Biostatistics in Public Health

Ralph Trane

Fall 2019(last compiled: 2019-09-30)

Contents

Chapter 1

Introduction to Notes

Welcome to Public Health 783. This is where you'll find my notes for the biostatistics part of this class for Fall 2019.

These notes are meant to be a supplement to lectures and Sullivan (2017). If you have any doubts as to what parts of the book I think are more important, this is a good place to look. There's a good chance that if it is not included here, I do not care too much about it. That being said, things included here might not be elaborated to a satisfying extent, and therefore I do *NOT* recommend that you rely solely on these notes, but rather use them in conjunction with the book, and, just as importantly, lectures.

I want to emphasize that these notes are not exhaustive, but are meant to complement lectures. We simply won't have time to cover everything in the amount of detail I would like. Therefore, I will use these notes to write down some thoughts on certain topics that I think you might as well read outside the classroom. At times I will ask you to read sections before showing up to lectures. Other times, I won't even make the material available to you until after lectures. There's good reason for that. Some things are easier to explain when you have already seen them before. Other concepts can be scary on paper, and are best introduced by another human being. In either case, it is extremely important that you show up to lectures.

Finally, note that this collection is a living thing. It will change throughout the semester as we move along. Sometimes I might go back and clarify sections, if it seems to be necessary. At the beginning of the class you'll only find the first part, but as we make our way through, I will publish more notes on the topics to be covered later.

Bounty Program

As already mentioned, this is a living document. Furthermore, it is very much so the first version. Therefore, there will inevitably be mistakes. Hopefully, most of them will be minor, but I would be surprised if there isn't a few major ones here and there. Therefore, I really hope that you will all help me out and let me know whenever you find mistakes, no matter how small they might seem.

Now, I know that noone wants to do that – who has the time to send an email just to let some fool know they can't spell "stastitics". But it's important to me to gradually improve these notes, and often being made aware of small mistakes opens up ones eyes to bigger ones. So to incentivize you to report mistakes, I have decided to start a bounty program: if you as a group gather more than 20 points (see below on how to collect points), there will be a reward at the end of the semester! If you

find a mistake, simply shoot me an email, let me know where the mistake is, and what type you think it is.

Type of Mistake	Points
Spelling/grammatical error	1
Math error	2
Conceptual nonsense in either text or math	3

Chapter 2

Introduction to Biostatistics

2.1 What is Biostatistics?

Biostatistics is "simply" statistics applied to a specific set of problems, namely problems related to biological questions. Hence, the question "what is biostatistics?" is quickly replaced by "what is statistics?"

From Wikipedia[1]: "Statistics is the discipline that concerns the collection, organization, displaying, analysis, interpretation, and presentation of data." That is quite the range, but all it is saying is that statistics is the science of making sense of data.

It is very hard to pinpoint exactly what statistics is, but it is rather easy to dismiss at least one very common misunderstanding: statistics is *NOT* an exact science. There is (almost) never just one answer to a question. Rather, statistics is a decision science in the sense that at every step of the way, from study design to data collection to data presentation to data analysis and interpretation, you have to make decisions. And inevitably, your conclusions depend on every single one of those decisions.

2.2 Biostatistics in PUBLHLTH 783

In this class, we'll be considering what might seem like a very simple setup with simple questions, and a general approach to answering said questions. The truth is basically all statistical methods, no matter how complicated they might seem, follow (to some extend) this exact pattern. In this class, we will work our way through this pattern, and talk about how the simplest of methods work. The hope is that when we're done, you take these simple methods with you, and whenever you encounter more complicated methods, you can draw parallels back to what you have seen here, which hopefully will help you make at least some sense of even the most complicated methods.

The story begins...

Any statistical analysis begins with a research question. This question always relates to a certain *feature* or *characteristic* of a certain *population*, or of multiple populations. The population is the group of entities one wants to learn more about. It could be that one is interested in the life expectancy (feature) of U.S. adults (population), risk of diabetes (feature) among people with a certain genetic profile (population), the resistance to a disease (feature) among ants in Governor Dodge State Park (population), strength (feature) of pipes produced by a certain manufacturer (population), etc.

---

[1] ... because where else do you go when you're not sure where to start?

Now, the first fundamental idea of the approach we'll pursue in this class is this: there is a single true answer to the question asked. The only one way to obtain the truth is to observe the entire population. If we could measure the disease status (diabetes/no diabetes) of every single individual with the genetic profile of interest, we could evaluate the true risk. The catch is, maybe obviously to someone, that this is simply not feasible. Often we do not even know exactly how big the population of interest is, which means it is impossible to survey every single subject in the population. And even if we knew exactly how many people were part of the population, how to get a hold of them, and how to convince them to participate in our study, it is very unlikely we would have the resources (read: $$$) to reach out and include every single one of them.

This leads us to the picture we will return to over and over again in this class: there is some population of interest. We want to say something about a certain feature of the population. This will usually be a mean of a measurement, the risk ratio or odds ratio of two groups, or something similar. Unfortunately, there is no way we can actually measure the feature for every single subject in the population – if we could, we would simply calculate the mean, and be done with it! Instead, we take a *sample* from the population. The hope is that we obtain this sample in such a way that the sample is representative of the entire population, and therefore we can use the information obtained from the sample to say something about the population.
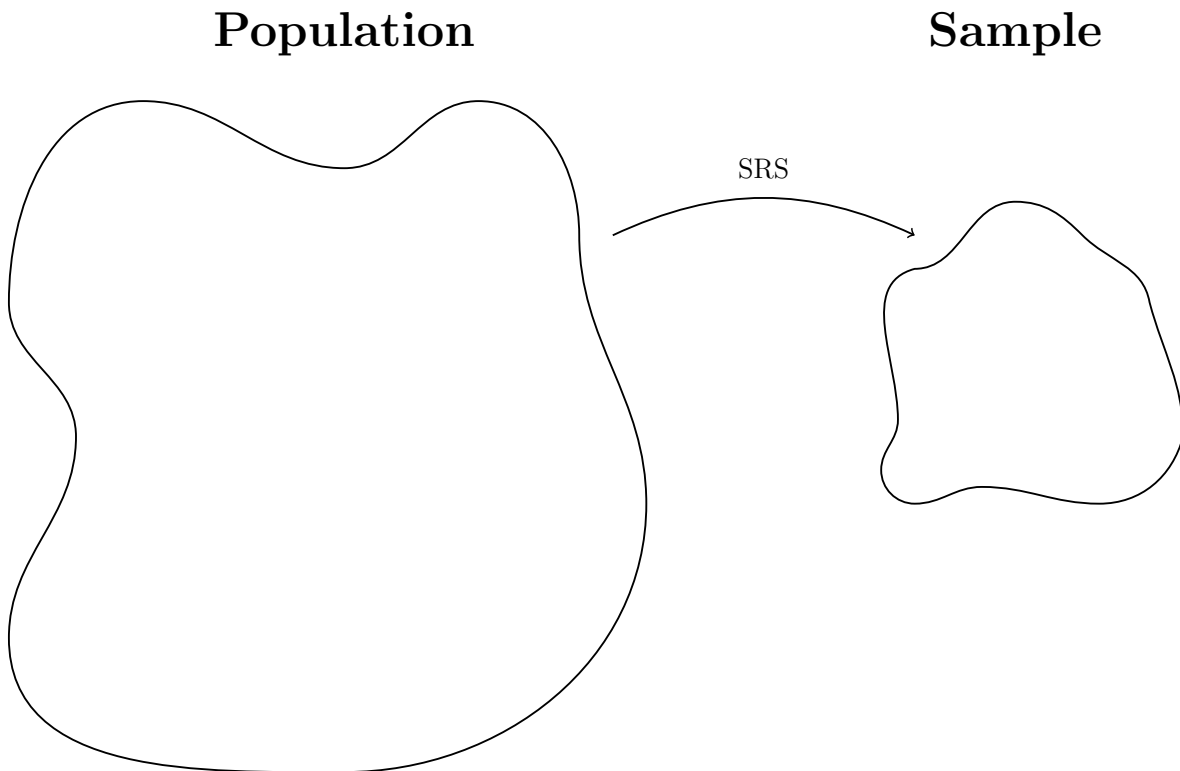
Population   Sample

When we talk about chances of observing something when sampling from the population, we talk about *probabilities*. What is the probability a sample of 23 people has more than 5 diabetics? What is the probability a randomly chosen individual from Minneapolis will develop cancer?

When we try to use a sample to say something about the general population from which the sample was created, we do *inference*.

The plot thickens...

We now have an idea of what it is we want to do: take a sample, obtain some information about the

thing of interest, then use that information to say something about the general population. Pretty simple. The problem is, how do we actually put the pieces together in a way that allows us to generalize to the population? Example: we are interested in estimating the prevalence of cardiovascular disease in the general population of U.S. adults. We have a hunch that the true prevalence is 11%. We take a sample of 3799 female U.S. adults of which 379 have a history of CVD. So, the prevalence is estimated to be $\frac{379}{3799} \approx 10\%^2$. The prevalence in our sample is clearly different from our hypothesis, so clearly our hypothesis is wrong…. right?

