Clinical statistics for non-statisticians: Day one

Steve Simon

One warning

- Lots of real world analogies, but
  - May be too specific to U.S.A.
  - Please ask about anything obscure

Speaker notes

Let me start off with a brief warning. I like to draw analogies a lot in my talks to various cultural references, such as books, television shows, or movies. I do this because it enlivens what sometimes can be a tedious topic. But I have to apologize if some of my cultural references may be too specific to the United States. Let me give an example.

There is a series, Ted Lasso, about a football coach, United States football, I mean. He is asked to coach a soccer team in England, soccer being the sport that the rest of the world calls football. Now I know that people outside the U.S.A. watch Ted Lasso. But part of the humor in that series is when Ted starts telling stories that have a point to them, and everyone stares at him in befuddlement because the only people who understand that point Ted is trying to make have been living in the United States their entire lives. As an example, Ted wears a t-shirt on one show that says “Arthur Joes Gates Stack barbecue.” It’s actually a joke that only those of us who live in Kansas City can laugh at. Look it up on the Internet if you are curious.

There’s a stereotype of people in the United States that they think the world revolves around them. It’s a stereotype that is not true for all of us, but it is true for many, including me. I try to fight that tendency, but it’s not always easy.

The point is, that I will try to be inclusive of those of you joining from outside the United States, but if I include a cultural reference that you are unfamiliar with, please don’t hesitate to ask me to explain it. It will be interesting to see what analogies translate to other countries.

Start with a bad joke

Put your reaction ("Ha ha", "Groan", etc.) in the chat box.

Two statistics are sitting in a bar. One turns to the other and asks, "So, how do you like married life?"

The other statistic responds …

Speaker notes

One more thing before I begin anything important. I like to start my talks with a silly joke. It always relates to something I am going to say later.

Now on Zoom, I often miss student reactions. So when I say something funny, I want you to type “Ha ha” or “Smile” or “LMFAO”. The acronym LMFAO means laughing my something … I forget how the rest of it goes.

Now if the joke is corny, like a really really bad pun, it’s okay to put “Groan”. The only thing bad is if I tell a joke and get no reaction at all.

I’ll be sneaking in some jokes throughout the talk and I really want a reaction from you, good or bad. If I don’t get any reaction to a bad pun, your “pun”ishment will be more bad puns.

So here’s the joke. It has been floating around on the Internet for quite a while, and I can’t find the person who gets credit for this. But here goes.

READ JOKE AND FINISH WITH “It’s okay but you lose a degree of freedom.”

Okay, I’m waiting for reactions.

Introduction

- Tell us one interesting number about yourself
- Examples
  - 8: I have traveled to eight countries outside the United States
    - (Canada, Italy, China, France, Russia, England, Holland, and Iceland)
  - 29: I did not learn how to drive until I was 29 years old
  - 1802: My highest chess rating was 1802, but I am not that good any more.

Speaker notes

I want to learn a bit about all of you, and I’m going to do this in a statistical way. Tell me one interesting number about yourself. It could be something simple, like the number of children you have or something exotic like the height of the highest mountain you have climbed.

Here are three numbers about me.

Your turn

Speaker notes

I want each of you to share one, just one, interesting number about yourself.

A bit more about myself

- PhD in Statistics in 1982 from the University of Iowa
- Currently full professor
- Part-time statistical consultant
- Funded on 18 research grants
- Over 100 peer-reviewed publications
- Website with over 2,000 pages
- Many invitations to talk at conferences

Speaker notes

I like to share my background. It’s not because I am conceited, though I am indeed quite conceited. The real reason is to establish what I know and why I am qualified to teach this class.

I have a PhD in Statistics from the University of Iowa. I have always had a strong interest in the computational side of Statistics. My dissertation was 150 pages, and 100 of those pages were computer generated graphs.

I am currently a full professor at the University of Missouri-Kansas City in the Department of Biomedical and Health Informatics. I also do statistical consulting on a part-time basis.

I have been a prolific researcher, receiving support from 18 different grants, and writing over 100 peer-reviewed publications.

I started a website in 1998, writing about data analysis, research ethics, and evidence based medicine. I wrote about two or three pages every week and my site now has over 2,000 pages. It shows the value of persistence.

I love to talk about Statistics and have given many presentations at regional, national, and international conferences. This ranges from short 15 minute talks to day long short courses.

Outline of the three day course

- Day one: Numerical summaries and data visualization
- Day two: Hypothesis testing and sampling
- Day three: Statistical tests to compare treatment to a control and regression models

My goal: help you to become a better consumer of statistics

Speaker notes

This course will cover three days. Here are the topics for each day

Day one topics

- Numerical summaries
  - When should you present the mean versus the median
  - When should you present the range versus standard deviation
  - How should you display percentages
  - Why should you round liberally

Speaker notes

Today, you will learn about numerical summaries.

Day one topics (continued)

- Data visualization
  - How should you display continuous data
  - Why is the normal bell-shaped curve important
  - How should you display categorical data
  - How do you illustrate trends and patterns
  - What are some common mistakes in the choice of colors

Speaker notes

Speaker notes

You will also learn a little bit about data visualization.

Quiz question 1

```
          No            Yes          Total
Female   154 (33%)    308 (67%)    462 (100%)
Male     709 (83%)    142 (17%)    851 (100%)
Total    863 (66%)    450 (34%)   1313 (100%)
```

This data table shows counts and …

1. cell percents
2. column percents
3. row percents
4. I do not know the answer

Speaker notes

Here is a question about the percentages shown in a table. If you do not know the answer, that’s okay. This is something you will learn about in this lecture and you should be able to answer correctly at the end of the class.

Quiz question 2

The median might be preferred to the mean if

1. a single extreme value distorts the mean
2. the data follows a bell shaped curve
3. there is very little variation in the data
4. you have a biased sample
5. I do not know the answer

Speaker notes

Here is a question about the median.

Quiz question 3

The problem with error bars is that they

1. fail to show if the data is skewed
2. have several competing definitions
3. use only two numbers to characterize your data
4. all of the above are correct
5. none of the above are correct
6. I do not know the answer

Speaker notes

Speaker notes

Here is a question about error bars

Counting and proportions

- Counts are the most common statistic
  - Counts are error prone
  - Counts require a solid operational definition

Speaker notes

Speaker notes

Let’s start with the simplest statistic of all a simple count. This is a very common statistic.

But counts can be tricky. The counting process is error prone and requires a solid operational definition.

Student exercise

Count the number of occurrences of the letter “e”.

```
A quality control  program is easiest
to implement from the top down.
Make sure that you understand the
the commitment of time and money
that is involved. Every workplace is
different, but think about allocating
10% of your time and 10% of the
time of all your employees to
quality control.
```

Speaker notes

Speaker notes

Here’s an exercise I want you to do. Just count the number of occurrences of the letter “e”. Once you have your answer, type it in the chat box.

PAUSE HERE.

The numbers are different because of two things. First, it is easy to make mistakes. Did anyone notice the repetition of the word “the” at the end of the third line and the beginning of the fourth. It would be easy to miss that and count one less “e”.

What did you do with the first e in “Every”?

Did you count the e’s in the quotes itself or also on the slide instructions and the slide header?

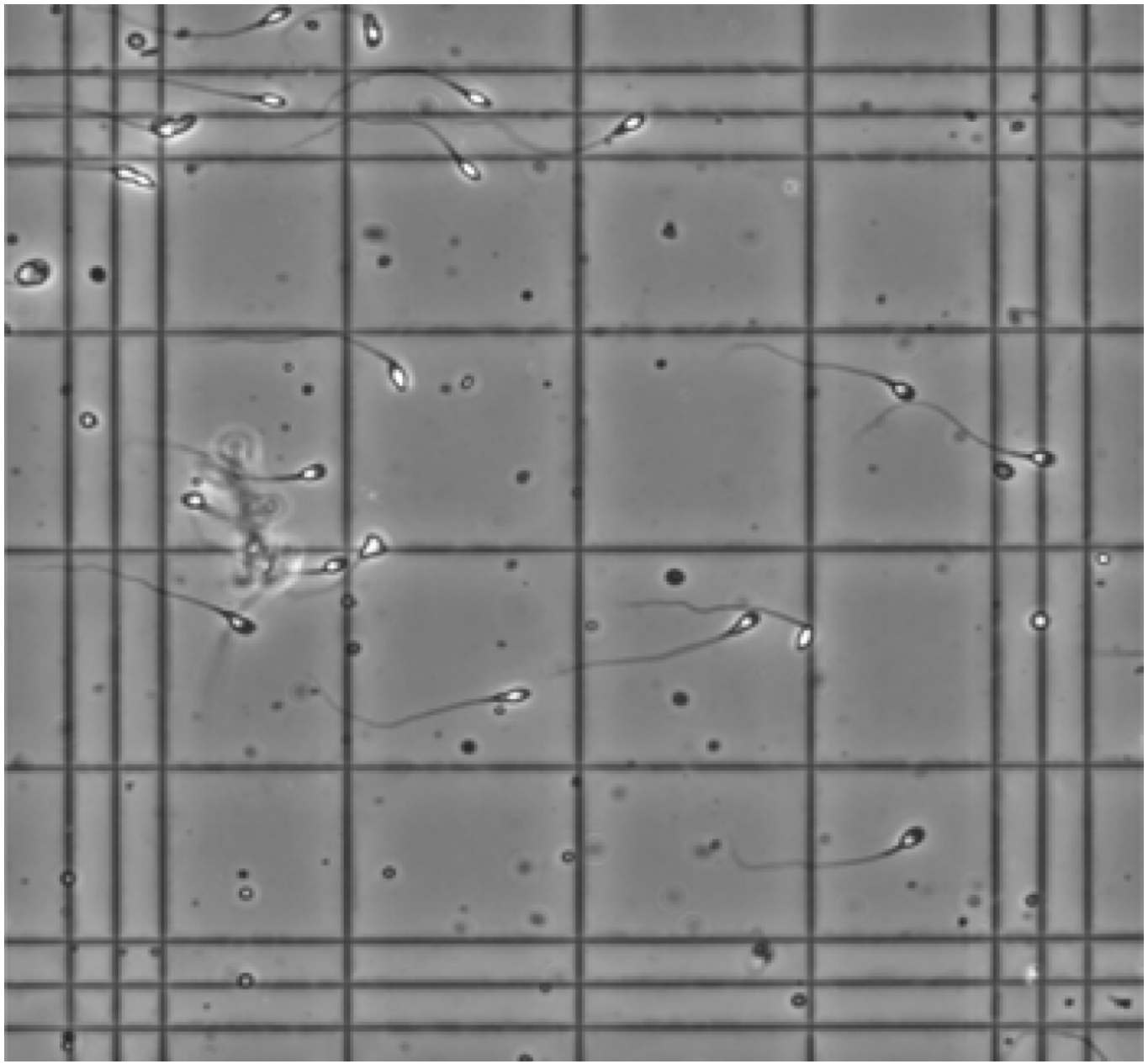


Figure 1: Image of a haemocytometer

Speaker notes

This image is take from the WHO laboratory manual for the examination and processing of human semen, published in 2021. It shows a haemocytometer, an instrument used for counting the number of cells. To get a proper count, you need to include any cells inside the four by four grid of large squares in the middle of this micrograph. But what does “inside” mean? Should you count only those cells entirely inside the four by four grid. Or should you include cells that are partially inside the grid?

One rule is to count cells if the head of the sperm cell touches the top or right side of a square, but not if it touches the bottom or left side of the square. And don’t count a sperm cell if only the tail is inside the square.

That’s not the only way you can do this, but just make sure that whatever convention you use for deciding “inside” versus “outside” is consistent across your laboratory.

Sex * Survived Crosstabulation

Count

		Survived		
		No	Yes	Total
Sex	female	154	308	462
	male	709	142	851
Total		863	450	1313

Figure 2: Titanic data: counts of survival by gender

Speaker notes

I have some data taken from the passenger ship Titanic. I have used this data in many talks and on many of my web pages. But in the middle of preparing this talk, a tragedy occurred. Titan, a 22 foot submarine was taking five people to view the wreckage of the Titanic last week when it had a serious accident. I can't talk about this data set without first acknowledging this tragedy. My heart goes out to the families and friends of those who lost their lives.

Now, I have a tendency to talk a bit flippantly about things, but I hope that none of my comments are interpreted as insensitive towards this recent tragedy. I had thought about changing this example, but it appears too many times in the talk to allow me to substitute another interesting data set.

The five people who died last week joined 863 who died on the Titanic back in 1912.

The Titanic was an enormous ship. It was bigger than any passenger ship ever built at the time. It was so large that they thought it was unsinkable. But in its first voyage across the Atlantic Ocean, it struck an iceberg and sunk.

They kept records on everyone on the ship: sex, age, and passenger class. There were 462 women on the ship. 308 of them survived, including Kate Winslet. The men did not fare as well. This was in a time when they really believed in the saying "Women and children first".

Among the 851 men, 709 died, including, sadly, Leonardo Di Caprio.

I'm making a reference to a popular movie, "Titanic" that was released in 1997. Has anyone seen that movie?

Now these numbers are just the passengers. Many of the crew died as well.

Anyway, you might want to examine mortality trends more closely by computing percentages. But there are three different ways you could compute these percentages.

Sex * Survived Crosstabulation

			Survived		
			No	Yes	Total
Sex	female	Count	154	308	462
		% within Survived	17.8%	68.4%	35.2%
	male	Count	709	142	851
		% within Survived	82.2%	31.6%	64.8%
Total		Count	863	450	1313
		% within Survived	100.0%	100.0%	100.0%

Figure 3: Titanic data with column percentages

Speaker notes

Here are the percentages computed by dividing by the column totals. Divide the 308 surviving females by the total number of survivors, 450, to get 68%. Divide the 142 surviving males by 450 to get 32%. So those lifeboats were mostly, but not entirely, filled with women.

These are called column percents. They add up to 100% within each column: 18% + 82% = 100% and 68% + 32% = 100%.

Sex * Survived Crosstabulation

			Survived		
			No	Yes	Total
Sex	female	Count	154	308	462
		% within Sex	33.3%	66.7%	100.0%
	male	Count	709	142	851
		% within Sex	83.3%	16.7%	100.0%
Total		Count	863	450	1313
		% within Sex	65.7%	34.3%	100.0%

Figure 4: Titanic data with row percentages

Speaker notes

You could also divide by the row totals. Divide the 308 surviving women by the total number of women, 462, to get a survival rate of 67%. Divide the 142 surviving men by the total number of men, 851, to get 17%.

17%! This shows how poorly the men fared on the Titanic. If you were female, you might have died, but more likely than not you DID survive. For the men, not such good news. Most of them died. Only a small fraction survived.

This is called the row percentages. These percentages add up to 100 within each row: 33% + 67% = 100% and 83% + 17% = 100%.

Percentages divided by grand total

Sex * Survived Crosstabulation

			Survived		
			No	Yes	Total
Sex	female	Count	154	308	462
		% of Total	11.7%	23.5%	35.2%
	male	Count	709	142	851
		% of Total	54.0%	10.8%	64.8%
Total		Count	863	450	1313
		% of Total	65.7%	34.3%	100.0%

Figure 5: Titanic data with cell percentages

Speaker notes

You could also divide all the numbers by the grand total of 1,313. The 308 female survivors represented a bit less than 24% of all the passengers that set sail from England.

The 142 male survivors represented a bit less than 11% of all the survivors.

These are called the cell percentages. They add up to 100% across the entire table: 12% + 54% + 24% + 11% = 101%. Close enough!

Which makes the most sense? It depends on your perspective. If you want to test the hypothesis that male passengers on the Titanic had a much smaller risk of dying, then the row percentages make the most sense.

But from the perspective of the Carpathia, the ship that rescued the survivors, the column percents make the most sense. They had to make room on their ship for 450 passengers, 68% who were female and 32% who were male. I bet that the lines for the women’s bathrooms on the Carpathia were really long.

My recommendations

- Treatment or exposure as rows
- Outcome as columns
- Usually report row percentages
  - Female survival rate: 67%
  - Male survival rate: 17%
- But sometimes column percentages
  - Survivors: 68% female, 32% male

Speaker notes

Speaker notes

These two by two tables occur a lot in Statistics. I have some general guidelines that I use with them. They don’t always work, but they work most of the time.

If you have a variable that represents a treatment or exposure, try using that as the rows of the table. If you have a variable that represents an outcome, try using that as the columns of the table. Sometimes, there are no clearly identified treatment variables and no clearly identified outcome variables. But try to categorize them this way, if you can.

With a table lined up with the treatments as the rows and the outcomes are the variables, calculate the row percentages.

In the Titanic data, survival is clearly an outcome. So arrange the table like I did with sex as the rows and survival as the columns and compare the two survival rates: a healthy 67% for females and a feeble 17% for males.

But sometimes you will find that the column percents make more sense. It does depend on what question you are trying to answer with the data.

Some rationale for these choices

My way

		Survived	
		No	Yes
Sex	Female	33% (154)	67% (308)
	Male	83% (863)	17% (142)

Not my way

		Sex	
		Female	Male
Survived	No	33% (154)	83% (863)
	Yes	67% (308)	17% (142)

Speaker notes

Speaker notes

Now, I believe it is important to think carefully about which is your rows and which is your columns. Here’s the layout that I recommend on the left and the layout that I don’t recommend on the right. The only difference is that the second table is transposed from the first. What was the rows becomes the columns and what was the columns becomes the rows.

The key comparison is among survival rates, 67% for females and only 17% for males. When you orient my way with the treatment/exposure (Sex) as rows and the outcome (Survived) as the columns, the numbers 67% and 17% are very close to one another. In the alternate layout the numbers you are most interested in comparing are not as close together.

Now this is not an absolute rule. Sometimes I’ll switch things up. But about 90% of the time, I find that the layout with the treatment or exposure as the rows and the outcome as the columns, the table just looks better.

Break

- What have you just learned?
  - Displaying percentages
- What is coming next?
  - Practice exercise
  - Calculation of the mean and median

Speaker notes

Let me pause here for a second. Are there any questions?

On your own

Calculate row and column percentages for the following tables. Interpret your results.

PClass * Survived Crosstabulation

Count

		Survived		
		No	Yes	Total
PClass	1st	129	193	322
	2nd	161	119	280
	3rd	573	138	711
Total		863	450	1313

Figure 6: Titanic passenger class counts

Child * Survived Crosstabulation

Count

		Survived		
		No	Yes	Total
Child	No	386	244	630
	Yes	57	69	126
Total		443	313	756

Figure 7: Titanic child counts

Speaker notes

Speaker notes

Now try to report both column and row percents for one of these two tables. Breakout room #1 work on the passenger class table and breakout room #2 work on the child data.

