \\ &= \frac{1 - \left(1 - e^{-\lambda(t+t_0)}\right)}{1 - \left(1 - e^{-\lambda t_0}\right)} \\ &= e^{-\lambda t} \\ &= 1 - F(t). \end{aligned}$$

That is, if the event is not observed before time, t_0, then the chance that it will not be observed for an addition amount of time, t, only depends on t. Note that not all time to event processes are consistent with this memoryless property. For example, the time until I die, in an epidemiological sense is characterized probabilistically by various age milestones that I survive. By surviving childhood, my life expectancy likely changed.

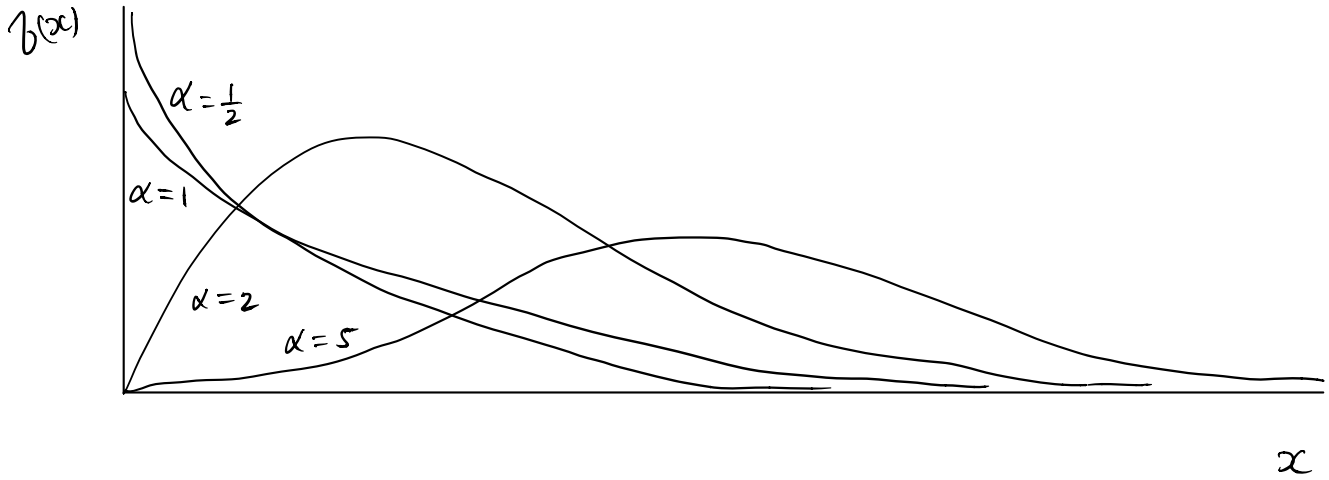
The next distribution we will look at is an extention of the exponential distribution.

DEFINITION. A continuous random variable X is said to follow the gamma distribution with parameters $\alpha > 0$ and $\beta > 0$ if the pdf of X is

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} \cdot 1\{x \geq 0\}.$$

Denote $X \sim \text{gamma}(\alpha, \beta)$.

The parameter α is called the shape parameter.



The parameter β is called the scale parameter, and plays a similar role as the rate parameter λ in the exponential distribution.

The special case, $\alpha = 1$ and $\beta = \frac{1}{\lambda}$ corresponds to the exponential distribution:

$$\text{exponential}(\lambda) = \text{gamma}(1, \tfrac{1}{\lambda}).$$

The function $\Gamma(\alpha)$ in the normalizing constant is defined for any $\alpha > 0$ as

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt.$$

It is called the gamma function (not to be confused with the gamma distribution), and arises in various contexts within the mathematical sciences (see its Wikipedia page).

The gamma function has the following nice properties:

(1) $\forall\, \alpha > 1$, $\Gamma(\alpha) = (\alpha-1)\,\Gamma(\alpha-1)$ (prove with integration by parts).

(2) For any $n \in \{1, 2, 3, \ldots\}$, $\Gamma(n) = (n-1)!$.

(3) $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Next, let's see how the gamma function arises in the density of the gamma

distribution by verifying that the density integrates to 1.