If only it was that simple. The problem here is that our prevalence estimate depends on the specific sample we got. If we were to repeat this experiment, we would ask a different group of people about their history of CVD, which would lead to a different estimate of the prevalence. There's simply no way (or it is at least very, very unlikely) we'll ever get a sample of people for which the prevalence matches our hypothesis exactly. So the question is not simply if our sample has the same prevalence as hypothesized, but rather is it "close enough" for us to believe our hypothesis.

Close enough? Did I read that right?

Yup. The majority of statistical methods, and definitely everything we'll be talking about in this class, are trying to decide if what we observe in our sample is "close enough" to our hypothesize about the population. The idea is that if what we observe is very far from our hypothesis, then it is unlikely that the hypothesis is true.

There are generally two ways of framing the question:

1. Is what we observe (think prevalence in sample) "close enough" to the hypothesis (that the prevalence in the population is 11%)?
2. What range of hypotheses (values for the prevalence) would we accept given what we observed?

The first approach is refered to as testing a *statistical hypothesis*, or *null hypothesis significance testing* (NHST), while the second is refered to as constructing a *confidence interval*. These two are closely related, as we will see later, but the information contained in the results differ drastically. While testing a hypothesis only provides you with a result concerning **one value**, the confidence interval gives you a wide range of values that you to some extent believe could be the true value. Why would anyone then ever report the result of a statistical hypothesis test without a confidence interval, you ask? That is a very, very, very good question to which I do not have an answer…

This has all been very abstract (and I think I got carried away in that last paragraph). We'll see examples of this over and over again throughout the semester, so hopefully it'll be easier to comprehend as we go along.

---

[2]This data was obtained from the Framingham Offspring Study, see table 3-1, page 24 of Sullivan (2017)

Part I

Data Types and Descriptive Statistics

Chapter 3

Before we get started...

This section is very minimal. I decided to go that route because 1) this is the "easiest" part of the material in the sense that there isn't a lot of things to really understand and think about, and 2) it is objectively speaking very, very, very boring... So, when you find my notes insufficient, take a look at Chapter 4 in (Sullivan, 2017).

Learning Objectives

1. Understand why descriptive statistics is important, and useful

2. Know the difference between discrete and continuous variables/data

3. Have some ideas of which summaries and figures are appropriate for different types of data

Chapter 4

Why Descriptive Statistics?

Descriptive Statistics, as the name implies, are about describing something. In general, we can only describe what we have at hand, so descriptive statistics only deal with the population in the very, very, very rare case when we have measured the entire population. It is way more common that descriptive statistics are about the sample at hand.

Often when conducting an experiment, what we are actually interested in is using the sample we obtain to say something about the general population. If we, for example, enroll 200 patients in a study to find out if a new drug works, we're not *really* interested in whether or not it works for those 200 patients specifically, but more so if it works for any individual from the population in general. With that in mind, the whole concept of descriptive statistics might not seem very (1) exciting, (2) useful, or (3) necessary. While (1) is highly subjective, (2) very much so depends on the specific situation, there is little question that it is in fact necessary.

Recall our general setup for any statistical analysis: we have some question about some feature of a population. To answer this question, we go get a sample of the population. If this is a good sample, the characteristics of this sample mimic those of the general population. If this is the case, then the hope is that we can draw conclusions about the sample, and generalize them to the general population. Conversely, if that is *not* the case, then no matter how convincing the evidence we find in the sample is for or against our hypotheses, it tells us nothing about the general population.

Population   # Sample

SRS
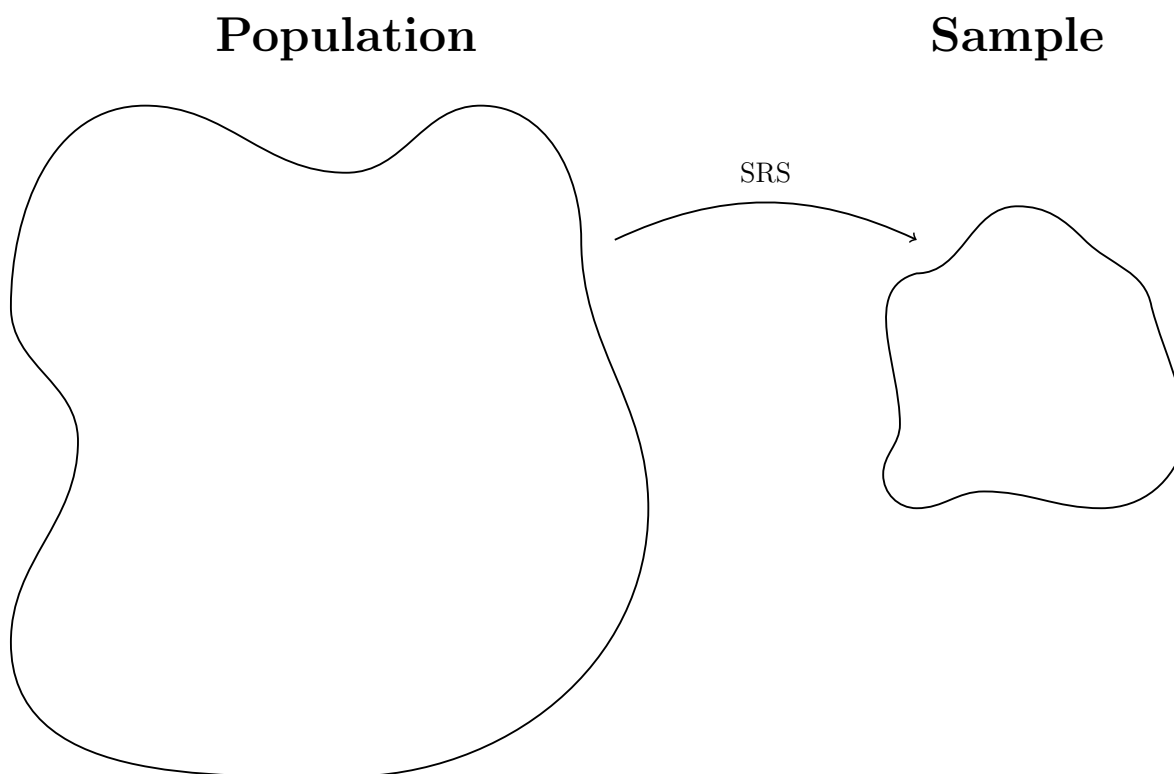
So how do we make sure the sample is representative of the population? The first step is to make sure our sampling scheme is good. Ideally, we sample completely at random, meaning that every single individual in the population has the same chance of ending up in the sample. As you might be able to guess, this is rarely the case, but if the sampling is done right it is either approximately true, or the sampling is done in a way that any biases introduced can be accounted for in the analysis.[1] Most commonly, sampling is done in such a way that the "equal chance" assumption isn't too crazy. But how do we know if this is actually the case?

The truth is, we don't really. What we can do, though, is describe the sample we obtain. That way we can make sure we don't generalize any results to an inappropriate population. Historically, this mistake has been made over and over again in medical research when excluding women and ethnic minorities from studies.[2] For this particular reason, producing descriptive statistics is often the first step in any data analysis.

The rest of this section will go through different types of data, and show how we, in each case, can describe (or summarize, if you will) the specific type of data.

In general, when we talk about data we often refer to *variables*. Variables are simply (and very vaguely) *things we measure*. You will see examples which hopefully helps understand exactly what is meant by a variable.

---

[1]**Full disclaimer**: sampling is hard. As in really, really, really hard. Very smart people spend a lot of time on making sure different sampling schemes work. It is well beyond the scope of this course, and me, how the details of this work out.

[2]See Liu and Mager (2016) for a more thorough discussion of the exclusion of women in particular.

Chapter 5

Discrete Data

A variable is called a *discrete* variable if the possible values of the variable are countable, that is if you can count them. Note that a discrete variable can technically have an infinite number of possible outcomes.

A discrete variable is of one of two subtypes: *categorical* or *ordinal*.

5.1 Categorical data

Categorical variables are discrete variables with no particular ordering of the categories.

5.1.1 Examples – categorical data

The classical example of a categorical variable is sex. For each subject, the value of this variable is one of two possible values: male or female.

Other examples:

- color
- race
- blood type
- country of origin
- political orientation

5.1.2 Binary/Dichotomous Data

Often researchers will refer to certain variables as *binary* or *dichotomous*. This simply means *categorical with two categories*.

5.2 How to describe categorical data

Categorical variables are often described using *frequency counts* and *relative frequencies*.

Frequency counts (or simply *frequencies*) are found by counting how many times each possible value is present in the data. *Relative frequencies* are found by dividing the frequency by the total number of observations.

5.2.1   Examples

Below are the frequencies for some categorical variables in the SHOW data set. One thing that often comes from this preliminary step of a data analysis is the realization that there are some kinks in your data. For example, notice how there are missing values for most of these variables (denoted by NA – "Not Available").