Put your percentages in a table using a word processing program or text editor so you can share your results with the group.

Be sure to interpret these numbers. Come back together again in about 10 minutes. I'll bounce back and forth between the two rooms to see if you have any questions.

Figure 8: Cartoon image of Professor Mean

Speaker notes

Speaker notes

Here’s a cartoon image of Professor Mean. I know this looks like it was drawn by a professional artist, but it was actually drawn by me. Really!

Professor Mean is my alter ego on the Internet. For those who don’t get the inside joke, I point out that Professor Mean is not just your average professor.

I will use the terms mean and average interchangeably througout this talk.

Figure 9: Road with a median strip

Speaker notes

Speaker notes

This is an image of a traffic median. This is a strip of land, typically raised from the road surface, that splits the road in half.

In Statistics, the median is the data value that splits the data in half. Half of the data is smaller than the median and half of the data is larger than the median.

Calculation of the mean and median

- Mean
  - Add up all the values, divide by the sample size
- Median
  - Sort the data
    - Select the middle value if n is odd
    - go halfway between the two middle values if n is even

Speaker notes

Speaker notes

You already know how to compute the average. Add up all the values and divide by the sample size.

The median is also simple. Sort the data and choose the “middle” value. If n is odd, there is one value that is right in the middle. With five data values, the median is the third value of the sorted list. The first and second values are smaller and the fourth and fifth values are larger.

With an even number, there are two middle values. Go halfway between them. If you have eight data values, the midpoint between the fourth and fifth values splits the data in half. The first through fourth values in the sorted list are smaller and the fifth through eighth values are larger.

Formal mathematical definitions

- Mean
  - $\bar{X} = \frac{1}{n}\Sigma X_i$
- Median
  - Sorted values $X_{[1]}, X_{[2]}, \ldots, X_{[n]}$
    - $X_{[(n+1)/2]}$ if n is odd,
    - $(X_{[n/2]} + X_{[n/2+1]})/2$ if n is even

Speaker notes

Here are the mathematical formulas for the mean and median. I know some people hate formulas, but I love them. With a few symbols and Greek letters, you can express really deep and beautiful ideas. Well these formulas aren’t all that deep.

Bacteria before and after A/C upgrade

Room	Before	After
121	11.8	10.1
125	7.1	3.8
163	8.2	7.2
218	10.1	10.5
233	10.8	8.3
264	14	12
324	14.6	12.1
325	14	13.7

Speaker notes

Here is some data that I got off the web.

https://dasl.datadescription.com/datafile/legionnaires-disease/

This represents bacteria counts before and after a new air conditioning unit was installed in a small hotel.

I want to illustrate the calculation of the mean and median.

Before remediation mean

11.8 + 7.1 + 8.2 + 10.1 + 10.8 + 14 + 14.6 + 14 = 90.6

90.6 / 8 = 11.325

Round to 11.3

Speaker notes

Here’s the data for bacterial counts before remediation. If you add the eight values up, you get 90.6. Divide this by eight to get 11.325. Always round liberally when you are talking about the mean.

After remediation mean

10.1 + 3.8 + 7.2 + 10.5 + 8.3 + 12 + 12.1 + 13.7 = 77.7

77.7 / 8 = 9.7125

Round to 9.7

Speaker notes

Here are the same calculations for the bacterial counts after remediation.

Before remediation median (1/4)

```
121  11.8

125   7.1

163   8.2

218  10.1

233  10.8

264  14.0

324  14.6

325  14.0
```

Speaker notes

Speaker notes

Now for the median.

Here is the data for bacteria counts before remediation. Notice that the data is arranged by room number.

Before remediation median (2/4)

```
125   7.1

163   8.2

218  10.1

233  10.8

121  11.8

264  14.0

325  14.0

324  14.6
```

Speaker notes

The first thing you do is sort the data from the lowest bacteria count to the highest bacteria count.

The data was arranged by room number, but now it is arranged by bacterial count. The smallest bacteria count is listed at the top and the largest is listed at the bottom.

Before remediation median (3/4)

```
125   7.1

163   8.2

218  10.1

233  10.8  10.8

121  11.8  11.8

264  14.0

325  14.0

324  14.6
```

Speaker notes

Speaker notes

Then pick out the middle value. If you have an even number of data points, there will be two middle values.

In this data set, the two middle values are the fourth and fifth largest values out of eight.

Before remediation median (4/4)

```
125   7.1

163   8.2

218  10.1

233  10.8  10.8
                   (10.8 + 11.8) / 2 = 11.3
121  11.8  11.8

264  14.0

325  14.0

324  14.6
```

Speaker notes

If there are two middle values, just average them.

After remediation median (1/4)

```
121  10.1

125   3.8

163   7.2

218  10.5

233   8.3

264  12.0

324  12.1

325  13.7
```

Speaker notes

Speaker notes

Here is the data for bacteria counts after remediation.

After remediation median (2/4)

```
125   3.8

163   7.2

233   8.3

121  10.1

218  10.5

264  12.0

324  12.1

325  13.7
```

Speaker notes

Just like before, you sort the data.

After remediation median (3/4)

```
125   3.8

163   7.2

233   8.3

121  10.1  10.1

218  10.5  10.5

264  12.0

324  12.1

325  13.7
```

Speaker notes

Then pick out the middle value. Here again, there are two middle values.

After remediation median (4/4)

```
125   3.8

163   7.2

233   8.3

121  10.1  10.1
                    (10.1 + 10.5) / 2 = 10.3
218  10.5  10.5

264  12.0

324  12.1

325  13.7
```

Speaker notes

Just average the two middle values.

Break

- What have you just learned?
  - Calculation of the mean and median
- What is coming next?
  - Criticisms of the mean and median

Speaker notes

Let me pause here for a second. Are there any questions?

Criticisms of the mean and median

- Are you combining apples and onions?
- Are you ignoring minorities?

Speaker notes

There’s a wonderful cartoon by Dana Fradon that appeared in The New Yorker in 1976. She shows a road going into town and the sign by the side of the road reads “Hillsdale, Founded 1802, Altitude 600, Population 3,700. Total 6,122.” You can’t add these things together.

It’s similar for means. There was a dataset showing housing prices for homes in Boston and none of the analyses seemed to make sense. The problem in Boston is that a small number of the houses had prices that were out of sync with their other homes. These were historical houses, such as Paul Revere’s house.

When you are averaging numbers, maybe it’s okay to have a few oranges in with the apples. A mix of apples and oranges is just fruit salad. You shouldn’t have a problem with that.

When it becomes a problem is when the data are so diverse that it becomes a mix of apples and onions. There are lots of great recipes that mix apples and oranges, but none that mix apples and onions.

The other problem is that an average may be a reasonable number to represent the majority of patients in your sample, but it may masks some important trends that appear in a minority.

This is a big problem in a larger context than just the mean or median. There are some very fancy high tech prediction models that work very well for most people and the statistics like the mean and median back this up quite nicely. But the prediction models perform terribly for minority groups. Something that does well for the average person may not be so great for a large segment of society.

Use of the mean for ordinal data

- Stevens scales of measurement (controversial!)
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- Addition/subtraction not allowed for ordinal data
  - Mean of ordinal data is meaningless

Speaker notes

A psychologist, Stanley Smith Stevens divided the entire universe of data into four categories: nominal, ordinal, interval, ratio. I won’t review the definitions for all of these, but ordinal data is categorical data where there is a natural ordering of categories. An important limitation to ordinal data, but where the spacing between successive units is not consistent.

The belief among many (but not all) researchers, is that

An example of ordinal data.

- “Do you agree or disagree with the following statements”
  - “I believe that knowledge of Statistics is important for my job.”
    - 1 = Strongly disagree,
    - 2 = Disagree
    - 3 = Neutral
    - 4 = Agree
    - 5 = Strongly agree

Speaker notes

Speaker notes

An example of ordinal data is the Likert scale. This takes various forms, but often it is used with group of questions on a questionnaire that reads something like

“Do you agree or disagree with the following statements”

You are asked to respond 1=Strongly disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly agree.

Now I’m sure everyone today is going to choose 5. But assigning numbers 1, 2, 3, 4, and 5 to categories of strongly disagree, disagree, neutral, agree, and strongly agree may falsely imply that a jump from 3 (neutral) to 4 (agree) is about the same amount of improvement as a jump from 4 (agree) to 5 (strongly agree). That’s probably not the case.

You can’t really average ordinal data, some people say because that implies that two responses of “Agree” are the same as one response of “Neutral” along with a response of “Strongly agree”.

Do you want everyone to be at least somewhat on your side or do you want to have a smaller number of very enthusiastic supporters.

If you believe that two 4’s are not the same as a 3 and a 5, then you can’t average.

Now I beg to disagree here, but I am part of a minority opinion. I think that if at the start of this class, your average rating was 3.2 and after I finish the lecture, your average rating climbs to 4.4, that I have done my job well.

If it only jumps to 3.6, then I have still done well, but not as much as that jump to 4.4.

Another example of ordinal data, course grades

- A = 4
- B = 3
- C = 2
- D = 1
- F = 0

Speaker notes

Speaker notes

Another example of ordinal data is grades assigned to students. Now everyone in this class is getting an A, but in other classes I teach I might assign different grades. You can attach a number to each of these grades, 4 for A, 3 for B, 2 for C, 1 for D, and 0 for F.

These numbers seem to imply that a student with two B's is as smart as a student with an A and a C.

It raises an interesting story. A colleague of mine told me that he would never hire anyone with a single F on their transcript. An F is a red flag, he felt. So he would not want to assign a value of 0 to F, because that implies that the difference between an F and a D is equivalent to the difference between a B and and A. He's want to assign a value like negative one million to an F so that the average would be pulled way down for a single F, no matter what the other grades would be.

Now I would never be so harsh, but there is really nothing wrong with his perspective. And I would certainly treat a student with three A's and one F differently from a student with two A's and two C's even though mathematically, both average out to 3.0.

Now, in spite of all the obvious problems with equivalence between different grades, most of us still accept a grade point average as a meaningful indicator of how well a student did in school.

Phoenix5 To help men and their companions overcome the effects of prostate cancer

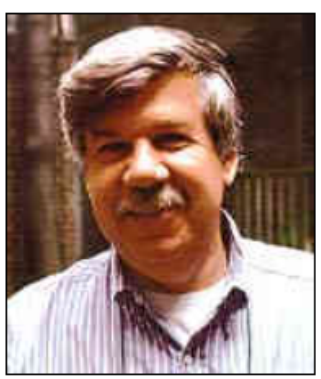
main menu - articles - prostate - stories - sexuality - resources - glossary - search

from
The Articles Menu

"This is a personal story of statistics..."

THE MEDIAN ISN'T THE MESSAGE

by Stephen Jay Gould



Born in 1941, Stephen Jay Gould was a geologist, zoologist, paleontologist and evolutionary biologist at Harvard. He was also one of the most noted, prolific and best-selling scientific writers of our day. He was diagnosed in 1982 with abdominal mesothelioma, a rare and very deadly form of cancer associated with exposure to asbestos. This is his story. It was first published in Discover magazine in June 1985 and was reprinted here at Phoenix5 with his kind permission. He beat the cancer for 20 years, finally passing on May 20, 2002, giving all of us a valuable lesson in beating the odds.

Figure 10: Excerpt from Gould 1985 publication

Speaker notes

Stephen Jay Gould was a famous Evolutionary Biologist. He was a prolific writer with 20 books and 300 essays. Much of his writing was for academic researchers, but just as much was for the general public.

One of his most famous essays was “The Median Isn’t the Message”. The title is a take-off of a quote by Marshall McLuhan, “The medium is the message” which itself has an interesting history that you should investigate on your own.

The Gould essay was written in 1985 for Discover Magazine. It has been reprinted many times, and you can easily find the full text with a simple Google search.

The image shown here is taken from phoenix5.org, an informational site for patients with prostate cancer.

Gould was diagnosed with a rare cancer, abdominal mesothelioma, with a very poor prognosis. Such a poor prognosis that Gould was actively discouraged by his physician from looking at any peer reviewed research about his cancer.

But Gould looked anyway. “Of course, trying to keep an intellectual away from literature works about as well as recommending chastity to Homo sapiens, the sexiest primate of all.”

But he found that the doctor had good reason to discourage this trip to the medical library.

“The literature couldn’t have been more brutally clear: Mesothelioma is incurable, with a median mortality of only eight months after discovery.”

Gould was momentarily distressed, but then he thought carefully about the problem.

“When I learned about the eight-month median, my first intellectual reaction was: Fine, half the people will live longer; now what are my chances of being in that half? I read for a furious and nervous hour and concluded, with relief: damned good. I possessed every one of the characteristics conferring a probability of longer life: I was young; my disease had been recognized in a relatively early stage; I would receive the nation’s best medical treatment; I had the world to live for; I knew how to read the data properly and not despair.”

He goes on to find a bit more reason for optimism.

“Another technical point then added even more solace. I immediately recognized that the distribution of variation about the eight-month median would almost surely be what statisticians call “right skewed.” (In a symmetrical distribution, the profile of variation to the left of the central tendency is a mirror image of variation to the right. Skewed distributions are asymmetrical, with variation stretching out more in one

direction than the other—left skewed if extended to the left, right skewed if stretched out to the right.) The distribution of variation had to be right skewed, I reasoned. After all, the left of the distribution contains an irrevocable lower boundary of zero (since mesothelioma can only be identified at death or before). Thus, little space exists for the distribution's lower (or left) half—it must be scrunched up between zero and eight months. But the upper (or right) half can extend out for years and years, even if nobody ultimately survives. The distribution must be right skewed, and I needed to know how long the extended tail ran—for I had already concluded that my favorable profile made me a good candidate for the right half of the curve."

Gould did indeed find himself on the happy side of the eight month median, a good 20 years beyond the median.

The median isn't the message. It is a single number with half the people on the lower side and half on the higher side. Don't think for a minute that single number like a median can characterize everyone in a group.

Choosing between the mean and median

Speaker notes

While there is some consensus on when to use the mean versus the median, the choice is not always obvious. Controversies often arise over this issue.

Here are some general guidelines.

Most of the time, either the mean or the median is fine.

One big advantage of the mean is that it allows extrapolation to totals. This is often important in the analysis of the economic effects of illness.

I found this data on the web site statista.com. The average cost per patient with at least one chronic disease was 696 euros. If you wanted to extrapolate this average and get a total cost for the whole country, multiply by the number of people in Italy times the proportion who have one or more chronic diseases.

The other issue is outliers. Extreme values tend to pull the mean towards that value.

We have this guy living in the United States named Elon Musk. My wife idolizes him. She bought a Tesla from his company and brags about it to all her friends. She's a big fan of his space exploration efforts and is fascinated by a possible manned flight to Mars.

Me, I think he is just a rich jerk. But suppose you are computing average net worth of a random sample of individuals and by your good luck (my wife's perspective) or bad luck (my perspective) Elon Musk gets to be part of your sample. The average net worth approaches a billion dollars because all the money that Musk has swamps the total. No one else in the sample has a net worth anywhere near a billion dollars, so the mean is not a fair reflection of the average person in the sample. The median net worth doesn't change if Musk's net worth is 400 billion dollars, before he bought Twitter or 200 billion after he bought Twitter.

Now the Elon Musk example is silly, but the issue of outliers having an effect on the mean is important in many applications.

Published in final edited form as:

Value Health. 2019 December ; 22(12): 1387–1395. doi:10.1016/j.jval.2019.08.005.

Trends in the Price per Median and Mean Life-Year Gained Among Newly Approved Cancer Therapies 1995 to 2017

Alice J. Chen, PhD[1,2,*], **Xiaohan Hu, MPH**[2], **Rena M. Conti, PhD**[3], **Anupam B. Jena, MD, PhD**[4,5], **Dana P. Goldman, PhD**[1,2,5]

Figure 11: Chen et al 2019

Speaker notes

Speaker notes

Here is an article I found on PubMed, one of my favorite websites that compares median and mean improvements in life expectancy in cancer patients.

Chen 2019, PMID: 31806195 (continued)

Background: The prices of newly approved cancer drugs have risen over the past decades. **A key policy question is whether the clinical gains offered by these drugs in treating specific cancer indications justify the price increases.**

Speaker notes

Here’s part of the abstract.

The United States is like a lot of first world countries in that we spend more and more money each year on cancer treatments. Are we getting our money’s worth?

Chen 2019, PMID: 31806195 (continued)

Results: We found that between 1995 and 2012, price increases outstripped median survival gains, a finding consistent with previous literature. **Nevertheless, price per mean life-year gained increased at a considerably slower rate, suggesting that new drugs have been more effective in achieving longer-term survival.** Between 2013 and 2017, price increases reflected equally large gains in median and mean survival, resulting in a flat profile for benefit-adjusted launch prices in recent years.

Speaker notes

Later on in the abstract, the authors point out that from the persective of the median, things are bleak. The median survival gains are not in line with the increasing amount of money spent on new treatments. But the mean survival gains show a different story. A flat profile means that increases in price are accompanied by an increase in benefits in terms of gains in the mean. What this implies is that the extreme tail of the distribution includes a number of Elon Musk types. A small number of people are showing amazingly big gains in survival, justifying the increase in cost.

Break

- What have you just learned?
  - Criticisms of the mean and median
- What is coming next?
  - Computing percentiles

Speaker notes

Let me pause here for a second. Are there any questions?

