Chapter 3.1 - Methodologies of Probability

> “The probability is like the stick used by the blind man to feel his way. If he could see, he would not need the cane, just as if we knew which horse runs faster, then we would not need probability theory.” - Stanislaw Lem

Probability is a trivial concept that statisticians have turned into a grotesque beast of unfathomable complexity. This is primarily for two reasons:

a. The average person is incapable of conceptualizing probability appropriately.

b. Statisticians needed job security.

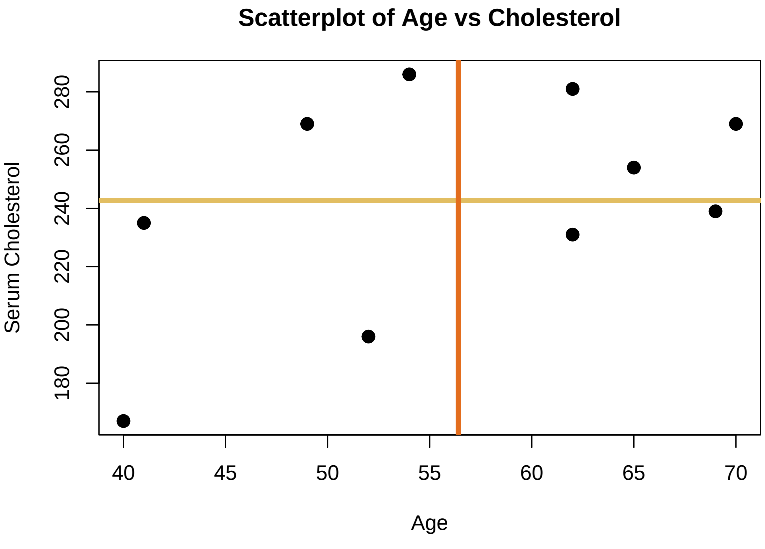
If something doesn’t happen, it clearly had no chance of occurring. If something happens, it’s already occurred so the chance was 100%. Obviously we can reduce this down to a binary outcome: Something happens or it doesn’t. Meaning that all probabilities are some form of coin flip, and coin flips have 50/50 odds. All occurrences are 50/50 odds coin flips. □

In case it wasn’t apparent, that “proof” shouldn’t be taken seriously at all.

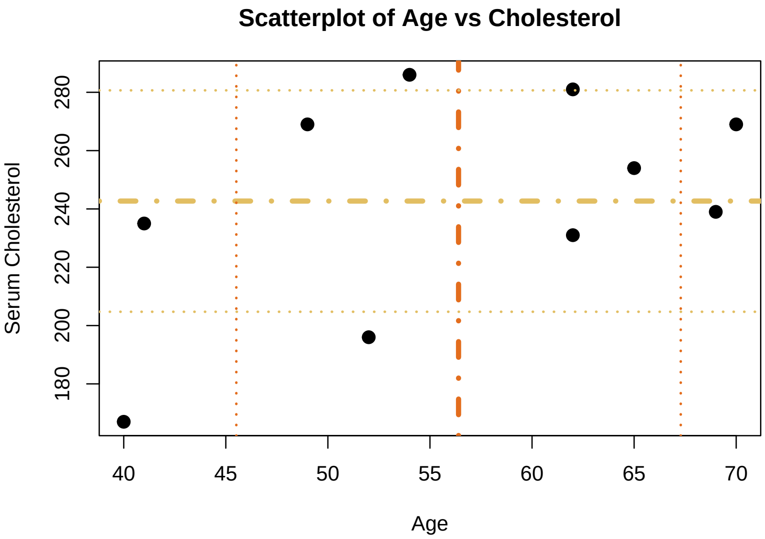
Why?

Why do statisticians study probability? So far all we’ve seen is graphical tools, summary statistics, and data collection. But these are just the basic vocabulary of statistics, we now dive head first into the true *science* of statistics.