- **edu**:

Value	Frequency
[0] Never attended/kindergarten only	2
[3] 3rd grade	2
[4] 4th grade	1
[6] 6th grade	4
[7] 7th grade	2
[8] 8th grade	20
[9] 9th grade	22
[10] 10th grade	49
[11] 11th grade	66
[12] 12th grade, No diploma	90
[13] High school graduate	624
[14] GED or equivalent	112
[15] Some college, no degree	680
[16] Associate degree: occupational, technical, or vocational program	437
[17] Associate degree: academic program	191
[18] Bachelor's degree	723
[19] Master's degree	263
[20] Professional degree	47
[21] Doctoral degree	40
[.D] Don't know	1
NA	5

- **gender**:

Value	Frequency
[1] Male	1479
[2] Female	1901
NA	1

- **marital**:

Value	Frequency
[1] Married	2075

Value	Frequency
[2] Widowed	113
[3] Divorced	416
[4] Separated	41
[5] Never married	603
[6] Living with partner	126
[.D] Don't know	2
NA	5

- **race**:

Value	Frequency
[1] Non-hispanic white	2870
[2] Non-hispanic African American	243
[3] Hispanic	108
[4] Other race or ethinicity	151
NA	9

We can add relative frequencies to this simply by dividing each frequency by the total number of observations. You can check a few of them yourself – simply divide the frequency by the total number of observations (found on the last line of the table).

- **edu**:

Value	Frequency	Relative Frequency
[0] Never attended/kindergarten only	2	0.0005915
[3] 3rd grade	2	0.0005915
[4] 4th grade	1	0.0002958
[6] 6th grade	4	0.001183
[7] 7th grade	2	0.0005915
[8] 8th grade	20	0.005915
[9] 9th grade	22	0.006507
[10] 10th grade	49	0.01449
[11] 11th grade	66	0.01952
[12] 12th grade, No diploma	90	0.02662
[13] High school graduate	624	0.1846
[14] GED or equivalent	112	0.03313
[15] Some college, no degree	680	0.2011
[16] Associate degree: occupational, technical, or vocational program	437	0.1293
[17] Associate degree: academic program	191	0.05649
[18] Bachelor's degree	723	0.2138
[19] Master's degree	263	0.07779
[20] Professional degree	47	0.0139
[21] Doctoral degree	40	0.01183
[.D] Don't know	1	0.0002958
NA	5	0.001479

- **gender**:

Value	Frequency	Relative Frequency
[1] Male	1479	0.4374
[2] Female	1901	0.5623
NA	1	0.0002958

- **marital**:

Value	Frequency	Relative Frequency
[1] Married	2075	0.6137
[2] Widowed	113	0.03342
[3] Divorced	416	0.123
[4] Separated	41	0.01213
[5] Never married	603	0.1783
[6] Living with partner	126	0.03727
[.D] Don't know	2	0.0005915
NA	5	0.001479

- **race**:

Value	Frequency	Relative Frequency
[1] Non-hispanic white	2870	0.8489
[2] Non-hispanic African American	243	0.07187
[3] Hispanic	108	0.03194
[4] Other race or ethinicity	151	0.04466
NA	9	0.002662

Relative frequencies are useful when trying to compare the values of a specific variable across groups. Say we want to investigate if there are any differences in the marital status between genders in this cohort. We could consider the frequency of marital status stratified by gender:

Table 5.9: Table continues below

gender	[.D] Don't know	[1] Married	[2] Widowed	[3] Divorced
[1] Male	0.001 (1)	0.627 (928)	0.021 (31)	0.103 (153)
[2] Female	0.001 (1)	0.603 (1147)	0.043 (82)	0.138 (263)
NA	0.000 (0)	0.000 (0)	0.000 (0)	0.000 (0)

Table 5.10: Table continues below

[4] Separated	[5] Never married	[6] Living with partner	NA__
0.005 (8)	0.205 (303)	0.036 (53)	0.001 (2)
0.017 (33)	0.158 (300)	0.038 (73)	0.001 (2)
0.000 (0)	0.000 (0)	0.000 (0)	1.000 (1)

Total
0.999 (1479)
0.999 (1901)
1.000 (1)

The table above shows the relative frequencies of marital status within each gender (frequency counts in parentheses). So we can see that the relative frequencies of men and women in the different groups are very close to each other. If you just look at the raw frequencies, this would not be obvious.

5.3   Ordinal Data

An *ordinal variable* is a discrete variable where the groups can easily be ordered in a meaningful sense. We won't distinguish between ordinal variables and categorical variables in this class, but there are methods out there that try to incorporate the extra information from ordinal data that you lose if you treat it as categorical data.
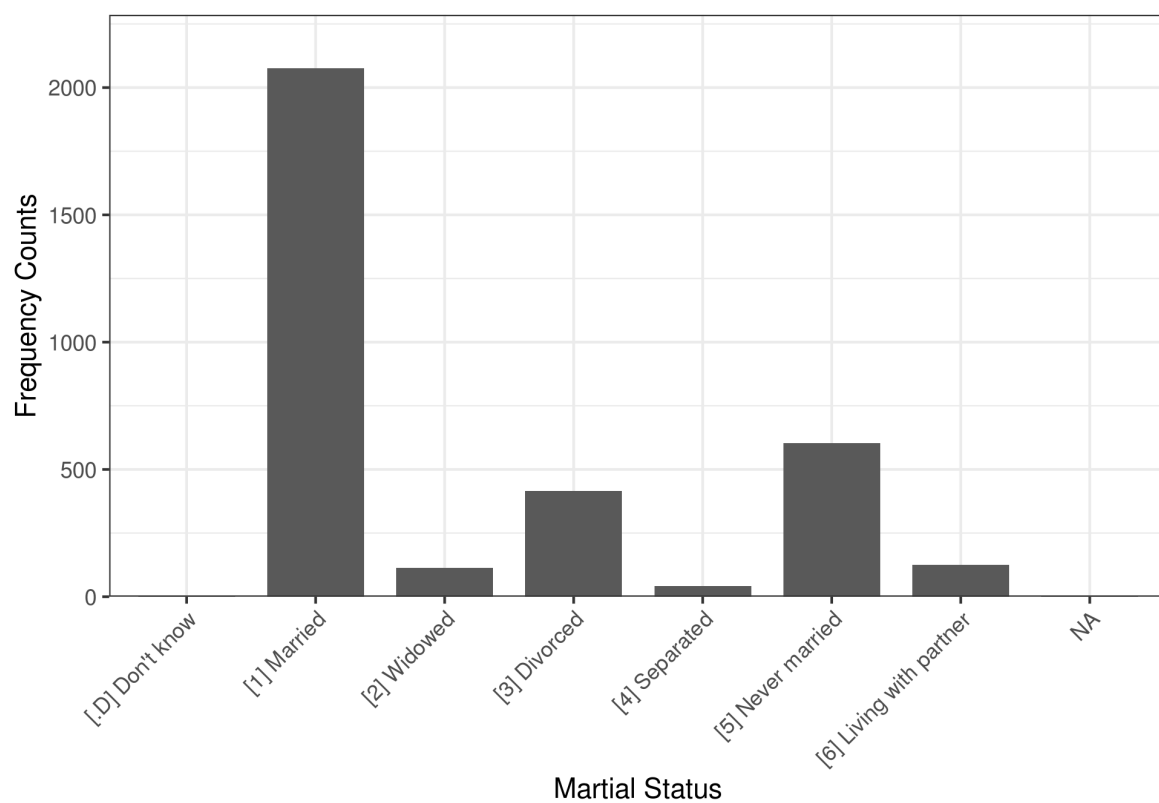
5.3.1   Examples

- age groups
- disease severity scales
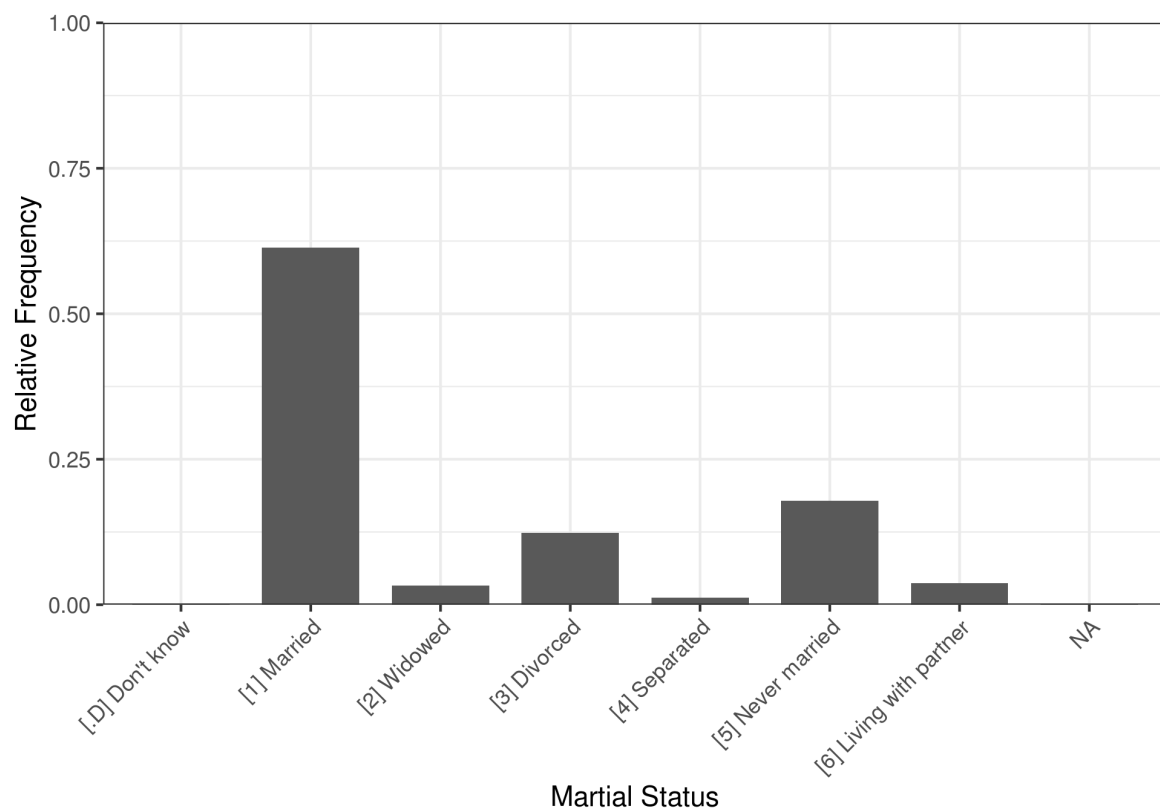
5.4   How to visualize discrete data

Discrete data is often best presented using a bar chart of either frequency counts or relative frequencies.