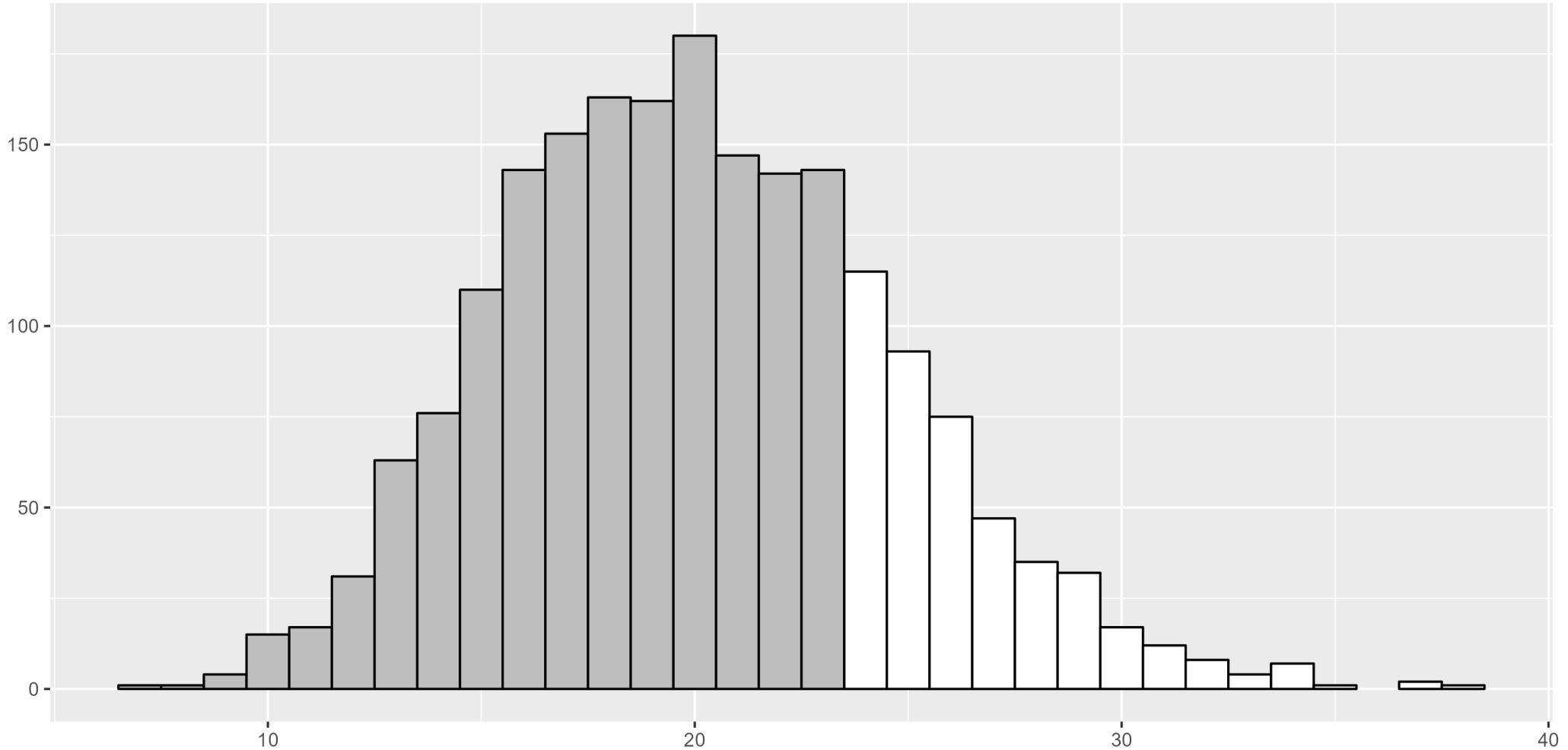


Figure 12: Illustration of the 75th percentile

Speaker notes

Speaker notes

I want to mention percentiles briefly. A percentile is a value that splits the data so that a certain percentage is smaller and a certain percentage is larger.

The 75th percentile, for example will be above 75% of the data and below 25% of the data. This graph illustrates the 75th percentile for some arbitrary data. The gray bars represent about 75% of the data and the white bars represent about 25% of the data.

I use a few weasel words like “roughly” and “about” because you can’t always get a perfect split. But you can usually come close.

Computing percentiles

- Many formulas
  - Differences are not worth fighting over
- My preference (pth quantile)
  - Sort the data
  - Calculate p*(n+1)
  - Is it a whole number?
    - Yes: Select that value, otherwise
    - No: Go halfway between
    - Special cases: p(n+1) < 1 or > n

Speaker notes

There are close to a dozen different ways to compute a percentile, but the differences between the values selected are small and not worth fussing about.

Here is my preference for choosing the pth quantile (remember that for quantiles, you range between 0 and 1, not between 0 and 100).

Calculate the quantity p*(n+1). If that value is a whole number, great! You just select that value. If it is a fractional value, round up and down and go halfway between.

Once in a while, you'll get an extreme case, where p(n+1) is less than 1 or greater than n. Just use a bit of common sense.

If you have nine values and p(n+1) is 9.2, you can't go halfway between the 9th and 10th observations. There is no 10th observation. So just choose the 9th or largest value.

Likewise if p(n+1) is 0.8, you can't go halfway between the zeroth and first observation. There is no zeroth observation. Just choose the first or smallest value.

Some examples of percentile calculations

- Example for n=39
  - For 5th percentile, p(n+1)=2 -> 2nd smallest value
  - For 4th percentile, p(n+1)=1.6 -> halfway between two smallest values
  - For 2nd percentile, p(n+1)=0.8 -> smallest value

Speaker notes

Suppose you have 39 observations. For the 5th percentile or the 0.05 quantile, p(n+1) equals 2. Lucky you. The second smallest observation is the 5th percentile. For the 4th percentile or the 0.04 quantile, you get p(n+1) equal to 1.6. Go halfway between 1, the smallest value, and 2, the second smallest value.

The 2nd percentile represents one of the special cases. You calculate p(n+1) and get 0.8. You can’t go halfway between 0 and 1, so just choose the smallest value.

Some terminology

- Percentile: goes from 0% to 100%
- Quantile: goes from 0.0 to 1.0
  - 90th percentile = 0.9 quantile
- 25th, 50th, and 75th percentiles: quartiles
  - 25th percentile: Q_1, $X_{0.25}$ or lower quartile
  - Median/50th percentiles: Q_2 or $X_{0.5}$
  - 75th percentile: Q_3, $X_{0.75}$ or upper quartile

Speaker notes

Speaker notes

A percentile always refers to a percentage. So it has to be between 0% and 100%. Sometimes, you may see references to a quantile. A quantile is a percentile, but is expressed as a proportion rather than a percent. A quantile goes from 0.0 to 1.0. The 90th percentile and the 0.90 quantile are the same thing.

You might see the term “quartiles”. These are the 25th, 50th, and 75th percentiles. These three values split the data into quarters.

If you see “lower quartile”, it means the 25th percentile. Likewise, “upper quartile” means the 75th percentile.

Let me be try to be careful about terminology here. But, sometimes I will mess up and use “percentile” when I mean “quantile”.

Before remediation upper quartile (1/4)

121	11.8
125	7.1
163	8.2
218	10.1
233	10.8
264	14.0
324	14.6
325	14.0

Speaker notes

Here is the data for bacteria counts before remediation. Let’s calculate the upper quartile, also known as the 0.75 quantile or the 75th percentile.

Before remediation upper quartile (2/4)

125	7.1
163	8.2
218	10.1
233	10.8
121	11.8
264	14.0
325	14.0
324	14.6

Speaker notes

Just like before, you sort the data.

Before remediation upper quartile (3/4)

125	7.1	
163	8.2	
218	10.1	
233	10.8	
121	11.8	
264	14.0	14
325	14.0	14
324	14.6	

Speaker notes

With n=8, you get p(n+1) = 6.75. So pick out the sixth and seventh values.

Before remediation upper quartile (4/4)

```
125   7.1

163   8.2

218  10.1

233  10.8

121  11.8

264  14.0  14
                   (14 + 14) / 2 = 14
325  14.0  14

324  14.6
```

Speaker notes

Speaker notes

Go halfway between these values.

After remediation upper quartile (1/4)

```
121  10.1

125   3.8

163   7.2

218  10.5

233   8.3

264  12.0

324  12.1

325  13.7
```

Speaker notes

Here are the same calculations the upper quartile of bacteria counts after remediation.

After remediation upper quartile (2/4)

```
125   3.8

163   7.2

233   8.3

121  10.1

218  10.5

264  12.0

324  12.1

325  13.7
```

Speaker notes

Just like before, you sort the data.

After remediation upper quartile (3/4)

```
125   3.8

163   7.2

233   8.3

121  10.1

218  10.5

264  12.0  12

324  12.1  12.1

325  13.7
```

Speaker notes

Speaker notes

Then pick out the sixth and seventh values.

After remediation upper quartile (4/4)

```
125   3.8

163   7.2

233   8.3

121  10.1

218  10.5

264  12.0  12
                   (12 + 12.1) / 2 = 12.05
324  12.1  12.1

325  13.7
```

Speaker notes

Speaker notes

Go halfway between these values.

When you should use percentiles

- Characterize variation
  - Middle 50% of the data
- Exposure issues
  - Not enough to control median exposure level
- Quantify extremes
  - What does “upper class” mean?
- Quality control
  - Almost all products must meet a minimum standard

Speaker notes

There are many reasons why you might be interested in percentiles rather than the mean or median. Actually, the median is a percentile, the 50th percentile, but I want to talk about percentiles other than 50%.

One important use of percentiles is looking at the middle 50% of the data. This is the data between the lower quartile (25th percentile) and the upper quartile (75th percentile). Is the middle 50% of the data bunched tightly together or spread widely apart?

Percentiles are also important in the study of exposures. If you work in an environment where the median worker has a safe level of exposure, you could easily end up with 20%, 30% or more of the workers dying from unsafe exposures. It is important to insure that not just the median, but a very high percentile like the 99th percentile of exposure levels is at a safe level.

Percentiles also help to define extreme groups. You can, for example, define the term upper class as anyone earning more than the 90th percentile of income.

Percentiles also can help with quality control. If you make a claim about a product, you want to make sure that that claim is not valid at a median level but at a much higher level. You don’t sell 500 mg bottles of liquid Tylenol is your factory is churning out a median fill level of 500 mg. Half of your customers would be cheated. Instead you insure that the 98th percentile coming out of the factory floor is at least 500 mg. You lose a bit of money because most bottles contain more than 500 mg, but the cost of an irate customer is worth more than the cost of 50 overfilled bottles.

Break

- What have you just learned?
  - Computing percentiles
- What is coming next?
  - Computing the standard deviation

Speaker notes

Let me pause here for a second. Are there any questions?

Standard deviation

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

At least one alternative formulas.

Speaker notes

The standard deviation is a commonly used measure of how spread out the data is. The formula is a bit messy, but if you look carefully at it, you will see that it is a measure of how far each individual value is from the overall mean.

Now, maybe you've seen or used a different formula. Don't worry about it. In a short course like this, I won't ask you to calculate anything as tedious as a standard deviation. Let the computer do all of the work.

Why is variation important

- Variation = Noise
  - Too much noise can hide signals
- Variation = Heterogeneity
  - Too little heterogeneity, hard to generalize
  - Too much heterogeneity, mixing apples and oranges
- Variation = Unpredictability
  - Too much unpredictability, hard to prepare for the future
- Variation = Risk
  - Too much risk can create a financial burden

Speaker notes

I want to discuss measures of variation now. Variation gets at the heart and soul of clinical statistics. A large portion of statistical analysis involves characterizing variation.

Variation can be thought of as a measure of noise. In general, but not always, noise is bad. Consider measuring a patient's glucose level, to see if you have early evidence of diabetes. Your glucose level varies a lot during the day based on whether you skipped breakfast or decided to get a mid-afternoon Snickers bar. Your glucose level is noisy. A high level might or might not mean trouble. A low value might or might not mean you are safe. The large standard deviation of your measures of blood glucose indicates noise.

That's why you are asked to take an overnight fast before testing your blood glucose level. Controlling your diet by not eating anything after midnight provides a more consistent measure of blood glucose. It has a smaller standard deviation and a high or low value is more helpful in diagnosis.

Variation can also be thought of as a measure of heterogeneity. Heterogeneity is also bad sometimes, but there are times when you want a fair amount of heterogeneity. A research study that has a lot of variation is better at providing a complete picture of what a typical patient is. Outcomes that are consistent in the presence of demographic heterogeneity give you more confidence in generalizing the results of a research study. You have some assurance that the therapy is not restricted to helping a small segment of patients.

Too much heterogeneity, though, can mean that any summary measure is a mixture of apples and oranges. You have to find the right balance.

Variation can be equated to unpredictability. The number of beds needed in a hospital does vary, and this makes it difficult to staff properly. The more variation in beds needed, the more headaches you have.

Variation can also be equated to risk. If you invest in a new drug, paying millions or even billions of dollars in testing, you are doing so with the hope that your investment will pay off. Unfortunately, the market for your drug is uncertain, and you might end up with no market at all if your clinical trials fail to convince FDA. There is variation in the return on your investment, and the more variation there is, the more risky your development plans are.

Should you try to minimize variation?

- Yes, for early studies
  - Easier to detect signals
  - Proof of concept trials
- No, for later studies
  - Easier to generalize results
  - Pragmatic trials

Speaker notes

It is a bit of a generalization, but most researchers try to avoid variation in early studies. By early studies, I mean studies of therapies that have not yet been extensively tested in a broad range of settings. Less variation means that there is a greater chance to detect signals. You remove variation by using very strict entry criteria on who can get into the study. You remove variation by tightly controlling what the patient is allowed to do (e.g., no concommitant medications). You remove variation by tightly standardizing the delivery of the intervention and the assessment of the outcome. You reduce variation by removing patients who deviate from the research protocol requirements.

These are known as proof of concept trials. If a new therapy cannot succeed even under the tight controls, there is no point in studying it futher. But success in a tightly controlled environment does not guarantee success in the real world.

If you are planning a trial that comes after many similar trials, you actually may want to encourage variation. Broaden the inclusion criteria so that the patients in the trial look no different than the patients you see every day in your clinic.

The bell shaped curve

- Does your variation follow a bell shaped curve?
- Synonyms: normality, normal distribution
  - Values in the middle are most common
  - Frequencies taper off away from the center
  - Symmetry on either side
- A bell shaped curve = better characterization of variation

Speaker notes

Much variation in the real world follows a bell shaped curve, alternately called a normal distribution. You can assess whether you have a bell shaped curve using a histogram. Look for values in the middle being most common. The frequencies should taper off slowly as you moved away from the middle. The histogram should have symmetry. The left side of the histogram should be roughly equivalent to the right side of the histogram.

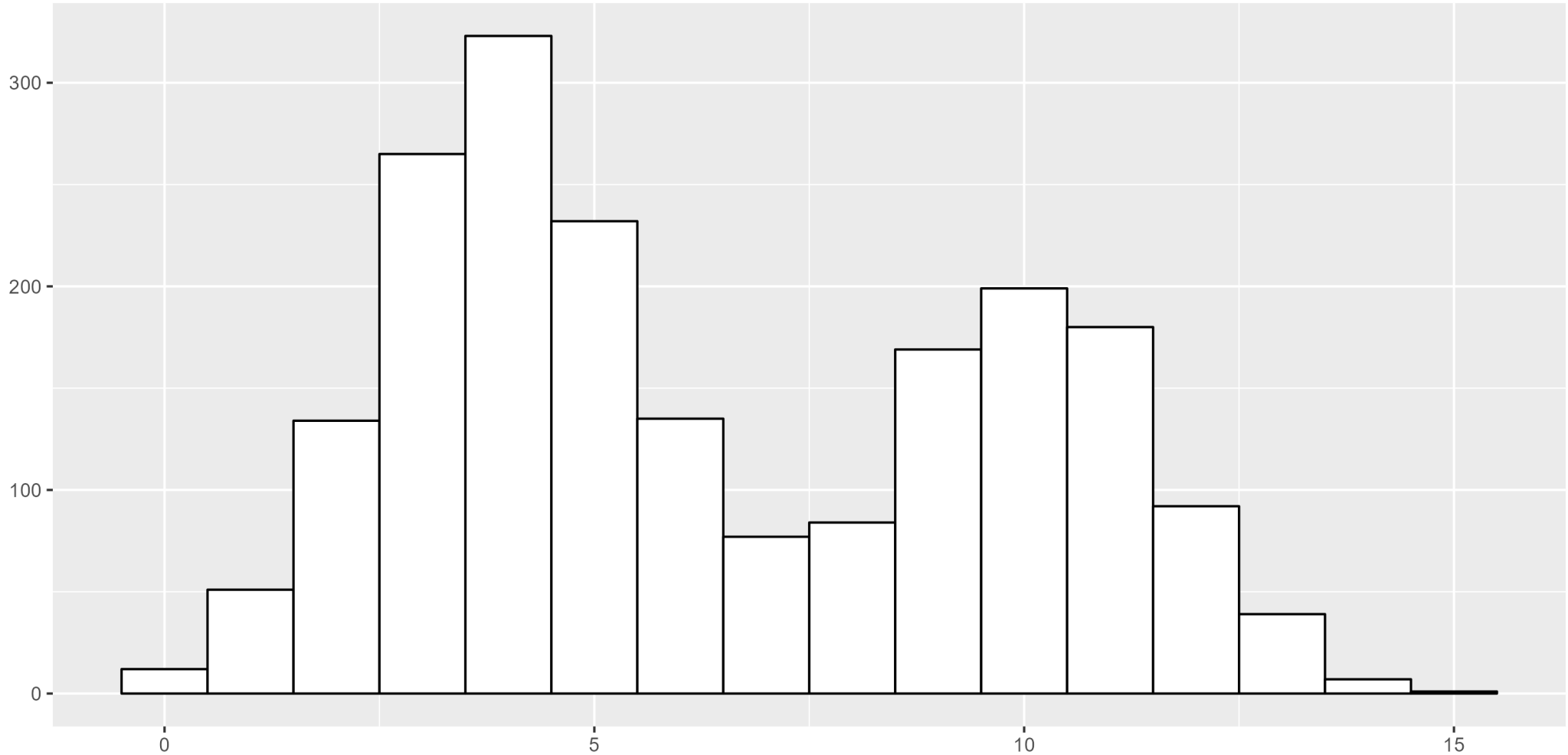


Figure 13: Bimodal histogram, not a bell shaped curve

Speaker notes

Speaker notes

Here’s a histogram that shows a bimodal distribution. The frequencies are not highest in the center of the data. This is not a bell shaped curve.