Say you wanted to describe the cholesterol study data with summary statistics, like the mean:



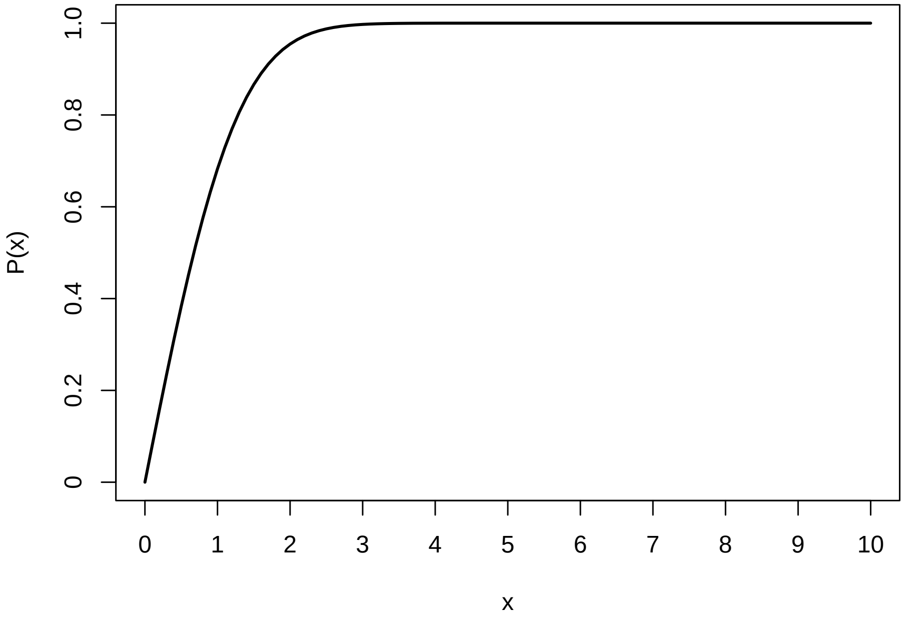
These lines are accurate to the data but inaccurate to anything **outside** of this study. Statisticians want our results to be *generalizable* because this lets us spend less time analyzing experiments and more time pretending to work. The basis behind studying probability comes from our addition of standard deviation lines into this scatterplot:



Probability allows us to answer the question, "How likely is it for a random person to fall within 0.8 standard deviations of the average cholesterol from this study?".

Basic Concepts of Probability

Probability: A number between 0 and 1 that tells us how *likely* a given "event" is to occur



We denote the probability of some event, x, occuring as $P(x)$. From here we can assign a value to $P(x)$ between 0 and 1 to describe the likelihood of event occurrence.

$P(x) = 0$	The event cannot occur
$P(x) = 0.5$	The event is as likely to occur as it is to not occur
$P(x) = 1$	The event must occur

As $P(x)$ gets *very close* to 0, it's possible for the event to occurr, but so unlikely that we'd be surprised if it did. For instance, it's *highly unlikely* that a shark attack occurrs on any given beach in the U.S., but it *is* possible:

$$x = \{\text{A shark attack on a beach in the U.S.}\}$$

$$P(x) \approx 0.00000008$$

As $P(x)$ approachs 1, it's not *impossible* for the event to occur although we're very confident it should. Like losing the lottery:

$$x = \{\text{You lose the national lottery you bought a ticket for}\}$$

$$P(x) \approx 0.999999997$$

Gambling isn't an advisable activity.

Probability Terminology

In order to formally study probability we need address some changes in vocabulary. Imagine you flip a *fair*, two-sided coin twice. This act of flipping the coin twice is considered an **experiment**. While not completely unlike the experiments we covered in Chapter 2 there's still an important distinction to be made:

Experiment (in context of probability): An activity that results in a definite outcome where the observed outcome is determined by chance.