5.4.1   Bar Charts

Below is a bar chart of the frequency counts of the marital status varaible from the SHOW data.

This can easily be turned into a bar chart of the relative frequencies.

Chapter 6

Continuous Data

A *continuous variable* is a numerical variable that can (at least theoretically) take on an infinite and uncountable number of possible values.

6.1 Examples

- age
- height
- speed
- blood pressure
- heart rate

6.2 How to describe continuous data

When dealing with continuous data, we are often interested in two aspects:

1. location
2. spread

6.2.1 Location

The location of continuous data is often described by one of two metrics: the *mean* (or *average*) and the *median*. For completion, these are briefly defined below. For a more in-depth discussion of the mean and the median, see (Sullivan, 2017) pages 50-57

6.2.1.1 Mean

The *mean* of a variable measured in a sample is also referred to as the *sample mean* or *average*. It is simply calculated as the sum of all observed values divided by the number of values. If we are interested in a variable X, the average is denoted $\bar{X}$ (read "X bar"). So $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.[1]

---

[1]If you are not familiar with this notation, fear not: it simply means "sum up all values of X". I.e., $\sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + ... + X_n$.

6.2.1.2 Median

The *median* is the middle point of the data. It is found by writing down all observations in order, then eliminating the most extreme pair (i.e. the smallest and largest values). Repeat until only one observation, or one pair of observations, is left. If you're left with one observation, congratulations, you found the median. If you're left with a pair, the median is the average of the two.

6.2.1.3 Other location related metrics

Sometimes we are interested in the most extreme values we can expect of a value. Here it would be of interest to find the *minimum* and *maximum* of the variable.

More generally, the minimum, median, and maximum values are examples of what is called *quantiles*. A quantile is a number that "cuts off" a certain proportion of the data (from the bottom). You can think of the median as the number that "cuts off" half the data. Therefore the median is also called the 0.5 quantile. The minimum is the 0 quantile (it cuts off nothing of the data), and the maximum the 1 quantile (it cuts off all the data). The most commonly used quantiles are the *quartiles*. This is the set of numbers that cut the data into four equally sized pieces. I.e. the 0.25, 0.5, and 0.75 quantiles are collectively known as the quartiles. These can be useful when talking about the location of the data, since indicating the 0.25 and 0.75 quantiles tells you where half the data is located – namely between those two values. The quartiles are also often referred to as Q_1, Q_2, and Q_3.

6.2.2 Spread

Once we have an idea of the location of a continuous variable, the next natural question is how large the spread (or variation) is.

We will here briefly introduce four (but kind of only three...) metrics for the spread of the data: variance, standard deviation, range, and interquartile range.

6.2.2.1 Variance/Standard Deviation

The *variance* of a continuous variable is in many ways "the average (squared) deviation from the mean". It is calculated as $\text{Var}(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$. So larger variance means larger spread, and vice versa.

The *standard deviation* is simply the square root of the variance: $\text{SD}(x) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$. Therefore, there's a one-to-one correspondance between the variance and the standard deviation. This also means that when the standard deviation is large, so is the spread.

A natural question is then: why do we need both? The variance is nice for mathematical reasons, as we will see later. It also provides this nice interpretation as an average, which we lose when converting to the standard deviation (because of taking the square root). On the other hand, the standard deviation is nice because it kind of encapsulates the same idea as the variance, but preserves the unit. We'll have a more detailed discussion of this in later sections.

6.2.2.2 Range

The *range* is simply the difference between the largest and the smallest value. Hopefully it is clear that this indeed is a measure for how spread out the data is. But it is not always a super useful measure – you could have a sample where 293 observations are the exact same, and the last two observations are

very, very different. In such a case, the range will indicate quite the spread, while in truth the data is not spread out very much at all.

### 6.2.2.3 Interquartile Range (IQR)

The interquartile range is simply the difference between the first and the third quartile. I.e. IQR $=$ $Q_3 - Q_1$. This is also the size of the box in a box plot (see section 6.3.2).

6.3 How to visualize continuous data

My favorite graphs to use with continuous data are scatter plots, boxplots, and histograms.