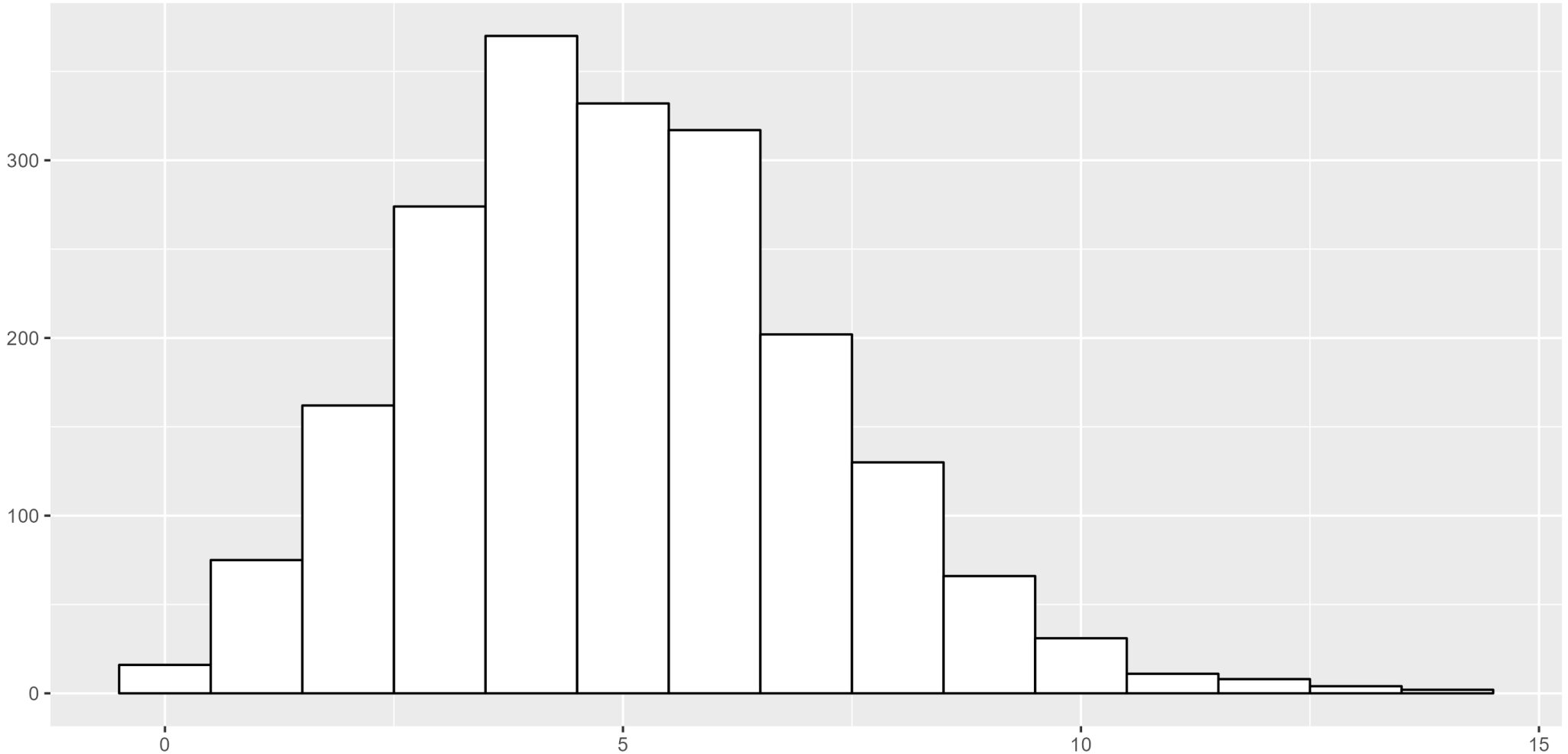


Figure 14: Skewed histogram, not a bell shaped curve

Speaker notes

Speaker notes

Here’s a histogram that shows a skewed or asymmetric distribution. This is not a bell shaped curve.

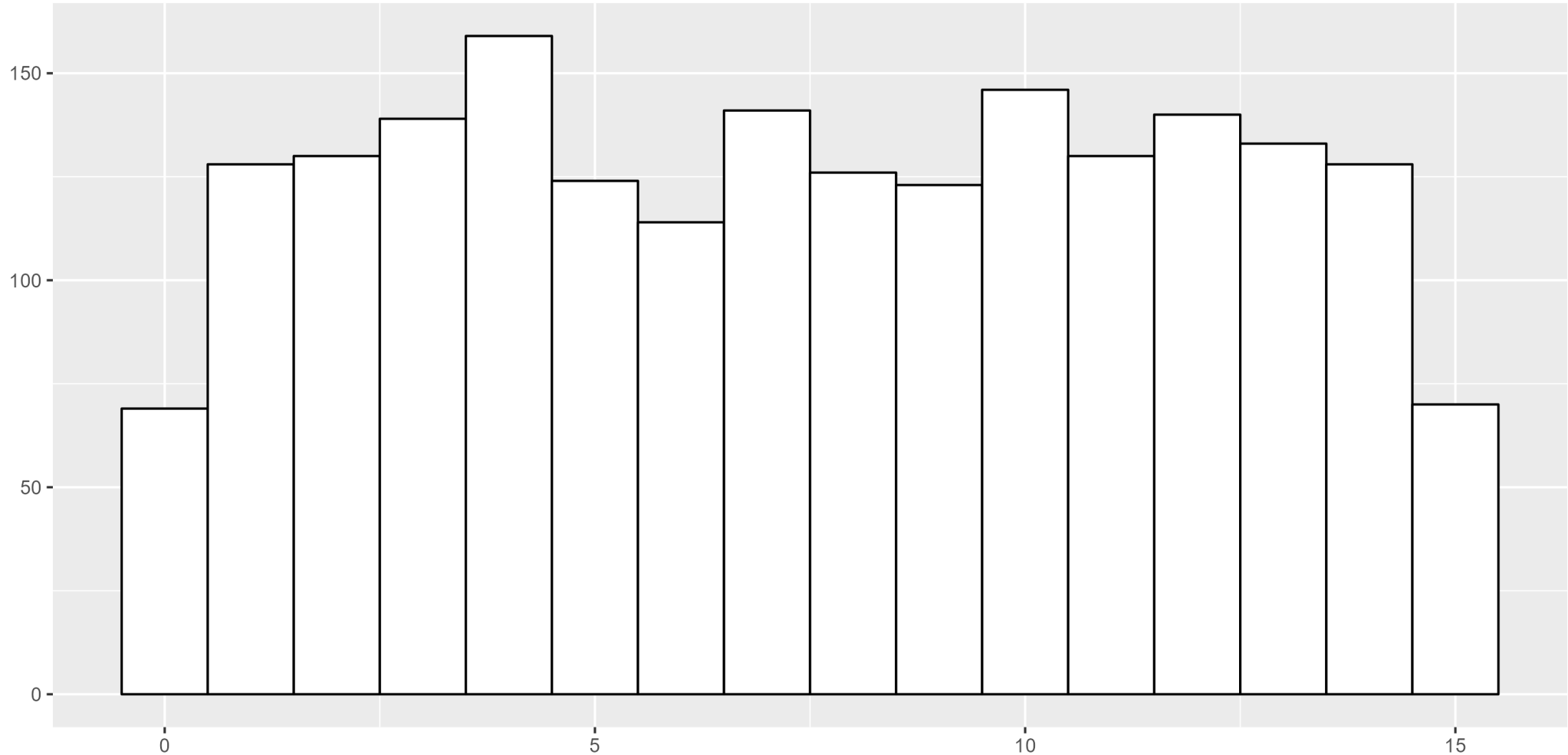


Figure 15: Uniform histogram, not a bell shaped curve

Speaker notes

Here’s a histogram that shows a symmetric distribution, but the frequencies do not taper off as you move away from the center. This is not a bell shaped curve.

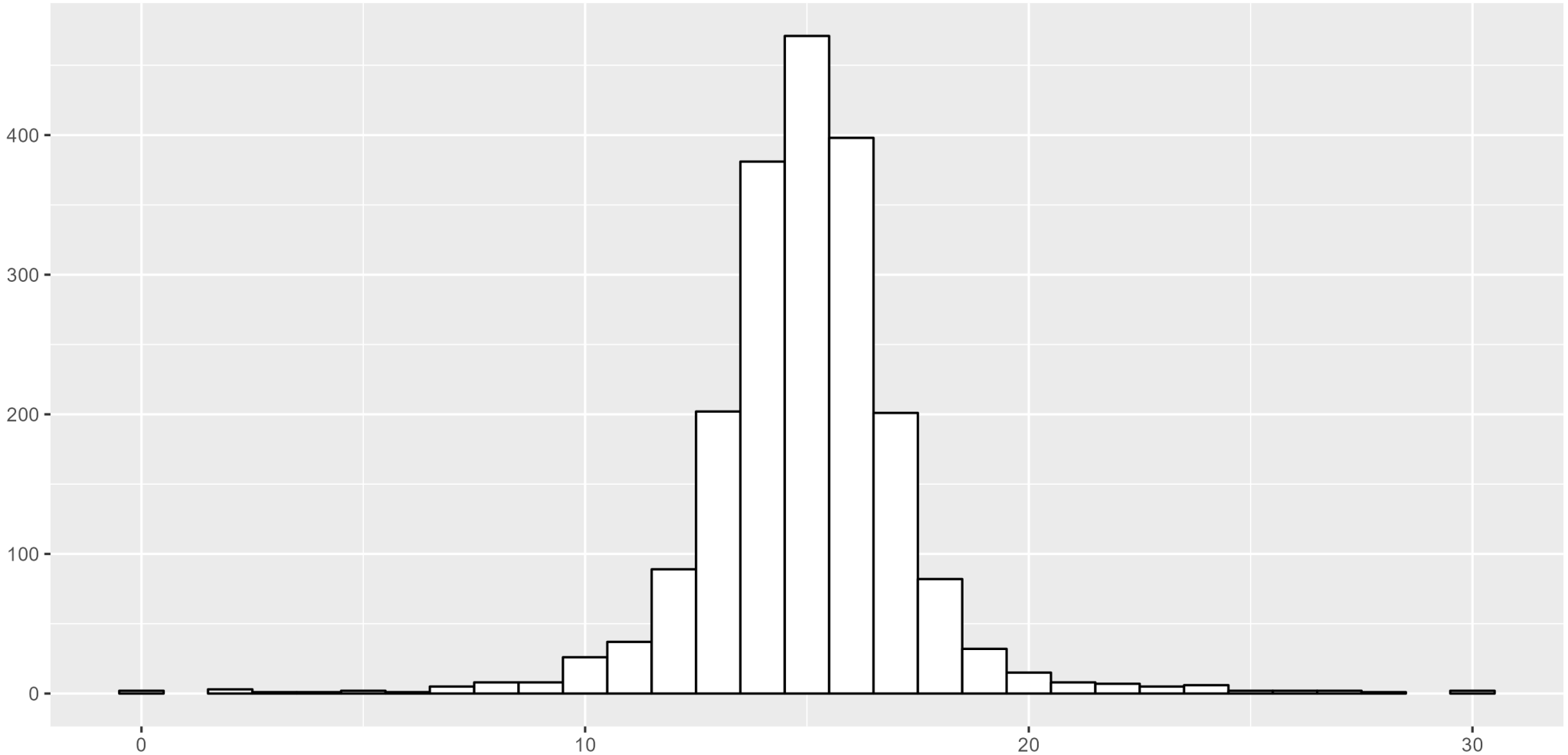


Figure 16: Heavy-tailed histogram, not a bell shaped curve

Speaker notes

Speaker notes

Here’s a histogram that shows a symmetric distibution, but the frequencies taper off at first, but then flatten out. This is called a heavy tailed distribution and it tends to produce outliers, extreme values, on both sides. This is not a bell shaped curve.

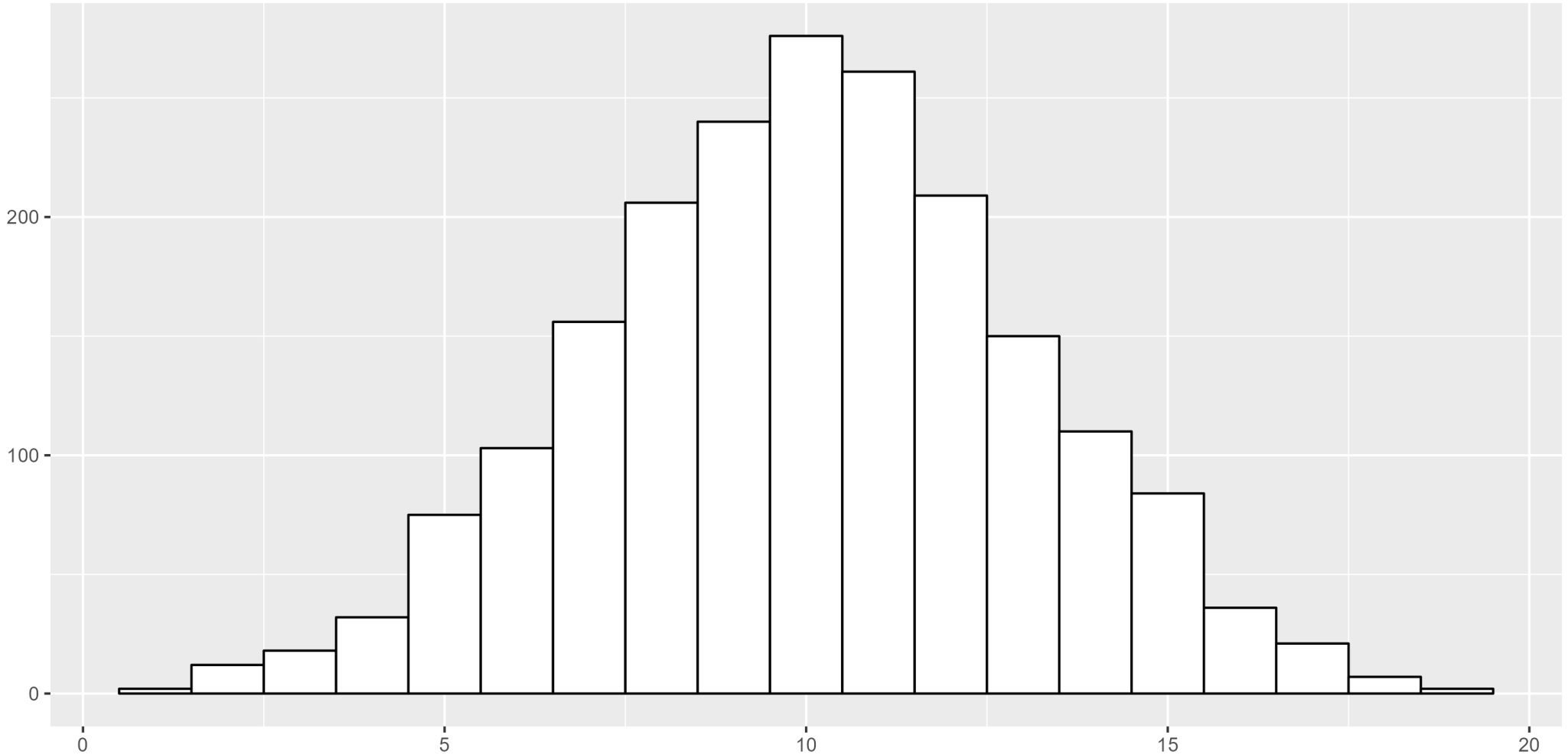


Figure 17: Bell-shaped histogram, finally!

Speaker notes

Speaker notes

Here’s a histogram that shows a symmetric distribution, with the most frequent values in the center and frequencies that taper off on either side. This is a bell shaped curve.

Why concern yourself with the bell shaped curve?

- You can characterize individual observations
- You can characterize summary measures

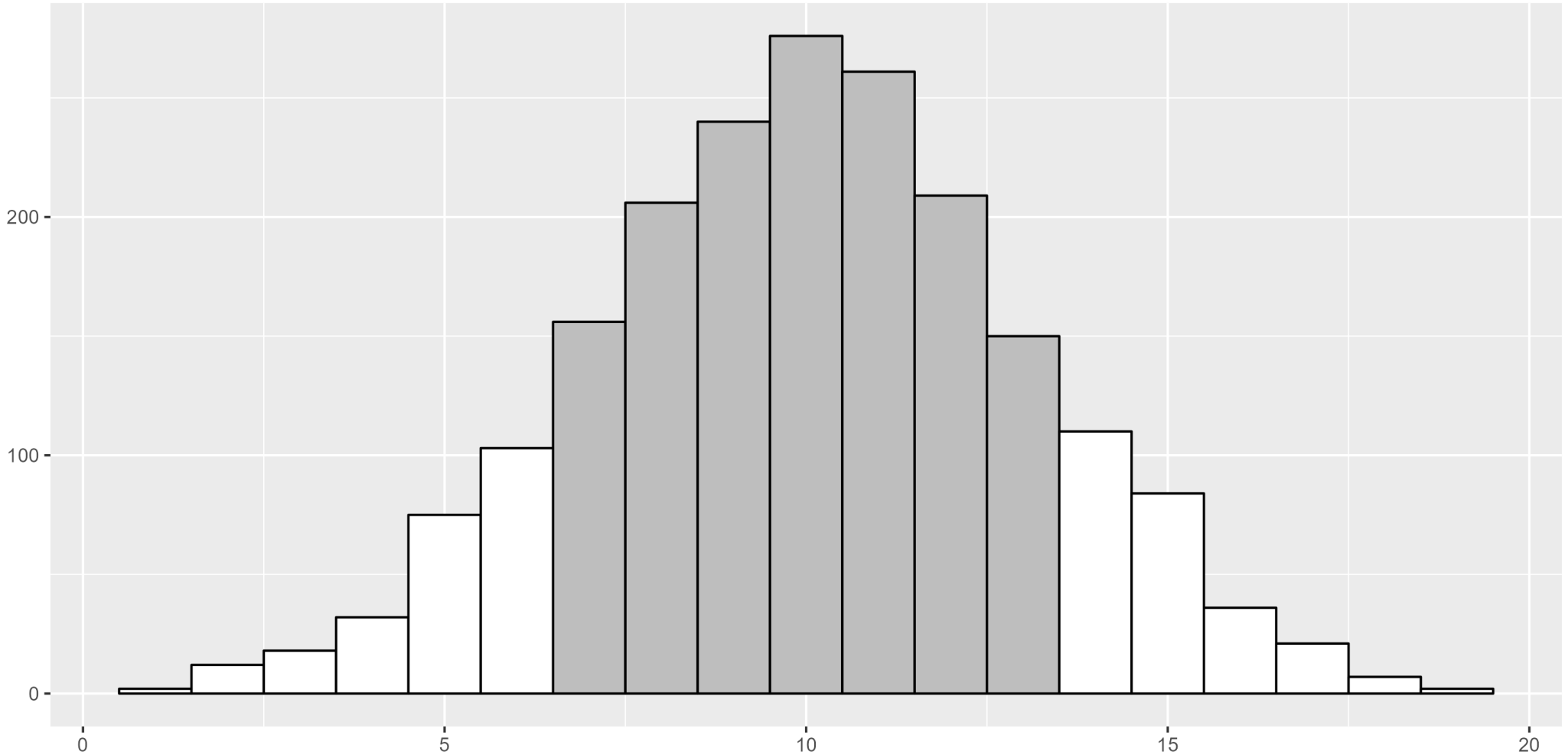


Figure 18: Percentage within one standard deviation

Speaker notes

This shows the bell shaped curve with the data within one standard deviation of the mean highlighted in gray. Roughly 68% of the data lies within one standard deviation of the mean. This is only true if the variation follows a bell shaped curve.