What are the possible outcomes for this experiment? If consider the coin landing on it's side to be an impossible outcome, there are 4 possible results from this experiment:

Flip 1	Flip 2
Heads	Heads
Heads	Tails
Tails	Heads
Tails	Tails

This list of possible outcomes is referred to as the **sample space**.

Sample space: The set of **ALL** possible outcomes of an experiment; denoted by S.

We can represent the sample space with tables, pictures, or notation. The sample space for the coin flip experiment would be:

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

If we want to observe a small piece of our sample space, such as all of the outcomes that include tails, we would refer to this as an **event**.

Event: A subset of outcomes belonging to sample space S.

Events are typically denoted by a capital letter towards the beginning of the alphabet:

- i.e. A, B, C, etc.

The possible outcomes in an event, A, where tails shows up in one of the two flips would be:

$$A = \{\text{TH, HT, TT}\}$$

Whereas the possible outcomes in an event, B, where *only* tails shows up would be:

$$B = \{\text{TT}\}$$

Since there's only one outcome from S in the event, B, this is considered a **simple event**.

Simple event: An event containing a single outcome in the sample space S

Meanwhile the event, A is considered a **compound event**.

Compound event: An event formed by combining two or more events (thereby containing two or more outcomes in the sample space S).

This can be thought of as any event that has more than one outcome since every outcome in an event can be deconstructed into it's own simple event.

$$A = \{\text{At least one flip is tails}\} = \{\text{TH, HT, TT}\}$$

$$B = \{\text{Both flips are tails}\} = \{\text{TT}\}$$

$$C = \{\text{The first flip is tails and at least one flip is heads}\} = \{\text{TH}\}$$

Probability Methods

When discussing probability it's good to establish how we're *assigning* probabilities to events. There are numerous methods for doing this but the majority of studies, papers, and analyses use one of the three we'll discuss below.

Subjective Probability: Probability is assigned based on judgement or experience.

- We refer to experts and ask them their opinion or observation regarding the probability of an event.
  - A doctor assessing the chance of a patient recovering from a medical procedure
  - A managerial team estimating the probability a project will achieve technical success

This probability may not be expressed in an actual number; instead, we may say "low", "high", "almost certain", etc.

Classical Probability: makes **assumptions** in order to build **mathematical models** from which probabilities can be **derived**.

Suppose we want to put a probability on the event of observing "tails" in one flip of a coin. We might assume the following:

- 2 possible outcomes: "heads" or "tails"
- The coin is "fair" (i.e., heads and tails have an equal chance of occurring)

Based on these assumptions we can develop the following model:

$$\text{The probability of observing tails} = P(\text{tails}) = \frac{1}{2}$$

This is the simplest probability model, equiprobable outcomes. This model can be generalized so that any experiment meeting the assumption of equal likelihood across outcomes can be modeled with it. Given any event, A:

$$P(A) = \frac{\text{number of outcomes in event } A}{\text{total number of outcomes in } S}$$

Relative or Empirical Probability: The probability of an event is the proportion of times that the event occurs.

This is a common method for assigning probability since we can apply it without knowing the "true" probability of an event. In this case if we flip a coin repeatedly to observe the probability of it landing on heads, the probability would be recorded as:

$$P(\text{Heads}) = \{\text{the proportion of all possible flips where the coin lands on "heads"}\}$$

- We could flip this coin many times (say 1000) and count the number of times it lands point up.
- This is like a simple random sample (SRS) from the population of all coin flips.

$$P(\text{heads}) \approx \frac{\text{number of times "heads" is observed}}{1000}$$

Law of Large Numbers

> There will be many instances in this book where I'll be using something called "R". R is a widely used statistical scripting language that offers a variety of tools to make statistical experiments and analyses possible. While this isn't a book on R programming there are certain examples that are made easier with its use. You (the reader) aren't required to code along, but you're more than welcome to.