6.3.1 Scatter Plots

A scatter plot is only really useful when you are considering the relationship between two variables where at least one is continuous. For example, consider the variables `height` and `weight` from the SHOW data. A scatter plot shows potential relationships between the two. Unsurprisingly, it seems that there is a positive correlation between the two – i.e. when one goes up, so does the other.
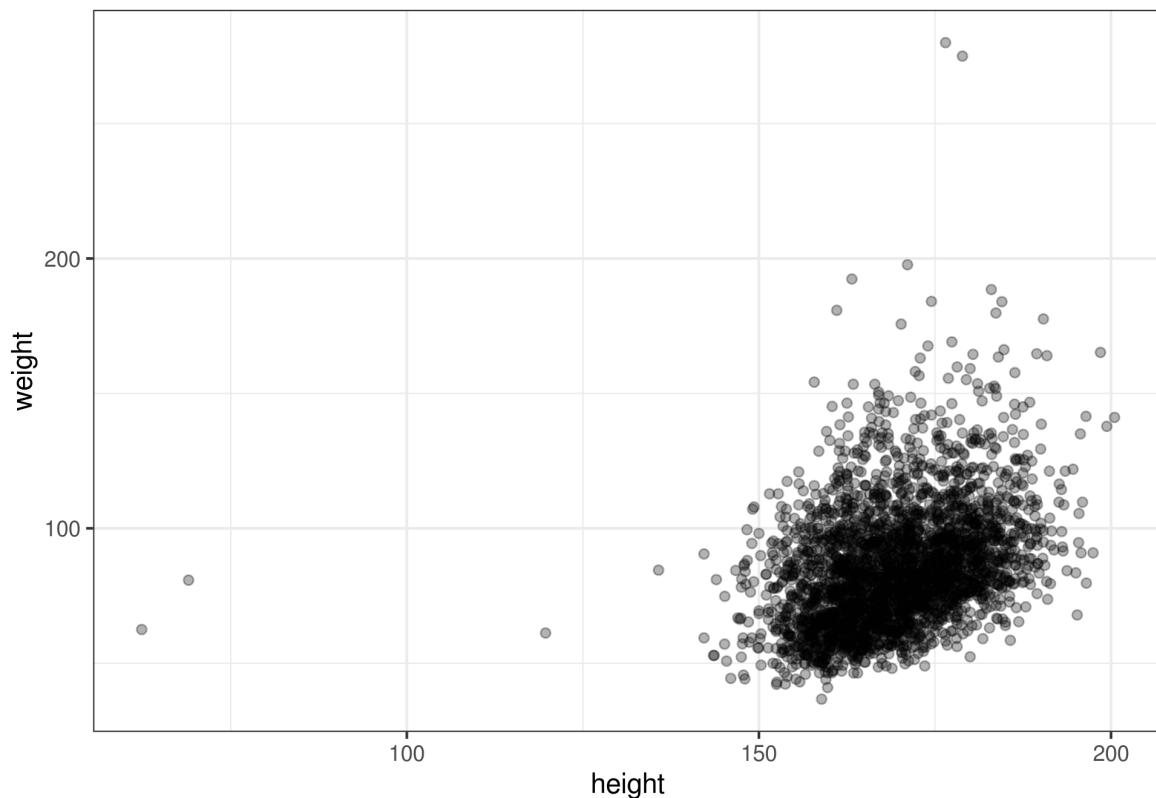
You can also utilize scatter plots when one of the variables is categorical. For example, we could be interested in the relationship between height and gender.
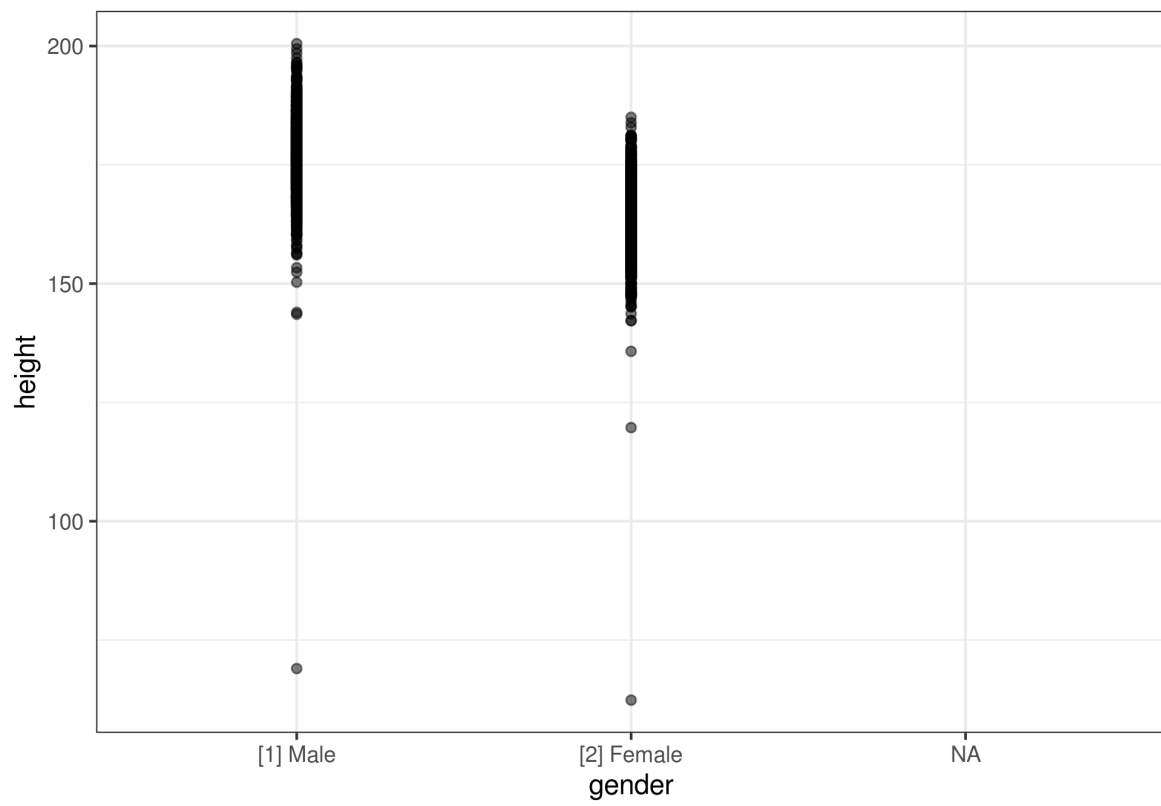
In this case, it can be beneficial to add a bit of jitter to the plot in the direction of the categorical variable.
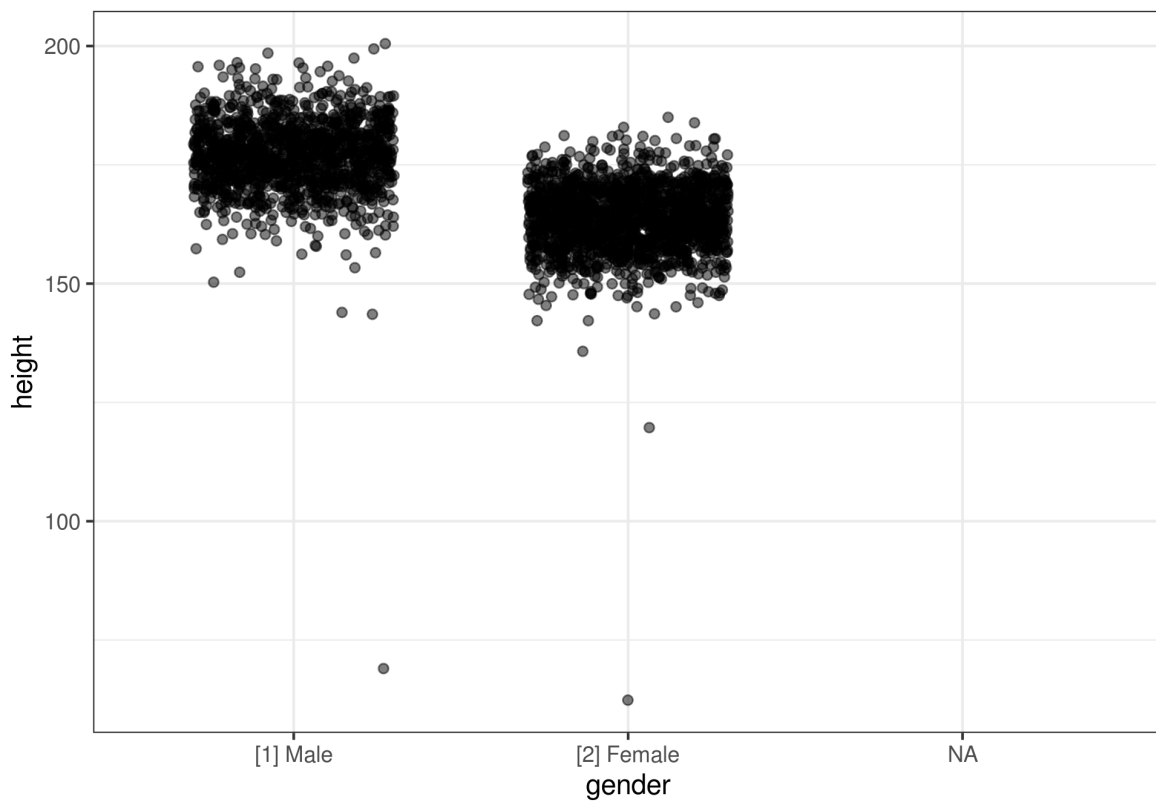
Even with a bit of jitter, it might be really hard to make anything of a scatter plot in this case, simply because we have "too much" data. In such a case, a boxplot might be a better choice.

6.3.2 Boxplots

Boxplots are great when you have a lot of data. They show the data through a set of summaries, namely the quartiles, and indicates if there are any *outliers*. Below are boxplots for the height of the SHOW population by gender.
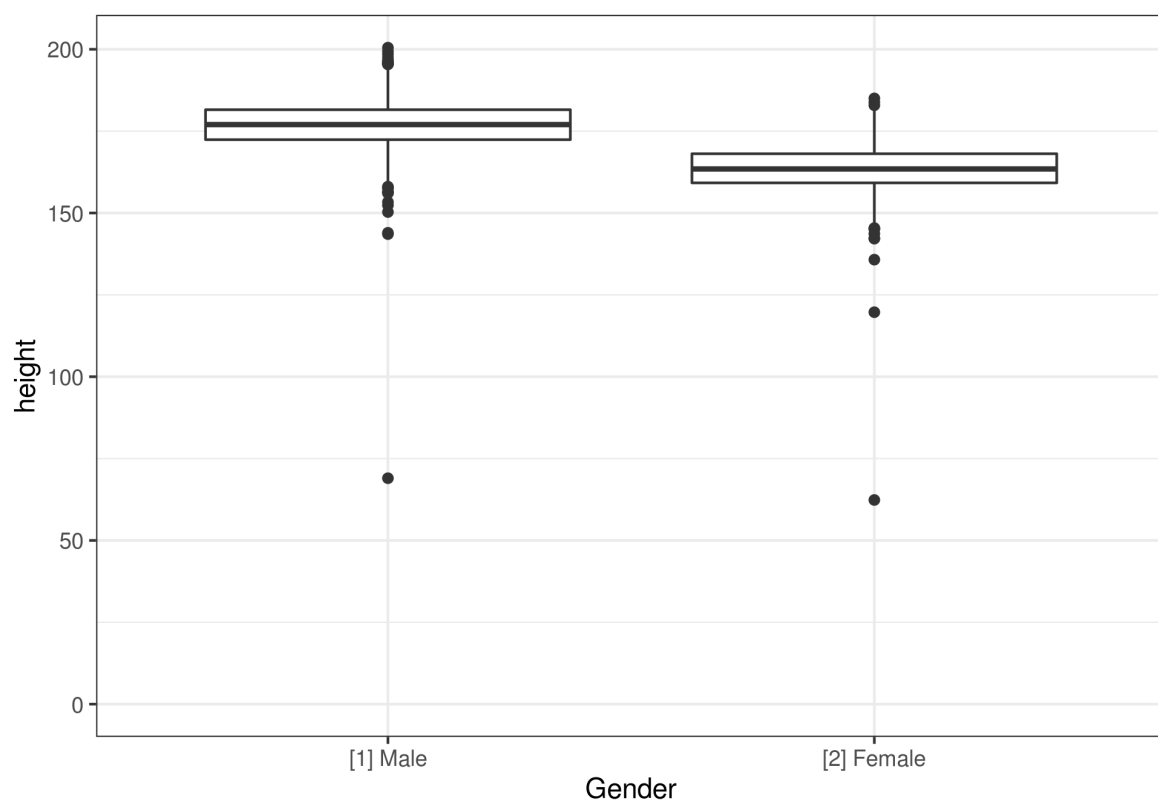
You can use the figure below to decipher the box plot:
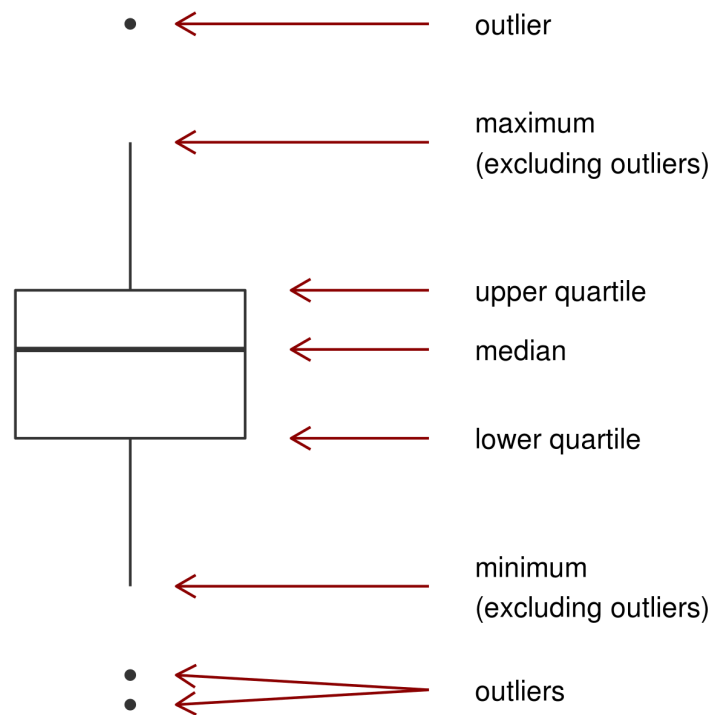
As you see on the boxplots of the SHOW data, it is a great tool to visualize continuous data when you have a lot of it. In a simple figure we can see that