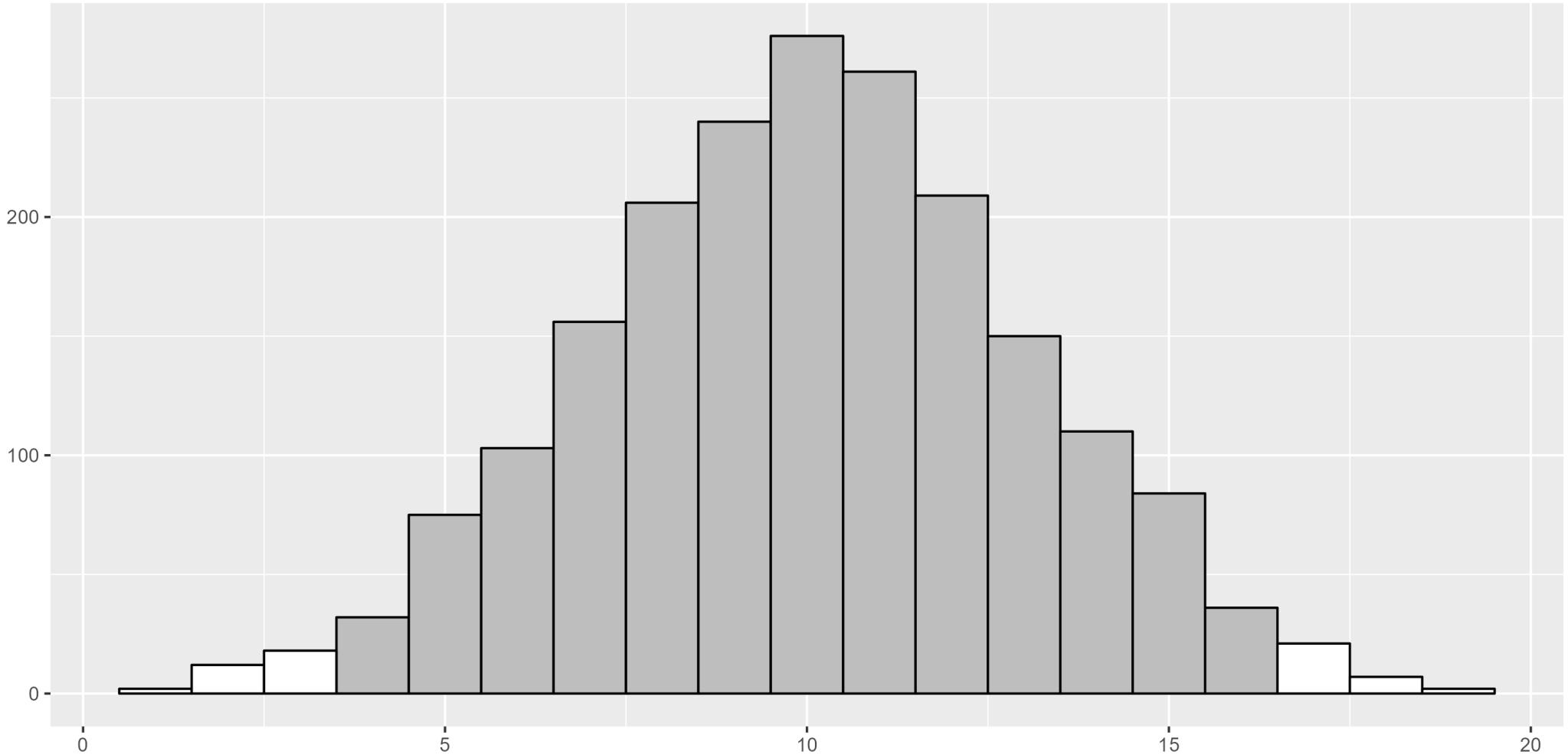


Figure 19: Percentage within two standard deviations

Speaker notes

This shows the bell shaped curve with the data within two standard deviations of the mean highlighted in gray. Roughly 95% of the data lies within two standard deviations of the mean. To repeat, this is only true if the variation follows a bell shaped curve.

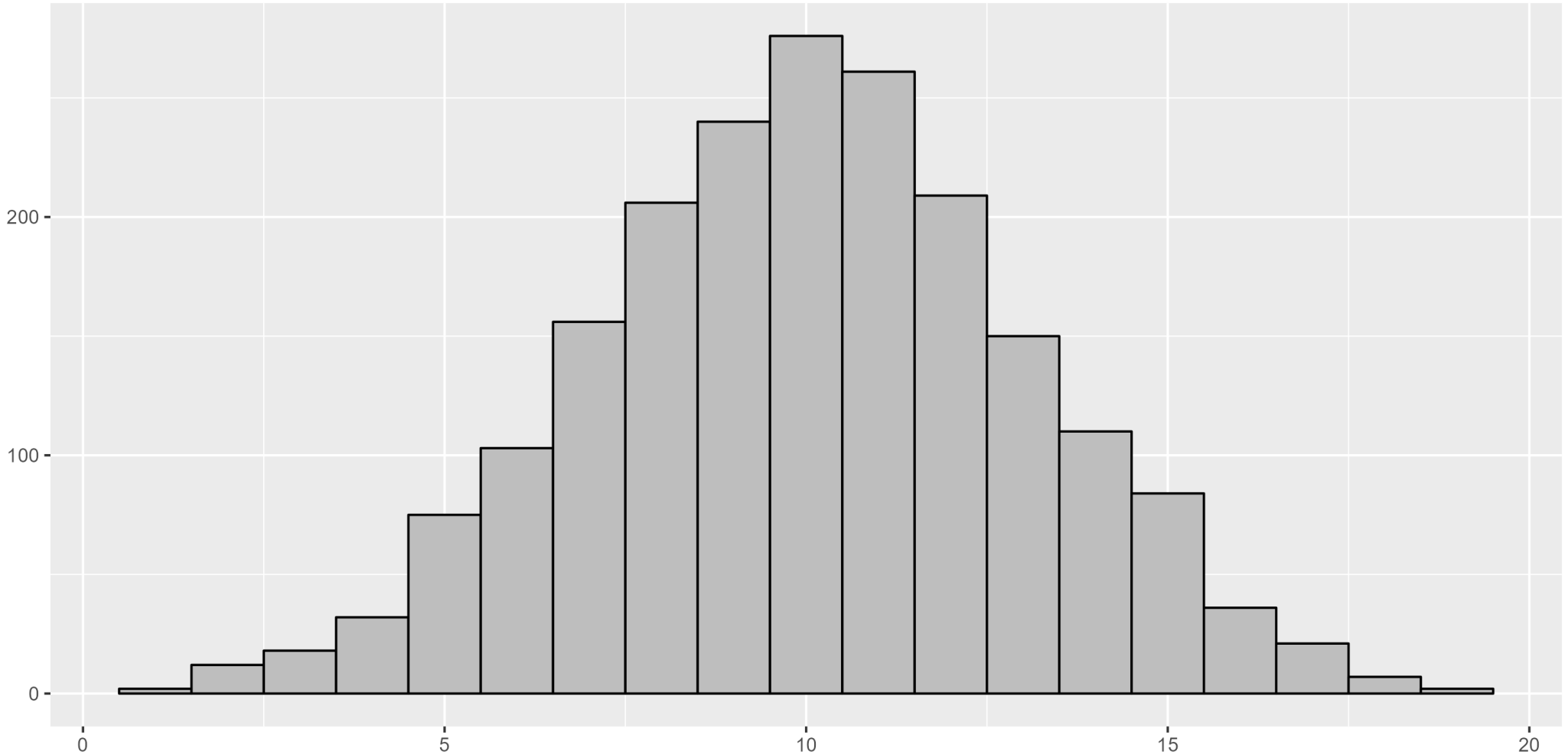


Figure 20: Percentage within three standard deviations

Speaker notes

This shows the bell shaped curve with the data within three standard deviations of the mean highlighted in gray. Almost all of the data lies within three standard deviations of the mean. Rember, and this is worth repeating. This is only true if the variation follows a bell shaped curve.

Behavior of the mean versus an individual

- Central Limit Theorem
  - Sample mean is approximately normal
  - Even if individual observations are not
- Standard error: $S/\sqrt{n}$

Speaker notes

Speaker notes

The histograms above show the behavior of individuals in a sample. The mean of a sample behaves differently.

For almost all settings, the sample mean follows a bell shaped curve. The extremes of bimodality, skewness, and other types of non-normality tend to average out across a sample.

The mean is also less variable than an individual observation, by a factor equal to the square root of the sample size.

What does this mean from a practical sense? It means that averages are much more predictable than individuals. The time that you spend getting a colonoscopy done can vary quite a bit. But the doctor who schedules eight colonoscopies in a day doesn't mind if yours takes a bit too long. From the perspective of the physician, the average time will be fairly predictable, even if an individual colonoscopy time varies quite a bit.

Now it is possible to get all eight colonoscopies taking longer than average. It doesn't happen very often, but on those days, the doctor earns every penny that they pay her.

Remember this perspective. Sometimes you care mostly about the individual value and sometimes you care mostly about the average value.

Diagnosing distributional issues (1/2)

- For all data
  - $\bar{X} \gg X_{0.5}$
  - $\bar{X}$ and/or $X_{0.5}$ not midway between Q_1 and Q_3
  - $\bar{X}$ and/or $X_{0.5}$ not midway between min and max

Speaker notes

Speaker notes

The best way to decide whether your data follows a bell shaped curve is by looking at a histogram. But this is not always an option. When you are reading a journal article, you probably only get descriptive statistics and not a histogram.

There are certain patterns, however, that you might see in the descriptive statistics that SOMETIMES help you assess when the data does not follow a bell shaped curve.

These don’t diagnose every possible deviation from the bell shaped curve, so some types of deviations might be undiagnosed. But if you see any of these patterns, they are definitely indicative of non-normality.

If the mean is a lot larger than the median, it is almost always because of some extreme values, but only on the high end.

If the data is lop-sided, either because the mean/median are not midway between the upper and lower quartiles or because they are not midway between the smallest value and the largest value, this indicates a lack of symmetry and is inconsistent with the bell shaped curve.

Diagnosing distributional issues (2/2)

- For non-negative data
  - $S > 0.5 \times \bar{X}$
- For data with an lower and/or upper bound
  - Q_1 = lower bound
  - Q_3 = upper bound
- Don’t overdiagnose, especially with small sample sizes!

Speaker notes

For non-negative data, data with a lower bound of zero, a large standard deviation, one that is half the size of the mean or greater, you would have the mean minus two standard deviations go negative. That’s a pretty good sign that trying to apply the mean plus or minus two standard deviations is not going to make much sense.

If there is another lower bound, or perhaps an upper bound to the data, see if one of the quartiles hits these bounds. If the 75th percentile equals the upper bound, that means that a quarter of your data is piled up at the upper bound. There’s no way this could be consistent with a bell shaped curve.

Be careful not to overdiagnose. Don’t try to look for subtle patterns. A mean that is just slightly larger than the median or not quite halfway between the lower and upper quartiles is not a cause for concern.

› Prim Care Companion CNS Disord. 2022 Sep 13;24(5):21m03173. doi: 10.4088/PCC.21m03173.

Unemployment, Homelessness, and Other Societal Outcomes Among US Veterans With Schizophrenia Relapse: A Retrospective Cohort Study

Dee Lin [1 2], Hyunchung Kim [3], Keiko Wada [3], Maya Aboumrad [4], Ethan Powell [4], Gabrielle Zwain [4], Carmela Benson [1], Aimee M Near [3]

Affiliations + expand

PMID: 36126916 DOI: 10.4088/PCC.21m03173

Figure 21: Lin et al 2022, PMID: 36126916

Speaker notes

Here is a research study looking at US veterans diagnosed with schizophrenia, comparing those who relapsed with those who did not.

Variable	Relapse Cohort (n = 16,862)	Nonrelapse Cohort (n = 16,862)
Age at index, y		
Mean (SD)	56.4 (13.3)	56.6 (13.0)
Median (Q1, Q3)	58 (51, 64)	59 (51, 65)
Minimum	20	20
Maximum	99	98

Figure 22: Excerpt from Table 1 of Lin et al 2022: ages

Speaker notes

Speaker notes

Here is part of a table for this article that gives a fair amount of detail about the ages of the veterans in this research study.

The data here looks fairl well behaved. The mean and the median are close to one another. Ages are non-negative, so you could look at the standard deviation. A really large standard deviation could be trouble, but here is is around 13, nowhere close to half of the mean. The mean and median both find themselves about midway between the two quartiles and between the minimum and maximum values.

CCI		
Mean (SD)	2.5 (2.7)	2.4 (2.4)
Median (Q1, Q3)	2 (0, 4)	2 (0, 4)
Minimum	0	0
Maximum	20	18

Figure 23: Excerpt from Table 1 of Lin et al 2022: CCI

Speaker notes

CCI is short for the Charlson Comorbidity Index. It is a weighted sum of nineteen medical conditions.The larger the value of CCI, the sicker the patient.

The CCI has a lower bound of zero, so you can see a problem right away. The standard deviations are about the same size or a bit bigger than the mean. The mean and medians are about midway between the two quartiles, but when you look at the minimum and maximum, there is quite a different story.

This is indeed not normally distributed.

PHQ-2 score, n (%)[f]	9,472 (56.1)	10,848 (64.4)
Mean (SD)	1.1 (1.7)	1.0 (1.6)
Median (Q1, Q3)	0 (0, 2)	0 (0, 2)
Minimum	0	0
Maximum	6	6

Figure 24: Excerpt from Table 1 of Lin et al 2022: PHQ-2

Speaker notes

PHQ-2 is a quality of life measure. It ranges from 0 to 6. The very large standard deviations, both quite a bit bigger than the means, indicates strong evidence of non-normality.

J Am Med Dir Assoc. 2021 Sep;22(9):1840-1844. doi: 10.1016/j.jamda.2021.07.003. Epub 2021 Jul 19.

Prevalence and Predictors of Persistence of COVID-19 Symptoms in Older Adults: A Single-Center Study

Matteo Tosato [1], Angelo Carfì [1], Ilaria Martis [1], Cristina Pais [1], Francesca Ciciarello [1], Elisabetta Rota [1], Marcello Tritto [1], Andrea Salerno [1], Maria Beatrice Zazzara [1], Anna Maria Martone [1], Annamaria Paglionico [1], Luca Petricca [1], Vincenzo Brandi [1], Gennaro Capalbo [1], Anna Picca [1], Riccardo Calvani [2], Emanuele Marzetti [3], Francesco Landi [3]; Gemelli Against COVID-19 Post-Acute Care Team

Affiliations + expand

PMID: 34352201 PMCID: PMC8286874 DOI: 10.1016/j.jamda.2021.07.003

Figure 25: Tosato et al 2021, PMID: 34352201

Speaker notes

Speaker notes

Tosato 2021, PMID: 34352201 (continued)

Symptom persistence weeks after laboratory-confirmed severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) clearance is a relatively common long-term complication of Coronavirus disease 2019 (COVID-19). Little is known about this phenomenon in older adults. The present study aimed at determining the prevalence of persistent symptoms among older COVID-19 survivors and identifying symptom patterns.

Speaker notes

Speaker notes

Here’s another article looking at long Covid, the persistence of symptoms long after infection.

Tosato 2021, PMID: 34352201 (continued)

The mean age was 73.1 ± 6.2 years (median 72, interquartile range 27), and 63 (38.4%) were women. The average time elapsed from hospital discharge was 76.8 ± 20.3 days (range 25-109 days).

Speaker notes

There are fewer statistics presented here. Age and time since discharge are both non-negative, so you can compare the standard deviation to the mean. No problems here. It might be worth calculating the mean plus or minus two standard deviations. Rounding a bit you get 61 to 85 for age and 37 to 117 for time

Ielapi 2021, PMID: 34968328

› Nurs Rep. 2021 Jul 12;11(3):530-535. doi: 10.3390/nursrep11030050.

Insomnia Prevalence among Italian Night-Shift Nurses

Nicola Ielapi [1 2], Michele Andreucci [3], Umberto Marcello Bracale [4], Davide Costa [2 5], Egidio Bevacqua [2 6], Andrea Bitonti [7], Sabrina Mellace [8], Gianluca Buffone [9], Stefano Candido [10], Michele Provenzano [6], Raffaele Serra [2 6]

Affiliations + expand

PMID: 34968328 PMCID: PMC8608071 DOI: 10.3390/nursrep11030050

Figure 26: Ielapi et al 2021, PMID: 34968328

Ielapi 2021, PMID: 34968328 (continued)

Background. Insomnia is one of the major health problems related with a decrease in quality of life (QOL) and also in poor functioning in night-shift nurses, that also may negatively affect patients' care. The aim of this study is to evaluate the prevalence of insomnia in night shift nurses.

Speaker notes

Here’s an article about insomnia among nurses.

Ielapi 2021, PMID: 34968328 (continued)

Excerpt from Table 1. Data reported as mean ± standard deviation or median [Q1-Q3]

Overall (n = 2′355)	
Age, years	40.4 ± 10.3
Months of work	168 [72–300]
Night shifts per month, number	6.3 ± 1.4
Time to reach workplace, minutes	45 [45–65]
Rest time, minutes	180 [4–240]
Rest in the afternoon, minutes	30 [0–120]
Number of coffees, mean	2.5 ± 1.5
Number of coffees during night shift, mean	1.4 ± 1.1

Speaker notes

Speaker notes

Break into small groups. The first group take the first four measures and the second group take the last four measures. Is there evidence that the data is non-normally distributed?

Also, calculate the mean plus of minus two standard deviations for any observations that report the mean and standard deviation.

Break

- What have you just learned?
  - Computing the standard deviation
- What is coming next?
  - Visualization

Speaker notes

Let me pause here for a second. Are there any questions?