Let's flip a coin. In order to do this we'll use a function that creates a 0.5 probability event with **two** possible outcomes; 0 and 1. We can define "heads" as 0 and "tails" as 1 inside the program for now, to keep things simple.

```r
set.seed(73) # seed for reproducibility
result=ifelse(rbinom(1,1,0.5) == 0, "Heads", "Tails")  # flip a coin once
cat(result)
```

```
## Heads
```

While we already know that the probability of heads is 0.5, if we were observing this from the empirical probability viewpoint what would the probability of heads be?

$$P(\text{heads}) \approx \frac{\text{number of times "heads" is observed}}{\text{number of times the coin if flipped}} = \frac{1}{1} = 1$$

Now if we flip this coin 5 times, would this probability change?

```r
set.seed(73)
result=ifelse(rbinom(5,1,0.5) == 0, "Heads", "Tails")  # flip a coin five times
cat(result)
```

```
## Heads Heads Tails Tails Tails
```

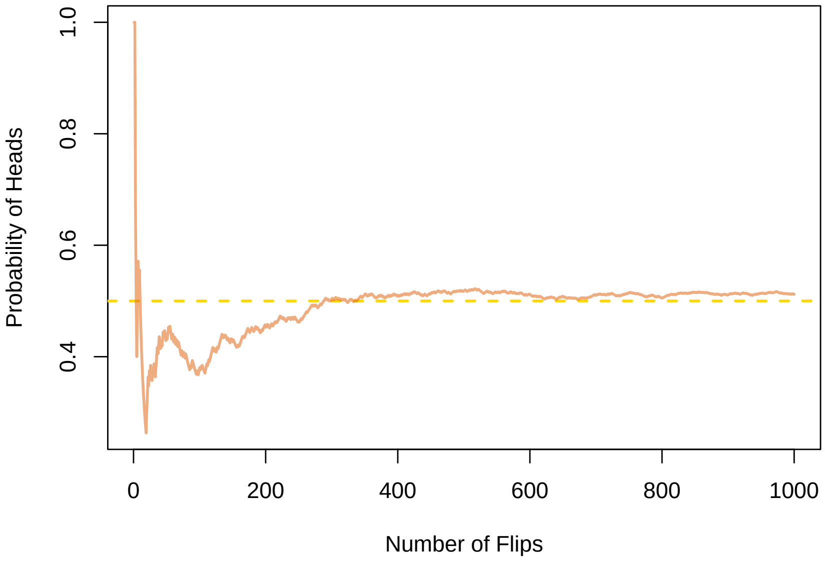
$$P(\text{heads}) \approx \frac{2}{3} \approx 0.67$$

A side effect of empirical probability is that it doesn't function very well under small sample sizes. Changing the component of the code that returns an answer of "heads" or "tails" to instead add up all of the 1's that occur will tell us the total observations of tails. If we subtract that number from the total number of tosses we'll end up with our count of heads. So let's flip this coin 1000 times, count up the number of heads, calculate the probability, and plot out the observations of our experiment:

```r
set.seed(73)
# this can also be done without the sum() function by using rbinom(1,1000,0.5)
result=1000-sum(rbinom(1000,1,0.5))
cat("Total heads:",result)
```

```
## Total heads: 512
```

$$P(\text{heads}) \approx \frac{512}{1000} \approx 0.512$$



What we've observed is something called the "Law of Large Numbers". As the size of our sample (i.e., number of experiments) gets larger and larger, the relative frequency of the event of our interest gets closer and closer to the **true probability**.

Assume that a fair die is rolled (i.e., all outcomes are *equally-likely*)

1. What is the **sample space**?
2. What's the **probability** of rolling a 5?
3. What's the probability of rolling an **even** number?
4. What's the probability of rolling a number **less than** 3?

An automobile insurance company divides customers into three categories: good risks, medium risks, and poor risks. Assume that of a total of $11,217$ customers, 7792 are good risks, 2478 are medium risks, and 947 are poor risks. As part of an audit, one customer is chosen at random.

a. What's the probability that the customer is a good risk?

b. What's the probability that the customer is not a poor risk?