- the median height is greater for men than women
- there is generally a shift upwards for men compared to women
- the 75% tallest men are all taller than 75% of women (compare the bottom of the box for men with the top of the box for women)

and much more.

One thing I haven't told you is the answer to a very important, very hard question: "how do we decide if a data point is an outlier?" We will simply adopt the practice that a data point is an outlier if it is more than 1.5 times the range of the box from the box. I.e. an observation is an outlier if it is greater than $Q_3 + 1.5 \cdot (Q_3 - Q_1)$ or less than $Q_1 - 1.5 \cdot (Q_3 - Q_1)$.

6.3.3 Histogram

At first, the *histogram* looks a lot like a bar chart, but there are a few very important differences. Before we go into details, lets take a look at a histogram. Below is a histogram of the depression scores in the SHOW data set.
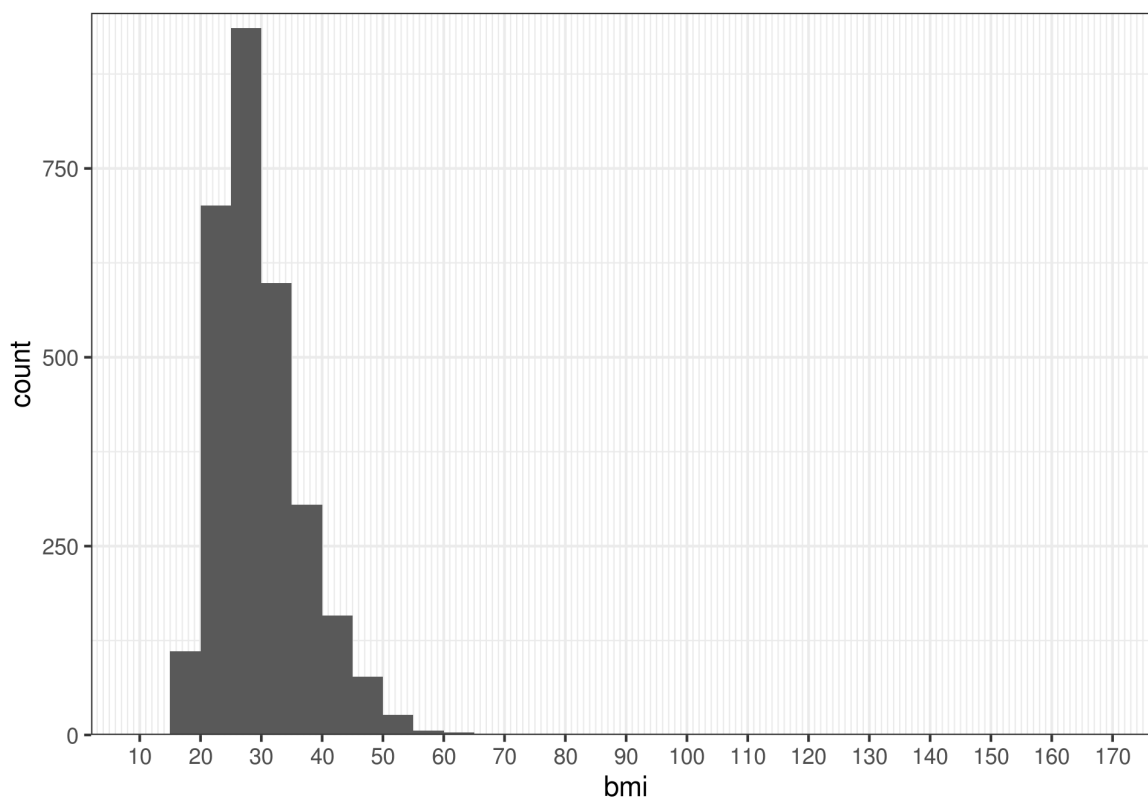
The main differences from a bar chart is that

1. there are no gaps on the x-axis
2. the *relative area* of a bar is the proportion of your sample that falls in the interval corresponding to that bar

Later on, we will use the histogram to answer questions like "what is the probability a randomly chosen individual from the SHOW population has a BMI greater than 40?" or "between 20 and 30?" etc. This is simply done by dividing the area of the bars that are specified (for example all bars with BMI greater than 40) with the total area.

The histogram will be **super** important to us moving forward, so make sure you know how to decipher it!

Chapter 7

Grey areas

An example of a variable that could easily be mistaken as categorical is age. Often when we think about age, we think about this in terms of years, or months, or even days. In that sense, age is a variable with a number of possible values that we could technically count – start with 0, 1, 2, 3, ..., 55, 56, 57, ... However, this is NOT the natural structure of the variable, but rather a limitation of the way it is measured and recorded. Technically, age is the time from birth till now, which if we could measure it with *infinite* precision, could be any possible number[1] you can think of.

Many other examples could be provided. In general, what's important to think about is the nature of the variable rather than what is measured. By nature, any measurement is going to be discrete, but some variables are continuous in nature.

---

[1] positive, real number, for those of you who want to be specific

Part II

Introduction to Probability

Loosely based on Sullivan (2017) chapter 5.

Chapter 8

What is "probability"?

A *probability* is a number between 0 and 1 that indicates how likely it is that a certain event happens. An event that has the probability of 1 **always** occurs, while an event with probability of 0 **never** occurs. Every number in between are a bit harder to interpret.

For example, an event with probability 0.5 supposedly happens every other time. This makes sense if you think about something that can be repeated, such as a coin flip, or the roll of a die, but how does that work if we consider an event that only occurs once? For example, how do we interpret a weather forecast that claims there's a 0.5 (i.e. 50%) chance of rain tomorrow? We can only observe if it rains tomorrow once, so the probability surely must be 0 (it doesn't rain) or 1 (it rains), right?

8.1 Definitions

As hinted at above, the concept of "probability" can be a bit challenging to wrap your head around. There are generally two ways that the term is introduced. Though they are very similar once you understand the concepts, they can seem radically different at first.

Definition 8.1. The probability of an event is the number of outcomes that ensure the event happens divided by the total number of possible outcomes, **IF** all outcomes are equally likely:

$$P(\text{event}) = \frac{\text{number of outcomes that result in event}}{\text{total number of possible outcomes}}.$$

We often refer to the numerator in this fraction as the number of favorable outcomes.

I want to take a second to draw your attention to that small, but incredibly important, final bit of the definition: "IF all outcomes are equally likely". We will later discuss what to do if this is not the case, but for now, this will be an underlying assumption.

The best way to become comfortable with this definition is by considering a few simple examples. The following two examples are the most commonly used, and (by far!!) most boring examples in the history of statistics. However, they are super useful for two reasons:

1. They are so simple that it is possible to better grasp what's going on
2. A lot of more complicated examples can be simplified by comparing them to these two

8.1.1   Examples

8.1.1.1   Coin Flip

We want to find the probability P(coin comes up heads). A natural assumption is that when flipping a coin, heads and tails are the only outcomes[1], and they are equally likely. Therefore,

$$P(\text{coin comes up heads}) = \frac{\text{number of possible outcomes that come up heads}}{\text{number of possible outcomes}}$$
$$= \frac{1}{2}$$
$$= 0.5.$$

Similarly, one can find the probability that the coin comes up tails:

$$P(\text{coin comes up tails}) = \frac{\text{number of possible outcomes that come up tails}}{\text{number of possible outcomes}} \tag{8.1}$$
$$= \frac{1}{2} \tag{8.2}$$
$$= 0.5. \tag{8.3}$$

8.1.1.2   Roll of a Die

Another classic example: calculate different probabilities when rolling a die. (Done in class – see lecture notes.)

––––––––––––––––––––––––––––

The two examples above show situations where all possible outcomes are equally likely. What if that is not the case?

8.1.2   Example: disease status

Let us consider the SHOW data set. We might be interested in the probability of a subject being obese. Now, there seems to be only two outcomes here: either the subject is obese, or the subject is not. So, using the same string of thoughts as above, one might conclude that the probability of a subject being obese if $\frac{1}{2}$, i.e. 0.5.