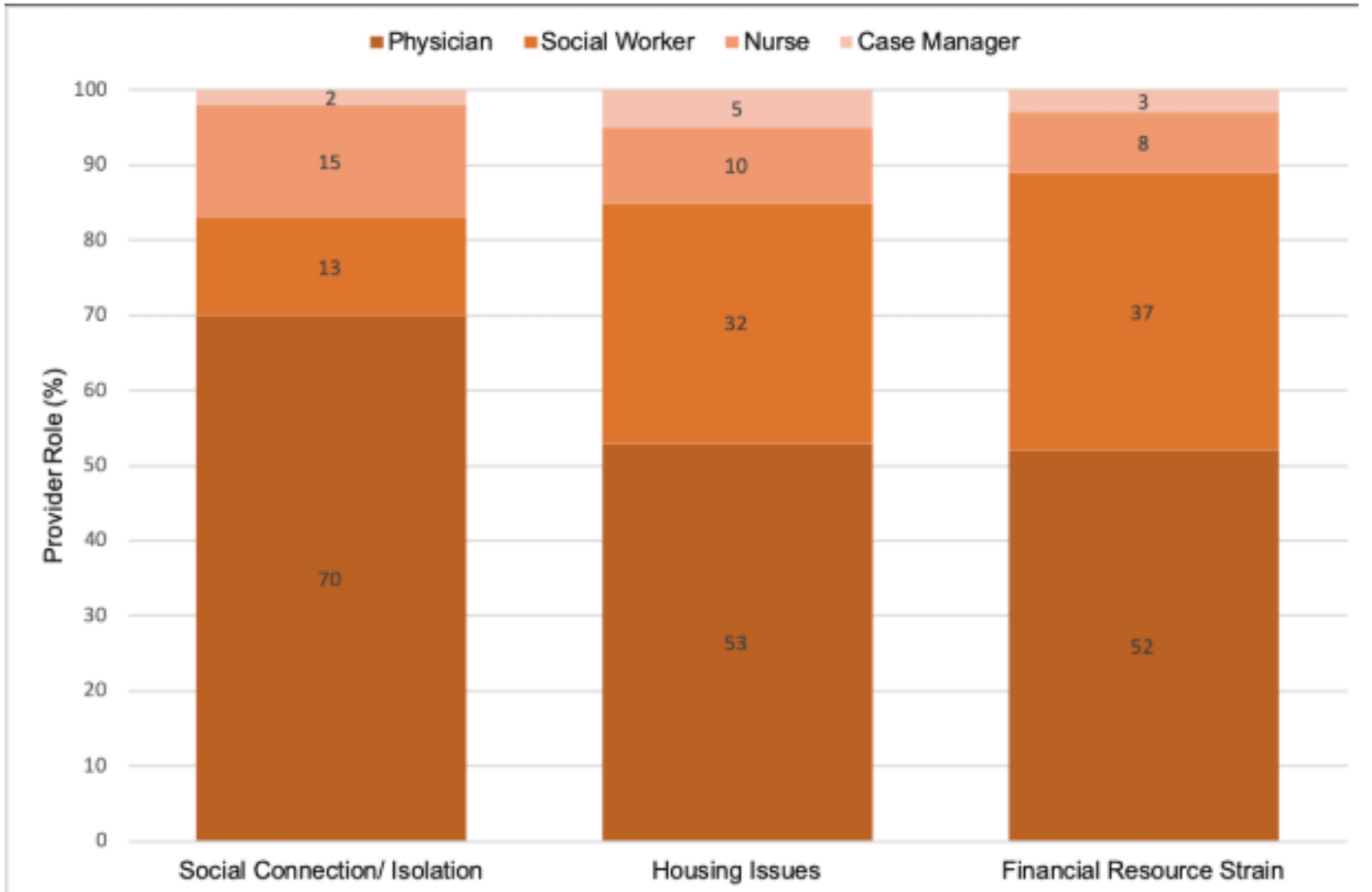


Figure 1. Characteristics of the electronic health record's unstructured data containing social and behavioral determinants of health, stratified by provider role.

Entries in the electronic health record by job title

Speaker notes

This bar chart is from an article about the unstructured part of the electronic health care records, often called the physician notes.

I’m going to ask you to interpret a few bar charts like this one. In this bar chart, you see that physicians are responsible for the bulk of unstructured data. This is especially true for comments about social connection and isolation. Case managers provide 5% or less of the unstructured comments.

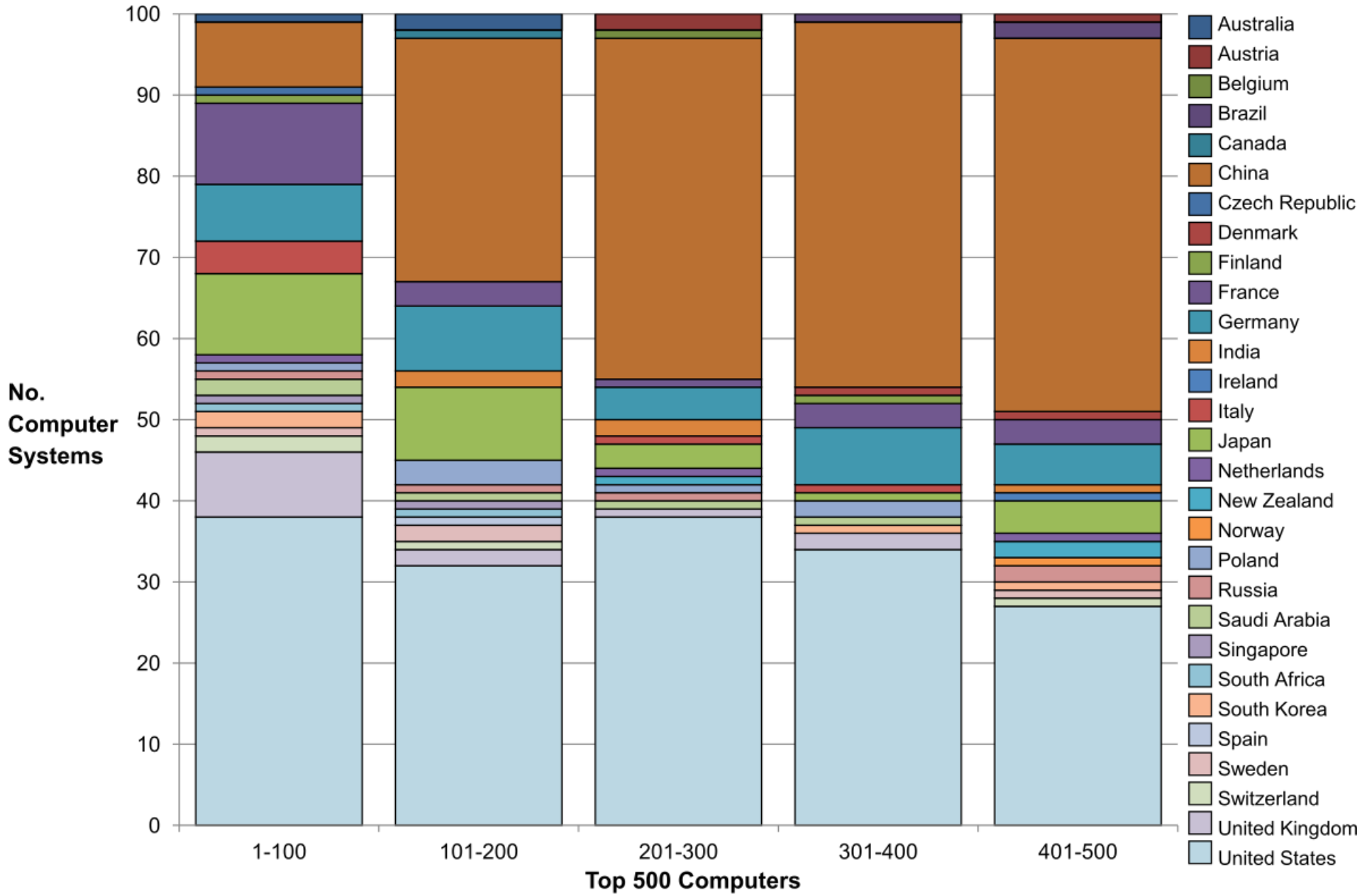


Figure 27: Fastest computers by country

Speaker notes

Interpret this bar chart.

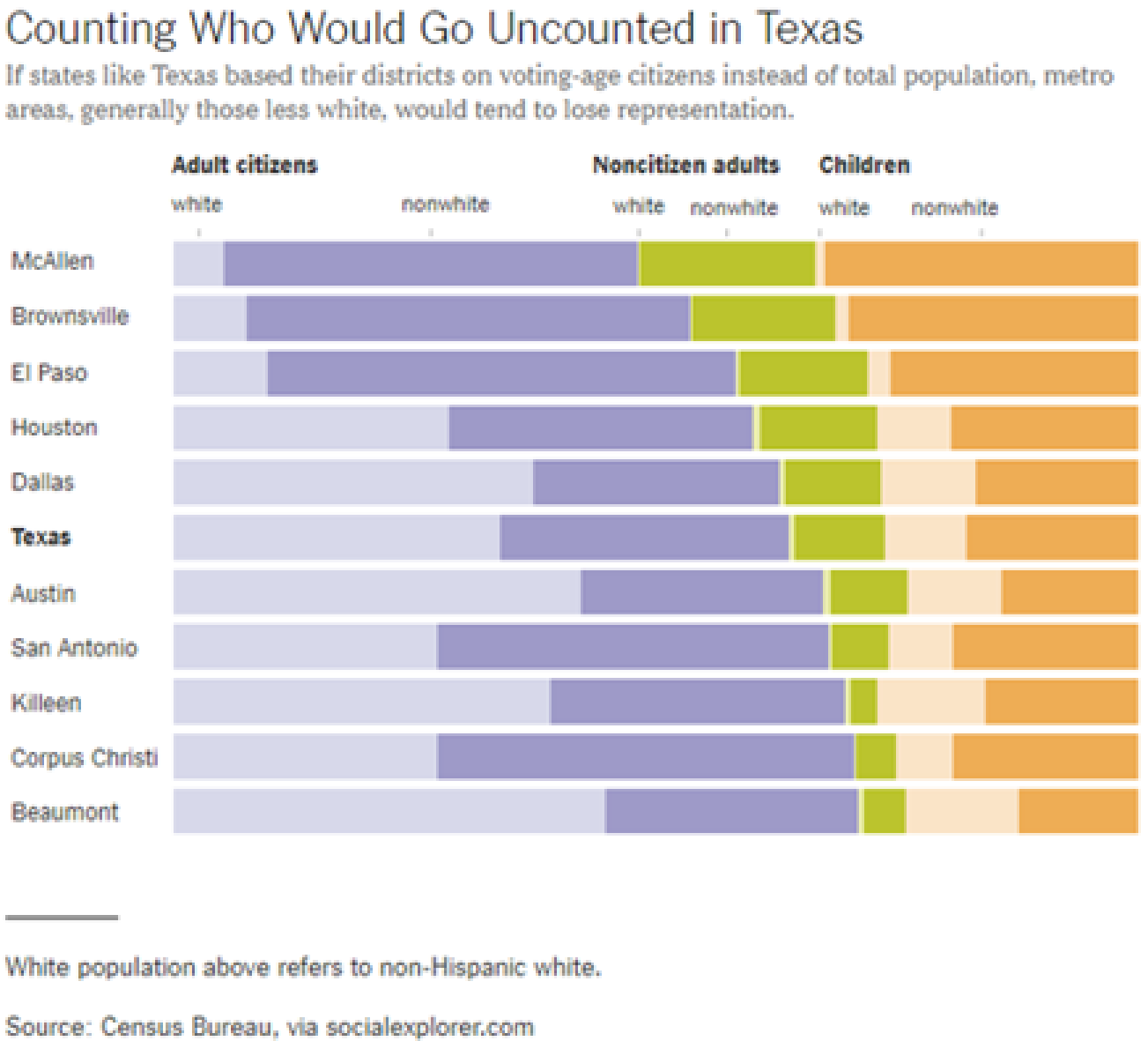


Figure 28: Demographic distribution of voters and non-voters in Texas

Speaker notes

Interpret this bar chart.

Visualization

- Categorical data
  - Pie charts
  - Bar charts
- Continuous data
  - Bar charts
  - Error bars
  - Boxplot
  - Histogram
  - Plot all the data

Speaker notes

Speaker notes

There are many ways to visualize data. Your choices depend on the type of data you are trying to visualize. If all your variables are categorical, the plots that you are most likely to see are pie charts and bar charts.

For continuous data, you might see bar charts (with or without error bars), boxplots, or histograms. All of these summarize the data, but often plotting all the individual data values is your best choice.

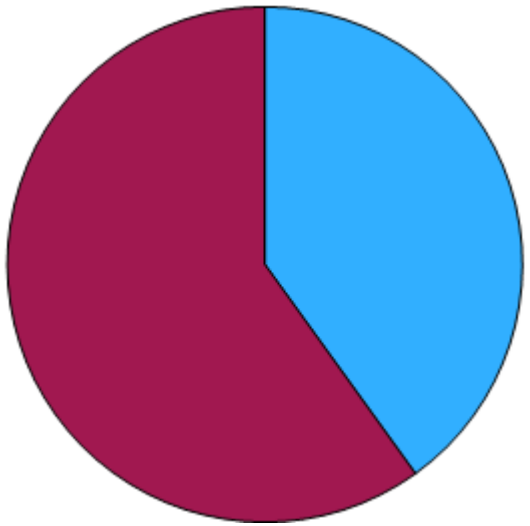


Figure 29: Survivors among first class passengers

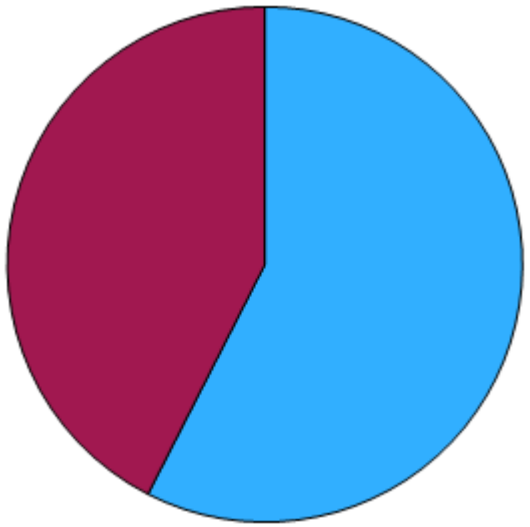


Figure 30: Survivors among second class passengers

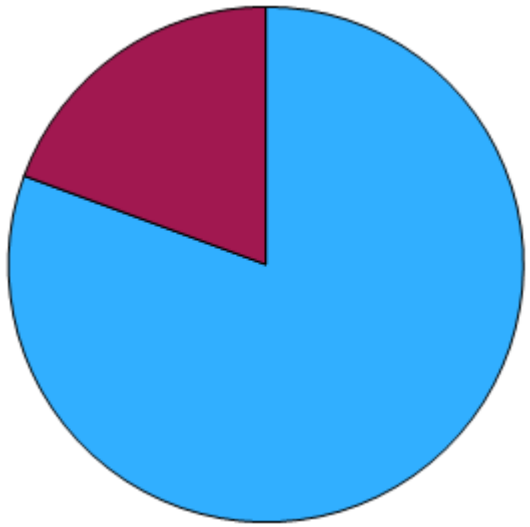


Figure 31: Survivors among third class passengers

Speaker notes

I do not recommend pie charts. Here are three pie charts showing survivors among the various passenger classes. The red portion represents those who survived and the blue portion represents those who died.

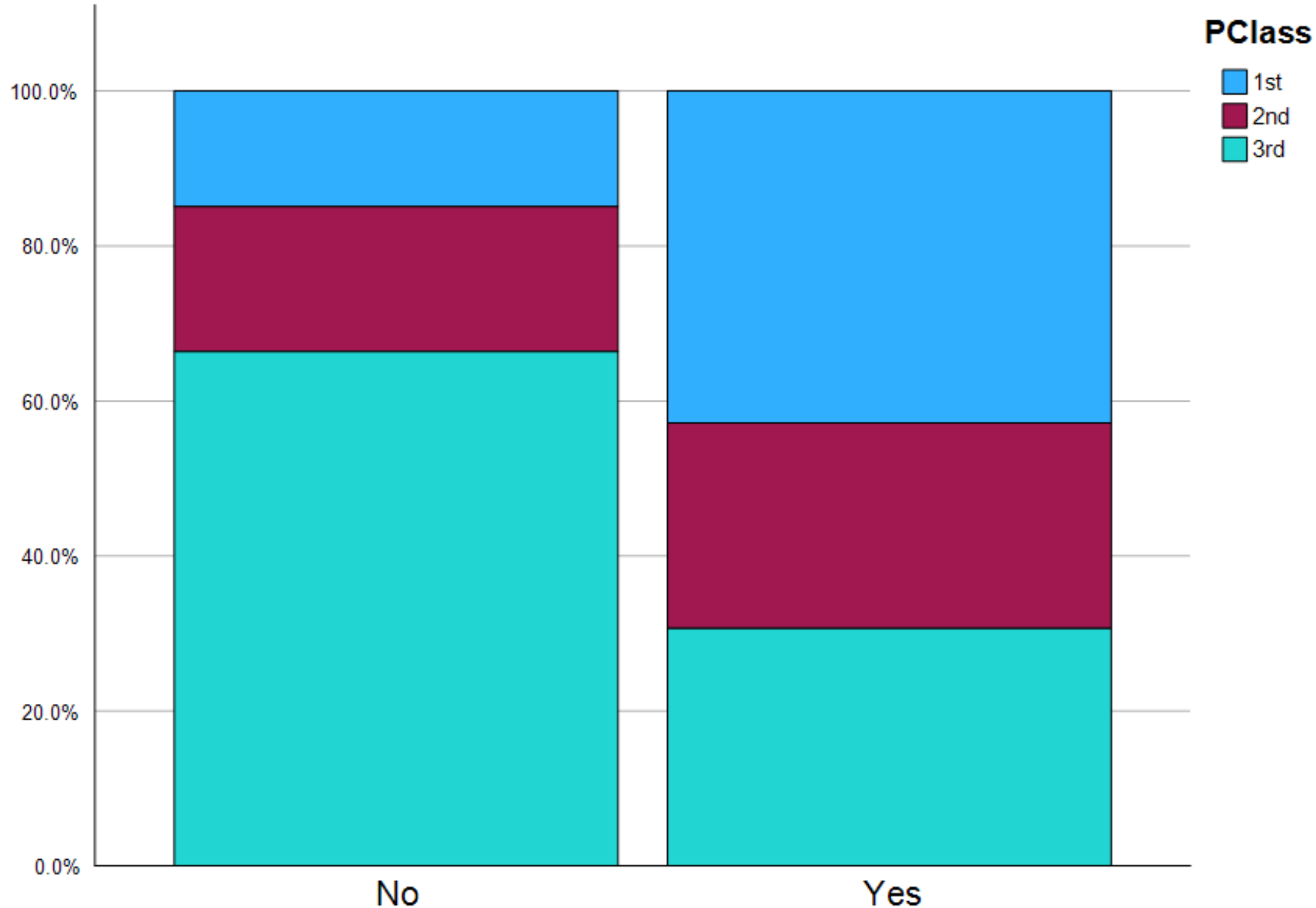
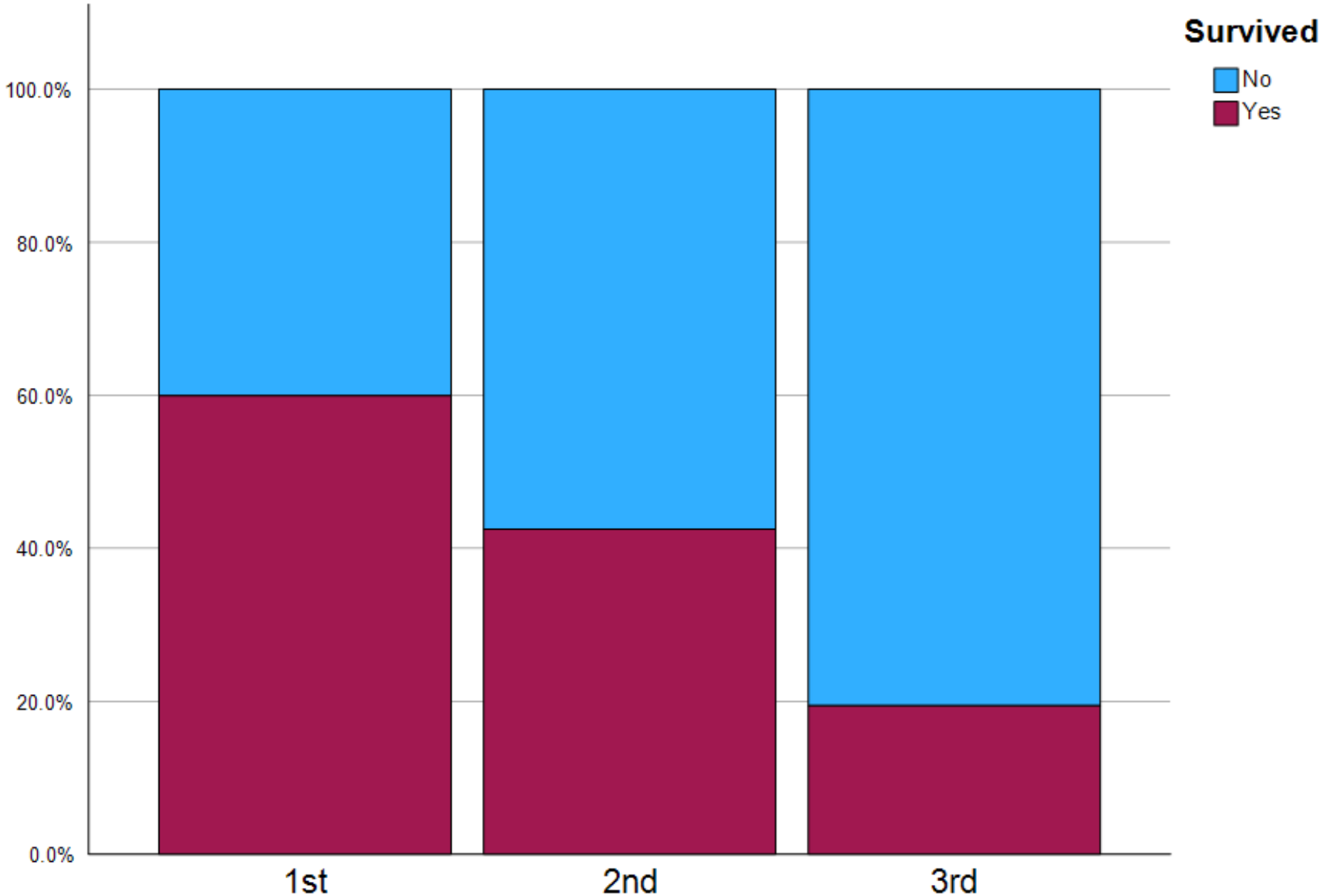


Figure 32: Bar chart showing proportion of passenger classes among deaths and survivors

Speaker notes

Here is the same data displayed as a bar chart. Survivors are evenly split betwen the three passenger classes, but fatilities were dominated by third class passengers.



Bar chart showing proportion of deaths and survivors among passenger classes

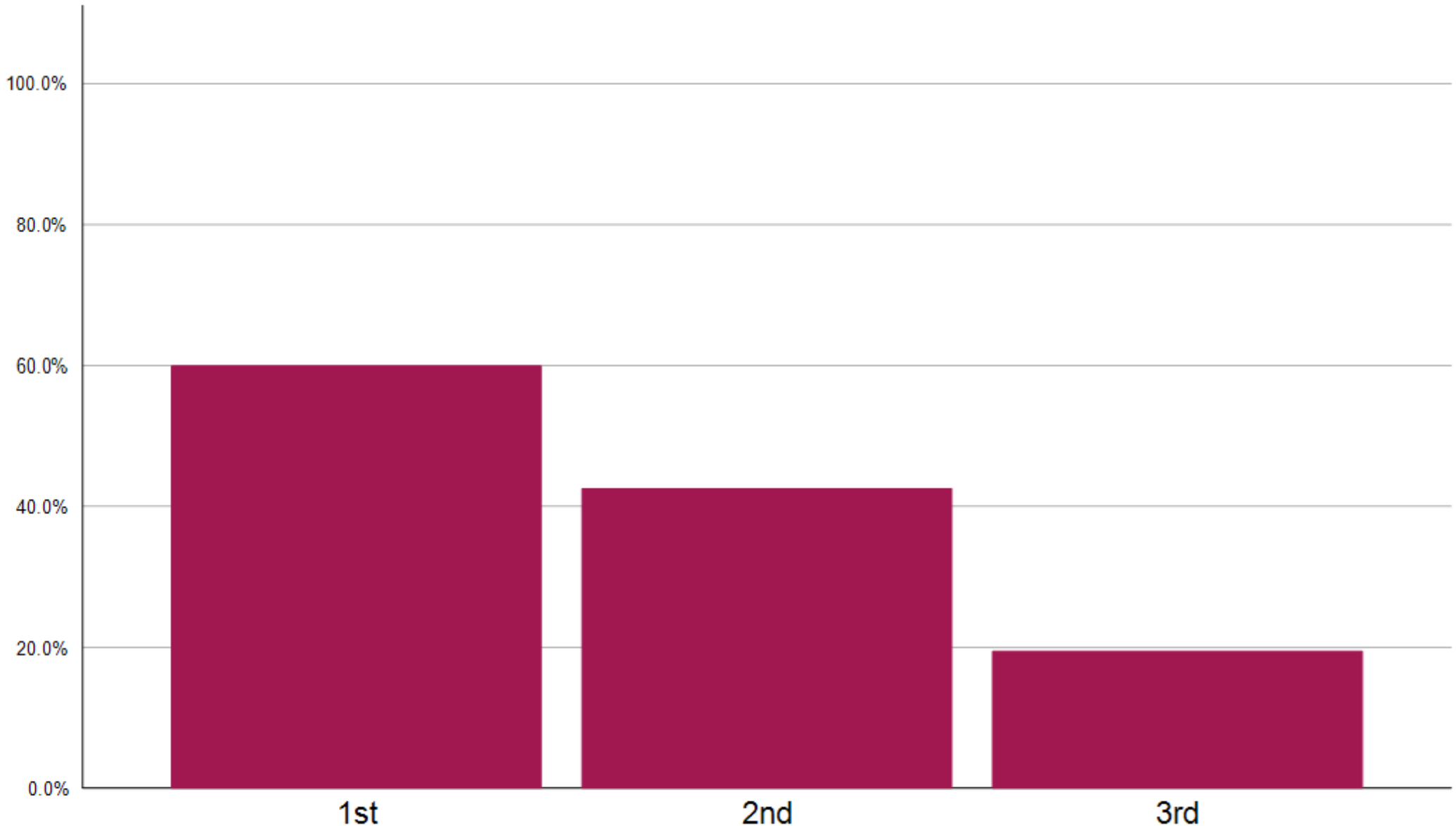
Speaker notes

Speaker notes

Now with a chart like this you could display the data as two bars fatalities and survivors to see how they split up among the three passenger classes. But there is an alternate way to display this data.

You could look at three bars, one for each passenger class with each bar split into two pieces, fatalities and survivors.

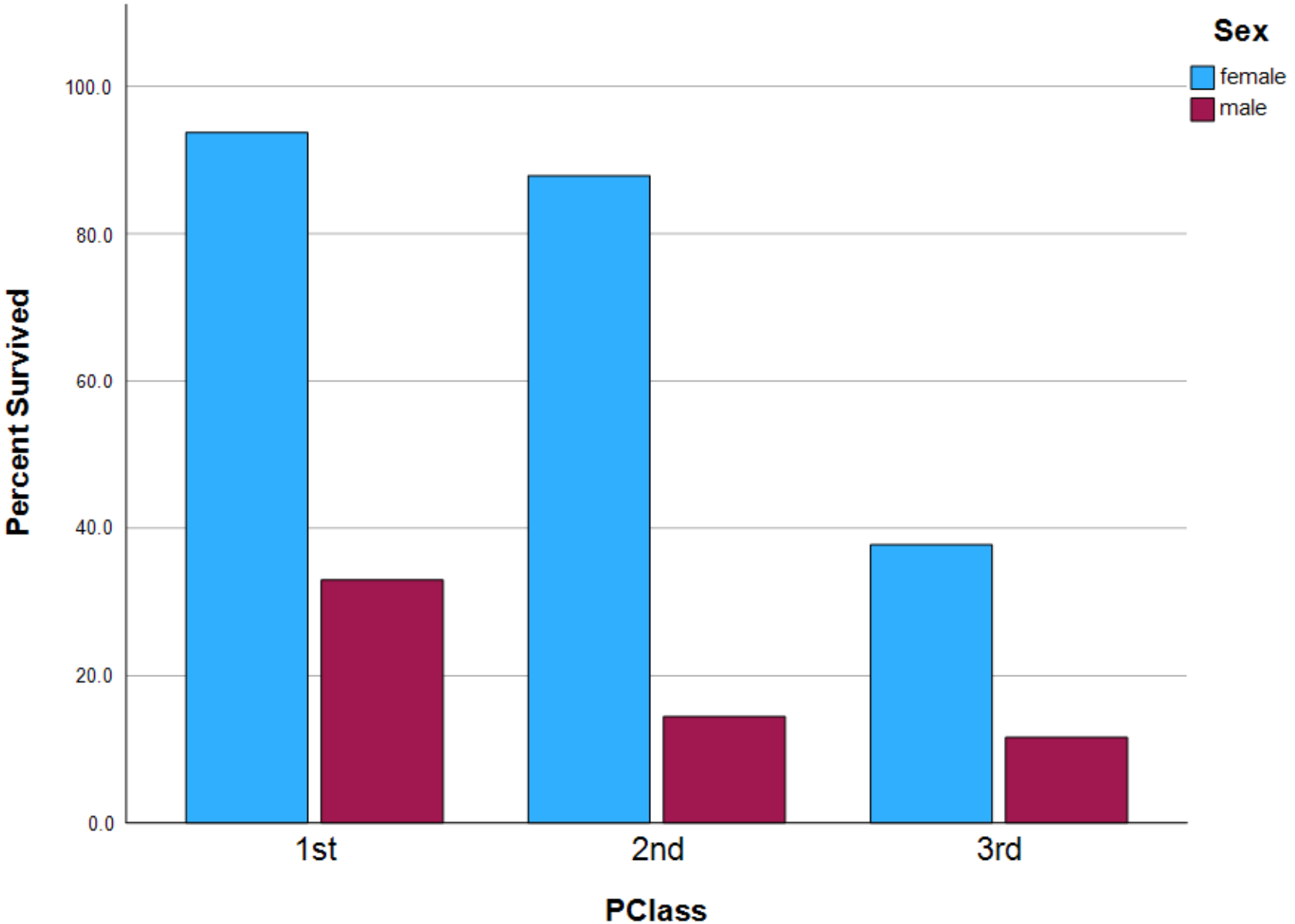
Which is correct? Well it is just like the question raised earlier about row percents and column percents. A lot depends on your perspective, but I like the chart currently being shown as better. It shows decreasing proportion of survivors when you consider first class versus second or third class. It may have been women and children first, but it was also rich people first.



Bar chart showing only proportion of survivors among passenger classes

Speaker notes

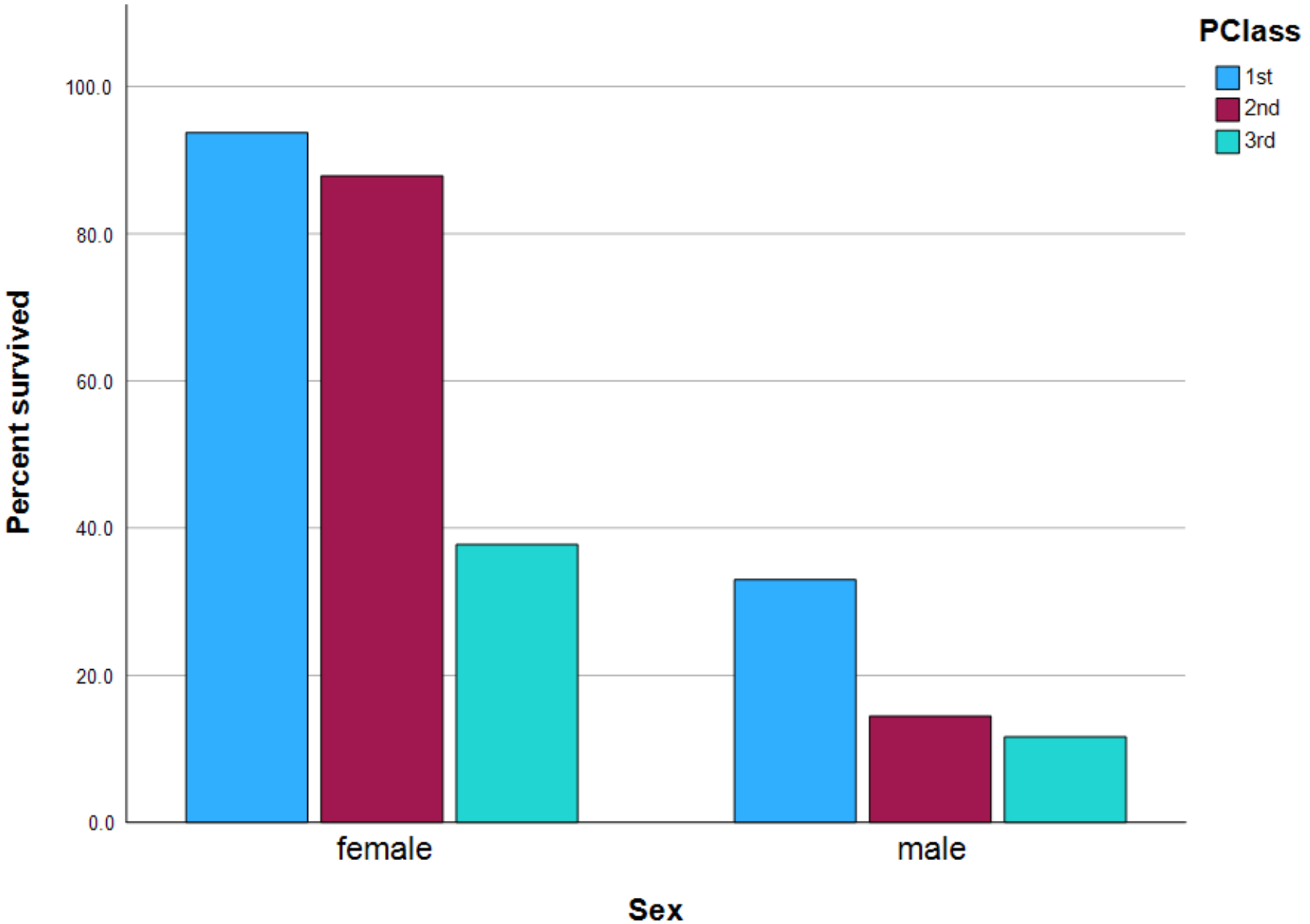
When you have bar charts divided into two pieces, there is no reason that you have to show both pieces. This is the exact same bar chart with the top part of the bar replaced. It shows the exact same trend, but with a bit less clutter.



Bar chart showing proportion of survivors among passenger classes and sex

Speaker notes

Let’s add another dimension. Let’s look at passenger class and gender. This is called a clustered bar chart. This shows that regardless of passenger class, women fared better than men.



Bar chart showing proportion of survivors among sex and passenger classes

Speaker notes

Speaker notes

There’s an alternate way to cluster the bar charts. Put the women in all three classes together in one cluster and the men in a differnt cluster. This plot shows that both men and women fared better if they paid for a more expensive cabin.

Which is best? Well it depends on what you want to emphasize. If you want to contrast men and women, put the male and female bars next to one another. If you want to contrast passenger classes, put them next to one another.

The rule of thumb here is that the things that are the closest are the things that are most easily compared.