This is obviously not the case. The problem with this approach is that the two outcomes – those being "the subject is obese", and "the subject is NOT obese" – are not equally likely, so the simple approach of simply dividing the number of favorable outcomes by the number of possible outcomes is not doing us any good.

––––––––––––––––––––––––––––

To find a more satisfying answer to the question asked in the last example, we need to consider a different approach to probabilities.

Definition 8.2. The probability of a specific outcome from an experiment is the proportion of times the outcome occurs if the experiment is repeated an *infinite number of times*.

––––––––––––––––––––––––––––

[1]i.e. it is NOT possible for the coin to land on the side

Repeating an experiment an infinite number of times is obviously not possible, so in practice "an infinite number of times" becomes "a very large number of times".

When introducing this different approach to probabilities, first we need to make sure it doesn't contradict our previous approach.

8.1.3   Example: coin flip (revisited)

We previously established that when flipping a coin, the probability of heads is 0.5. Hopefully this new definition will yield a similar answer.

To find out if that is actually the case, we would have to flip a coin "an infinite number of times". Obviously, this is not possible, so we will have to settle for "a very large number of times". So, imagine we flip a coin 100000 times. Every time it is flipped, we write down the result, and count how many times we've seen heads, and how many times we've seen tails so far. If the probability of seeing heads is 0.5, we should eventually see about as many heads as tails.

Below is an animation that shows the results of such an experiment. The bars show you the proportion of heads and tails, which in the end (by the definition above) will converge to the probability. The first 100 flips are all shown, then only the results after every 100 flips, and finally results after every 1000 flips are shown. Note how at the very end the two bars are both very close to 0.5.

8.1.4   Example: roll of a die (revisited)

Similarly to what we did above for the coin flip, we will do here for the roll of a die.

8.1.5   Example: disease status

Okay, so both when flipping a coin and rolling a die, the second definition agrees with the first one. But how can we use this way of thinking in the disease status example? What does it even mean to "repeat the experiment", let alone "repeat an infinite number of times"?!

In such a situation, we make a (very crude, but very necessary) assumption: we assume that all the subjects in the cohort are "similar enough" that we can pretend that observing the disease status of multiple people constitutes multiple experiments. We then estimate the probability of having the disease as the proportion of subjects with the disease.

Let's consider the probability that a person from the SHOW population is mildly depressed. To estimate this, we simply divide the number of individuals in the population who are mildly depressed with the total number of people in the population.

Below are estimated probabilities for all depression severity levels. Note: $P(\text{mildly depressed}) = \frac{454}{3381} \approx 0.134$.

Depression Severity	Count	Estimated Probability
[1] No depression	1629	0.482
[2] Mild depression	454	0.134
[3] Moderate depression	125	0.037
[4] Moderately severe depression	52	0.0154
[5] Severe depression	15	0.00444
NA	1106	0.327

Depression Severity	Count	Estimated Probability
Total	3381	1

Chapter 9

Conditional Probability

So far, we have talked about probabilities in a context where no additional information is available about the experiment. This is of course not always the case, and also not always what we are interested in.

A useful concept in these cases is the concept of *conditional probabilities*. In a nutshell, conditional probabilities deal with the chances of something happening given something else has already happened. If we consider two events, A and B, then we write $P(A|B)$ for the conditional probability of A given that B has happened, and read it as "the (conditional) probability of A given B".

9.1 Example: roll a die

Previously, we considered the probabilities associated with the roll of a die. We found that the probability of rolling a six is $\frac{1}{6}$. What if we somehow knew that the outcome turned out to be an even number, but simply didn't know which even number? Well, using this information, we know there are only three possible outcomes, namely $2, 4, 6$. They are all equally likely, so using definition 8.1, we find that the probability of rolling a six given the roll comes up even is

$$P(\text{roll a 6} \mid \text{roll is even}) = \frac{1}{3}.$$

9.2 Example: disease status

In the last section we found $P(\text{mild depression}) \approx 0.134$. Let us try to calculate the conditional probability of having a mild depression *given* the subject is divorced. The way to do this is to first create the two way contingency table:

Depression Severity	[.D] Don't know	[1] Married	[2] Widowed	[3] Divorced	[4] Separated	[5] Never married	[6] Living with partner	NA_	Total
[1] No depression	1	1102	57	174	12	239	43	1	1629
[2] Mild depression	0	253	17	76	4	84	19	1	454

Depression Severity	[.D] Don't know	[1] Married	[2] Widowed	[3] Divorced	[4] Separated	[5] Never married	[6] Living with partner	NA_	Total
[3] Moderate depression	0	55	1	27	3	33	6	0	125
[4] Moderately severe depression	0	13	4	15	1	13	6	0	52
[5] Severe depression	0	7	1	3	1	2	1	0	15
NA	1	645	33	121	20	232	51	3	1106
Total	2	2075	113	416	41	603	126	5	3381

To find the conditional probability, you basically narrow down the universe you operate in. Instead of asking "how many individuals have mild depression out of all individuals?" you ask "how many individuals have mild depression out of **individuals that are divorced**?". So, in other words, all you worry about is the column in the table corresponding to the divorced subjects. We estimate the conditional probability of having a mild depression given the subject is divorced as $P(\text{mild depression}|\text{divorced}) = \frac{76}{416} \approx 0.183$.

Depression Severity	[.D] Don't know	[1] Married	[2] Widowed	[3] Divorced	[4] Separated	[5] Never married	[6] Living with partner	NA_	Total
[1] No depression	1	1102	57	**174**	12	239	43	1	1629
[2] Mild depression	0	253	17	**76**	4	84	19	1	454
[3] Moderate depression	0	55	1	**27**	3	33	6	0	125
[4] Moderately severe depression	0	13	4	**15**	1	13	6	0	52
[5] Severe depression	0	7	1	**3**	1	2	1	0	15
NA	1	645	33	**121**	20	232	51	3	1106
Total	2	2075	113	**416**	41	603	126	5	3381

9.3   Example: Sensitivity/specificity

Two important examples of conditional probabilities are the so-called sensitivity and specificity. These are particularly useful when discussing the accuracy of screening tests.

The *sensitivity* of a test is the *true positive rate* (or fraction). That is, out of the tests performed on individuals with the disease of interest, how many come out positive. I.e.

sensitivity $= P(\text{test positive} \mid \text{individual diseased})$.

Similarly, the *specificity* of a test is the *true negative rate* (or fraction), i.e. the proportion of tests performed on healthy individuals that come out negative: specificity $= P(\text{test negative} \mid \text{individual healthy})$.

It is also often useful to consider the *false positive rate* (FPR) and *false negative rate* (FNR). These are defined as follows:

$$\text{FPR} = P(\text{test positive} \mid \text{individual healthy}),$$
$$\text{FNR} = P(\text{test negative} \mid \text{individual diseased}).$$

Let's consider a concrete example. Below is table 5-5 from Sullivan (2017). This table shows the results of screenings of 4810 pregnant women to assess if their fetus is likely to have Down Syndrome. After birth, it is determined if the child actually has Down Syndrome, provided a ground truth that we can check our screening method against. Ideally, the test is positive for all kids with Down Syndrome, and negative for alld kids without Down Syndrome.

Let us calculate the specificity, sensitivity, FNR, and FPR:

$$\begin{aligned}
\text{specificity} &= P(\text{test negative} \mid \text{child healthy}) \\
&= \frac{\text{number of negative tests among healthy children}}{\text{number of healthy children}} \\
&= \frac{4449}{4800} = 0.927
\end{aligned}$$

$$\begin{aligned}
\text{sensitivity} &= P(\text{test positive} \mid \text{child has Down Syndrome}) \\
&= \frac{\text{number of positive tests among children with Down Syndrome}}{\text{number of children with Down Syndrome}} \\
&= \frac{9}{10} = 0.9
\end{aligned}$$

$$\begin{aligned}
\text{FPR} &= P(\text{test positive} \mid \text{individual healthy}) \\
&= \frac{\text{number of positive tests among healthy children}}{\text{number of healthy children}} \\
&= \frac{351}{4800} = 0.073
\end{aligned}$$

$$\begin{aligned}
\text{FNR} &= P(\text{test negative} \mid \text{individual diseased}) \\
&= \frac{\text{number of negative tests among children with Down Syndrome}}{\text{number of children with Down Syndrome}} \\
&= \frac{1}{10} = 0.1.
\end{aligned}$$

We see that the test has some very desirable attributes, in high specificity AND high sensitivity. At this point, some might stop and wonder for a second: the end goal is to determine if the test is accurate, so why don't we just calculate the accuracy of the test? I.e. what's wrong at simply looking at the number of corret test results out of the total number of tests? Let's take a look.

$$\text{test accuracy} = \frac{\text{number of correct results}}{\text{number of tests performed}}$$
$$= \frac{9 + 4449}{4810}$$
$$= \frac{4458}{4810} = 0.927$$

That's pretty impressive. The test has an accuracy rate of almost 93%, i.e. almost 93% of tests yield the correct result. Now, let us consider a different test for the same disease. Tested on the same 4810 women, and pretend it yields the following results:

Now, the accuracy rate of this test is $\frac{1+4449}{4810} = 0.925$, i.e. almost the same as the first test. That's, again, really impressive! But upon further investigation, something is off. The sensitivity is way off. Out of 10 children with Down Syndrome, the test only came back positive for 1, which yields a sensitivity of only only 0.1. In other words, if a fetus actually is a affected, the test only has a 10% chance of detecting it. That's not very comforting.

This is a common problem with rare diseases. Since by far most individuals will not be diseased, a test that is good at predicting healthy individuals, but awful at predicting diseased individuals, will have a high accuracy, but such a test is not very desirable. Consider this last test for Down Syndrome: no test is performed, and we just always say the fetus is unaffected. Since 4800 out of 4810 fetuses were unaffected, we have an accuracy of $\frac{4800}{4810} = 0.998$. Pretty impressive accuracy rate, absolutely useless test...

9.4   Example: positive/negative predictive value

The specificity is the answer to the question "what is the probability the test will be correct when the patient is actually healthy?" This is of course a very important thing to know, and if this probability is very low, the test might not be particularly useful. However, a just as important, and sometimes more relevant, measure is the *negative predictive value*. This relates to the question "what is the probability the patient is actually healthy when the test comes back negative?"

Similarly, we can talk about the *positive predictive value*. Where the sensitivity is the probability that the test is positive if the patient has the disease, the positive predictive value is the probability that a patient has the disease if the test comes back positive.

Let us again consider the Down Syndrome data. Since the positive predictive value is the probability a child has Down Syndrome given the test is positive, it is calculated as the proportion of children with positive tests that actually had the disease. So,

$$\text{Positive Predictive Value} = P(\text{child diseased} \mid \text{test positive})$$
$$= \frac{9}{360}$$
$$= 0.025.$$

Similarly, the negative predictive value is the probability a child is healthy given the test was negative. This is calculated as the proportion of children with negative tests that actually are healthy. So,

$$\text{Negative Predictive Value} = P(\text{child healthy} \mid \text{test negative})$$
$$= \frac{4449}{4450}$$
$$= 0.999.$$

9.5 Bayes' Theorem

We have seen a few examples of some very useful and meaningful quantities that are actually conditional probabilities. We've seen how we, in general, calculate these conditional probabilities, but only in a setting where we know everything. The following theorem[1] provides a powerful way of finding conditional probabilities, and it also provides a very useful connection between conditional probabilities, and marginal probabilities (i.e. probabilities that are not conditional).

Theorem 9.1 (Bayes' Theorem). *Bayes' Theorem simply states that*

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

or equivalently

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

Since $P(A \text{ and } B) = P(B \text{ and } A)$, equation (9.1) gives us that $P(B|A)P(A) = P(A \text{ and } B)$, and so equation (9.1) follows by plugging this into the numerator in equation (9.1).

Especially the latter formulation is very powerful, as we shall see in this next example.

9.5.1 Example (5.8 in Sullivan (2017)): positive predictive value from sensitivity

Bayes' Theorem allows us to calculate the positive predictive value using the sensitivity of a test, the prevalence of the disease we're testing for, and how often the test itself is positive (regardless of patient status).

Consider a situation where a disease is really rare with a prevalence of 0.2% (i.e. 2 in 1000 individuals have the disease). A screening test for this disease has a reported sensitivity of 85%, comes back positive 8% of the time, and negative 92% of the time.

We would like to know what the positive predictive value is, i.e. $P(\text{patient has the disease}|\text{screen positive})$. Using Bayes' rule, we know

$$P(\text{patient has the disease}|\text{screen positive}) =$$
$$\frac{P(\text{screen positive}|\text{patient has the disease}) \cdot P(\text{patient has the disease})}{P(\text{screen positive})}$$

---

[1]to those who are not familiar with the math jargon: theorem = very big and important result

Notice: P(screen positive|patient has the disease) is exactly the sensitivity, P(patient has the disease) is the prevalence, and P(screen positive) is the probability the screening test comes back positive. We know all these probabilities, and so we can calculate the positive predictive value.

$$
\begin{aligned}
\text{PPV} &= P(\text{patient has the disease}|\text{screen positive}) \\
&= \frac{P(\text{screen positive}|\text{patient has the disease}) \cdot P(\text{patient has the disease})}{P(\text{screen positive})} \\
&= \frac{0.85 \cdot 0.002}{0.08} \approx 0.021
\end{aligned}
$$

9.6   Independence

One of the big concepts in statistics in general is the concept of independence. When things are independent, all the math simplifies a great deal, which is the main reason why a lot of the methods we will consider later on are based on the assumption that observations are independent of one another.

Loosely speaking, two events are said to be *independent* if knowledge about one of the events does not provide any information about the other. I.e. if I ask you what the probability of event A before and after I tell you whether event B happened or not, your answers should be the same.

9.6.1   Example: independent events

Event A: I walk around Madison one day, stop a random stranger, and ask: "are you taller than 6ft?"

Event B: I flip a coin, and it comes up tails.

Events A and B are independent. The probability that a random person is taller than 6ft is not altered by the fact that a coin flip comes up tails.

9.6.2   Example: dependent events

Event A: I walk around Madison one day, stop a random stranger, and ask: "are you taller than 6ft?"

Event B: The random stranger I stop is male.

Events A and B are NOT independent. The probability a random stranger is taller than 6ft is about 0.16 if the person is male, but less than 0.01 if the person is female.[2] So the probability of event A being 'yes' depends on the outcome of event B. Therefore, they are not independent.

We will work with two definitions of independence. (Fortunately, they are equivalent, i.e. if one holds, the other holds.)

Definition 9.1. Two events are independent if and only if $P(A \text{ and } B) = P(A)P(B)$.

Definition 9.2. Two events are independent if and only if $P(A|B) = P(A)$ **AND** $P(B|A) = P(B)$.

---

[2]loosely based on data from https://dqydj.com/height-percentile-calculator-for-men-and-women/

9.6.3 Example: are "depression severity = mild depression" and "marital status = divorced" independent?

Let's say that A = subject is divorced and B = subject is mildly depressed. For simplicity, we only consider subjects for which we know both marital status and depression severity, i.e. any subjects with missing data in one of the two variables have been removed.

The contingency table:

Depression Severity	Divorced	Not divorced	Total
mild depression	76	377	453
not mild depression	219	1601	1820
Total	295	1978	2273

Now, we can test for independence in two different ways: either using definition 9.1 or definition 9.2. Let's do both.

Using definition 9.1, we need to calculate three probabilities:

$$P(A \text{ and } B) = P(\text{divorced and mildly depressed}) = \frac{76}{2273} \approx 0.0334$$

$$P(A) = P(\text{divorced}) = \frac{295}{2273} \approx 0.1298$$

$$P(B) = P(\text{mild depression}) = \frac{453}{2273} \approx 0.1993$$

Since $P(A) \cdot P(B) \approx 0.1298 \cdot 0.1993 \approx 0.0258691$, which is NOT the same as $P(A \text{ and } B) = 0.0334$, these two events are not independent of each other.

Using definition 9.2, we need to calculate four probabilities. Two of them, $P(A)$ and $P(B)$, we already calculated. Let's calculate the remaining two:

$$P(A|B) = P(\text{divorced}|\text{mildly depressed}) = \frac{76}{453} \approx 0.1678$$

$$P(B|A) = P(\text{mildly depressed}|\text{divorced}) = \frac{76}{295} \approx 0.2576$$

As you can see, $P(A|B) \neq P(A)$ and $P(B|A) \neq P(B)$, so again, the conclusion is that the two events are not independent of one another.

What if we instead let A = subject is male and B = subject is married? The contingency table:

gender	[.D] Don't know	[1] Married	[2] Widowed	[3] Divorced	[4] Separated	[5] Never married	[6] Living with partner	Total
[1] Male	1	928	31	153	8	303	53	1477

gender	[.D] Don't know	[1] Married	[2] Widowed	[3] Divorced	[4] Separated	[5] Never married	[6] Living with partner	Total
[2] Female	1	1147	82	263	33	300	73	1899
Total	2	2075	113	416	41	603	126	3376

Let us just check using definition 9.1:

$$P(A \text{ and } B) = P(\text{male and married}) = \frac{928}{3376} \approx 0.2749$$

$$P(A) = P(\text{male}) = \frac{1477}{3376} \approx 0.4375$$

$$P(B) = P(\text{married}) = \frac{2075}{3376} \approx 0.6146$$

So $P(A) \cdot P(B) \approx 0.2689$, which is not very different from $P(A \text{ and } B)$. Seems reasonable to say that these events indeed are independent.

_____

While definition 9.1 is very useful when doing math, it might not make a whole lot of sense intuitively. Definition 9.2, on the other hand, aligns with our intuitive understanding of independence.

Intuitively, two events, A and B, are independent if A provides no information about B, and vice versa. Now, if A provides no information about B, then what would the difference between $P(B)$ and $P(B|A)$ be? Remember, the latter is basically the probability of B if A happens. But if A and B are independent, A provides no information about B, so the probability of B doesn't change if A happens.

Appendix A

Lecture Slides

Lecture 1 (9/12) slides

Lecture 2 (9/26) slides

References

Bibliography

Liu, K. A. and Mager, N. A. D. (2016). Women's involvement in clinical trials: historical perspective and future implications. *Pharmacy Practice*, 14(1).

Sullivan, L. (2017). *Essentials of Biostatistics in Public Health*. Jones & Bartlett Learning.