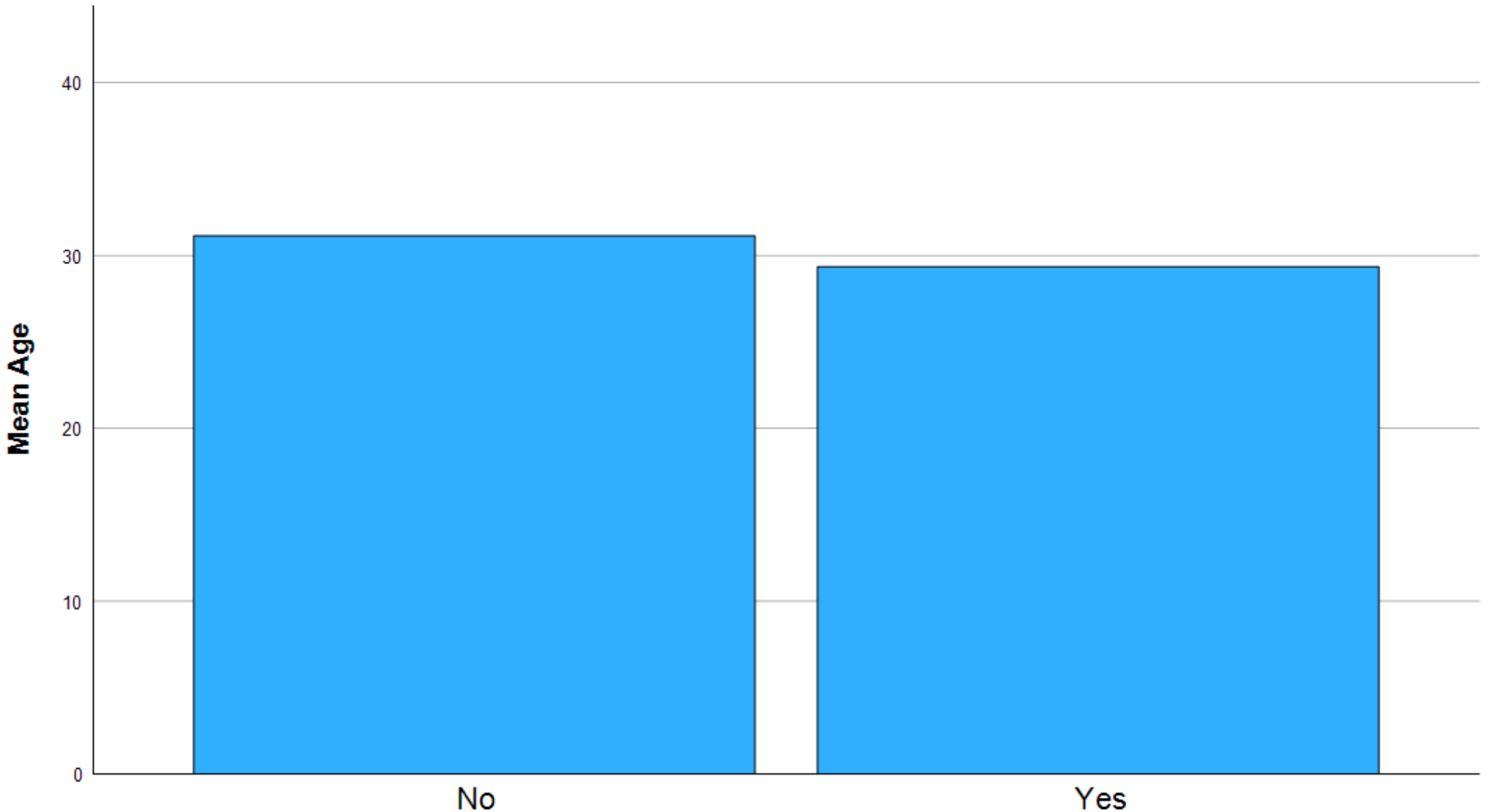


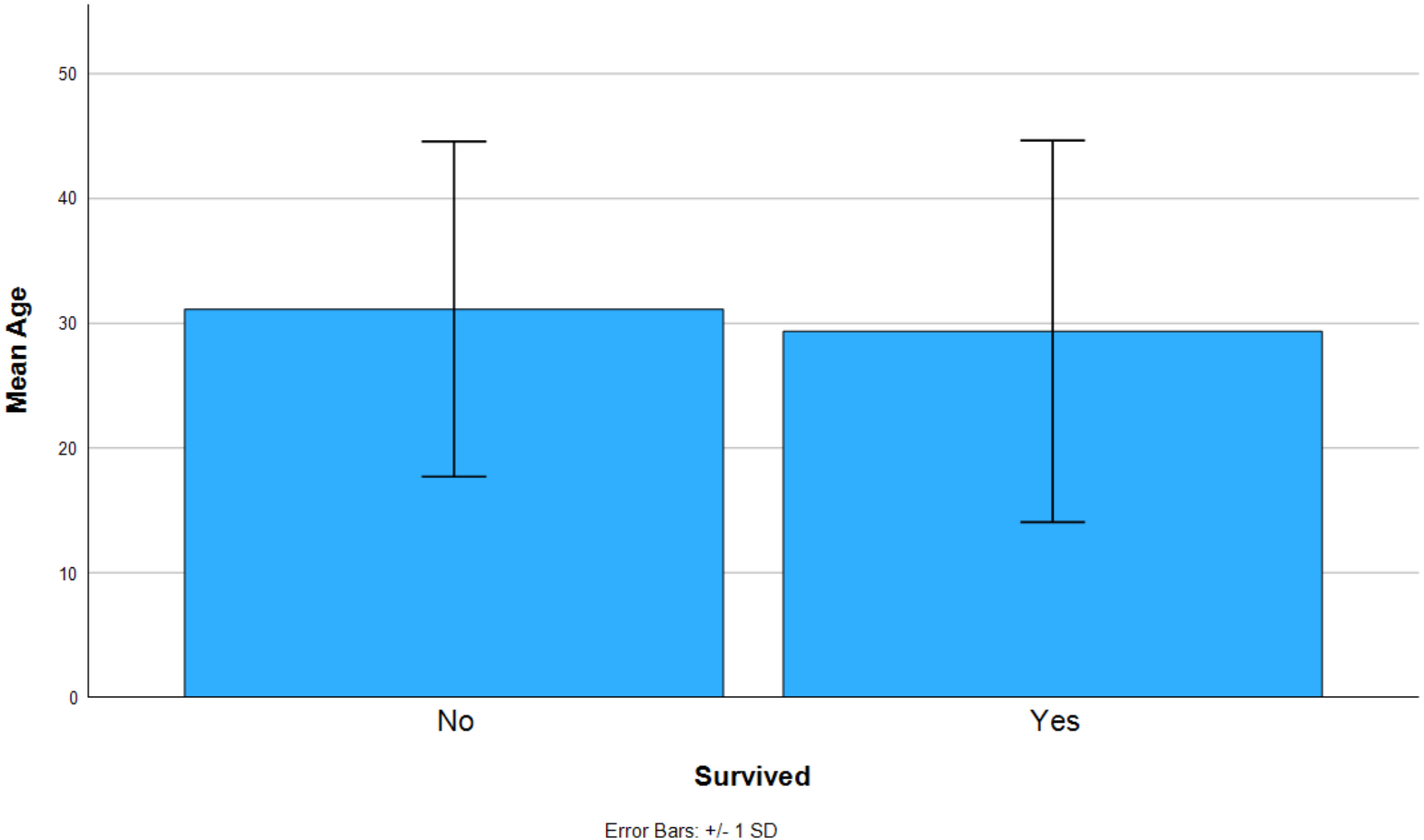
Figure 33: Bar chart showing average age among deaths and survivors

Speaker notes

Speaker notes

Now let’s look at a continuous variable, age. This bar chart shows the average age among those who died and those who survived. A very small difference.

Now I cannot recommend this bar chart because it shows a single number and cannot offer information about how variable the data is.



Bar chart with error bars showing proportion of survivors among sex and passenger classes

Speaker notes

Speaker notes

A common choice is to include error bars. In this example, the error bar shows plus and minus one standard deviation.

Two numbers is better than one, but it is still not good. The error bars, by definition are symmetric, but the data that generated the error bars may or may not be symmetric.

Another problem is that some researchers will use error bars to represent standard errors, confidence intervals, or ranges.

I do not recommend that you use error bars.

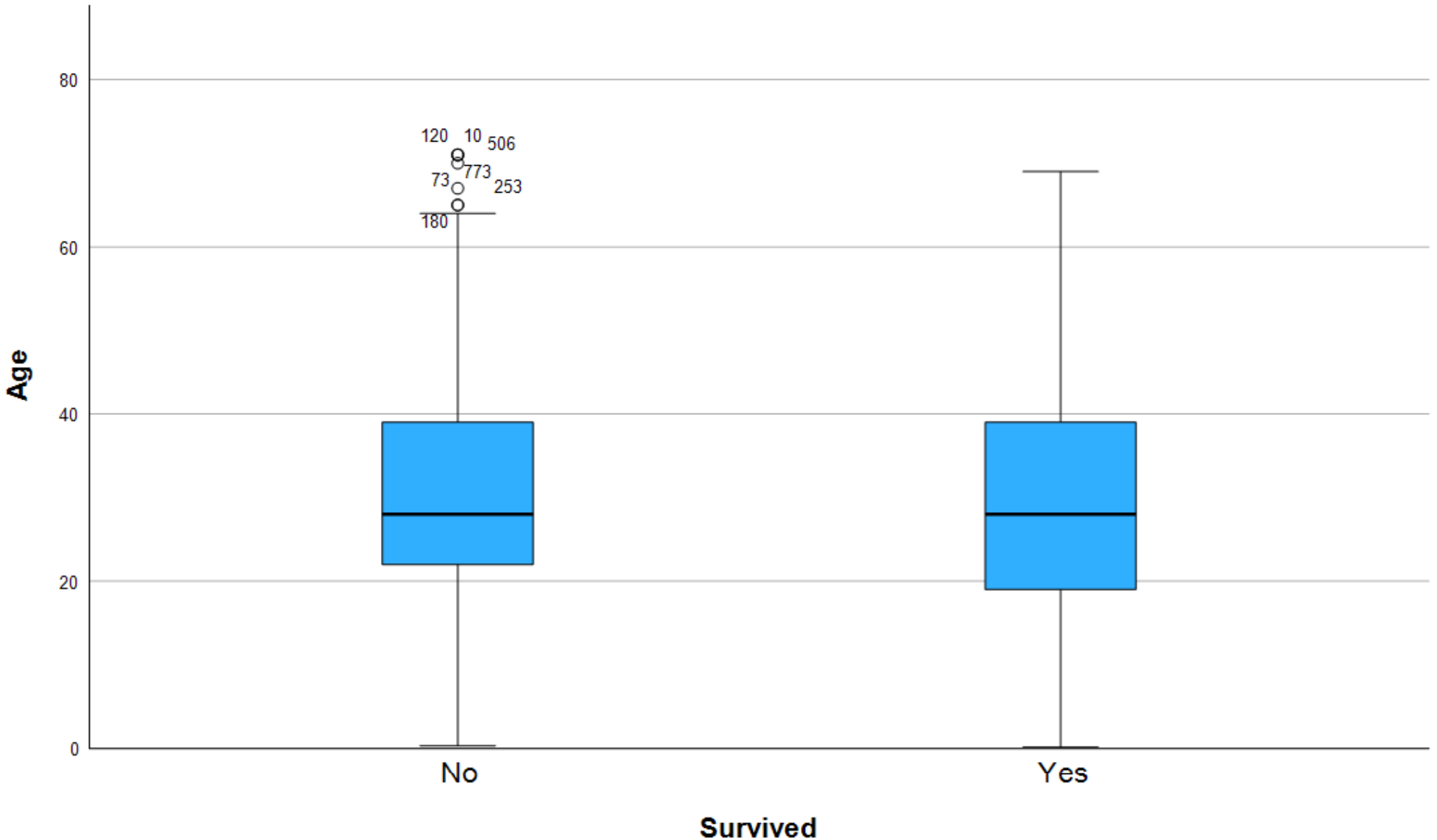


Figure 34: Boxplot showing ages of deaths and survivors

Speaker notes

Speaker notes

For continuous data, I recommend boxplots. Here are the box plots of ages.

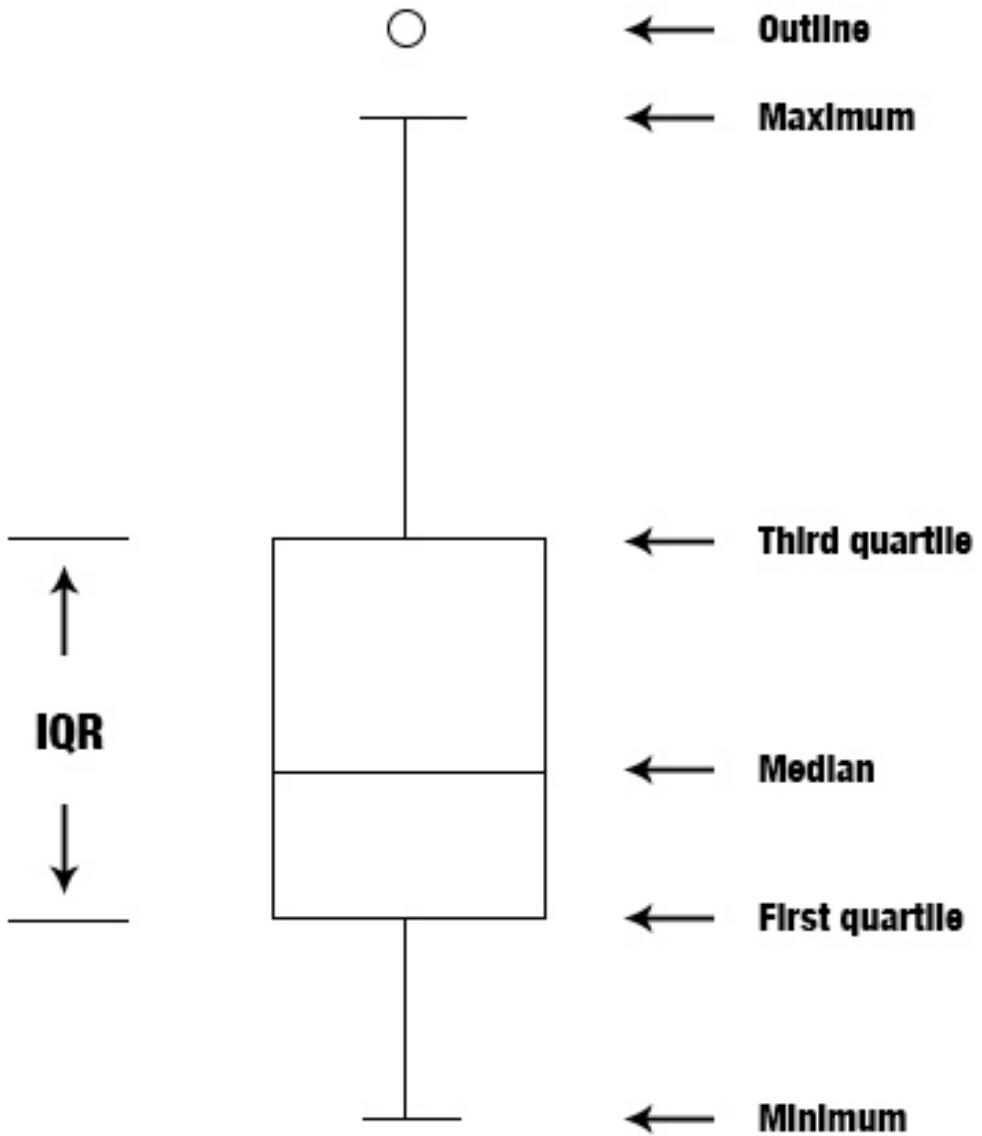


Figure 35: Annotated boxplot

Speaker notes

Speaker notes

If you have not seen a boxplot before, here is a nice schematic. The boxplot uses five numbers to summarize the data: the minimum value, the lower quartile, the median, the upper quartile and the maximum. The size of the box represents the interquartile range, the middle 50% of the data. If there are outliers, points that stray too far from the middle 50%, these are highlighted individually.

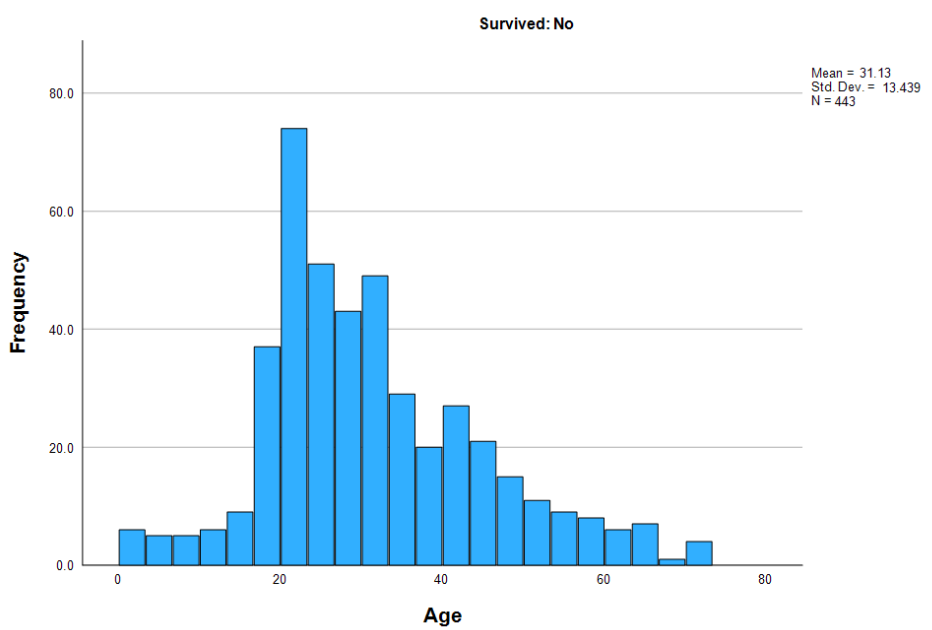


Figure 36: Histogram of ages of passengers that died

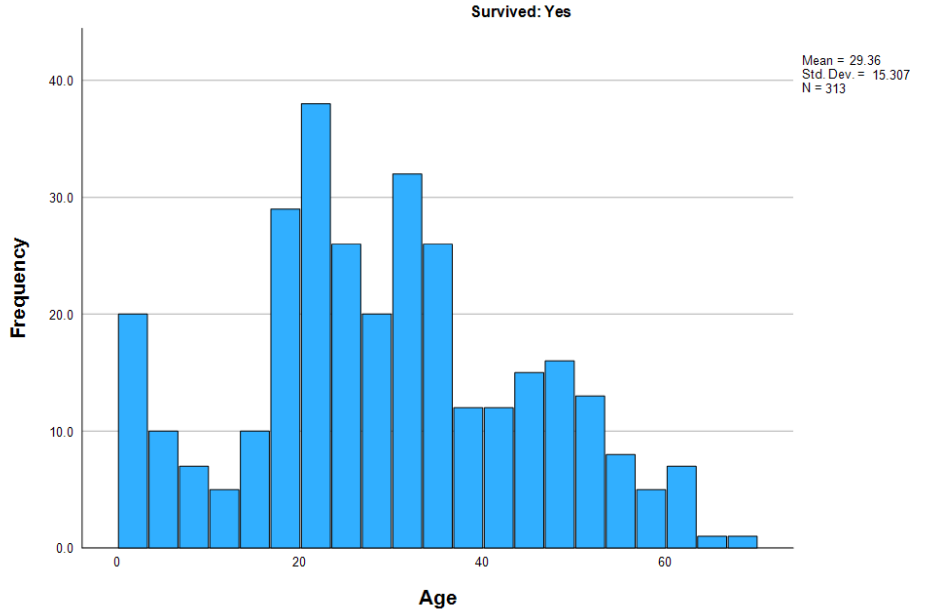


Figure 37: Histogram of ages of passengers that survived

Speaker notes

Speaker notes

Here are two histograms. They are great for looking at individual patterns, but do not allow for an easy comparison between two groups.

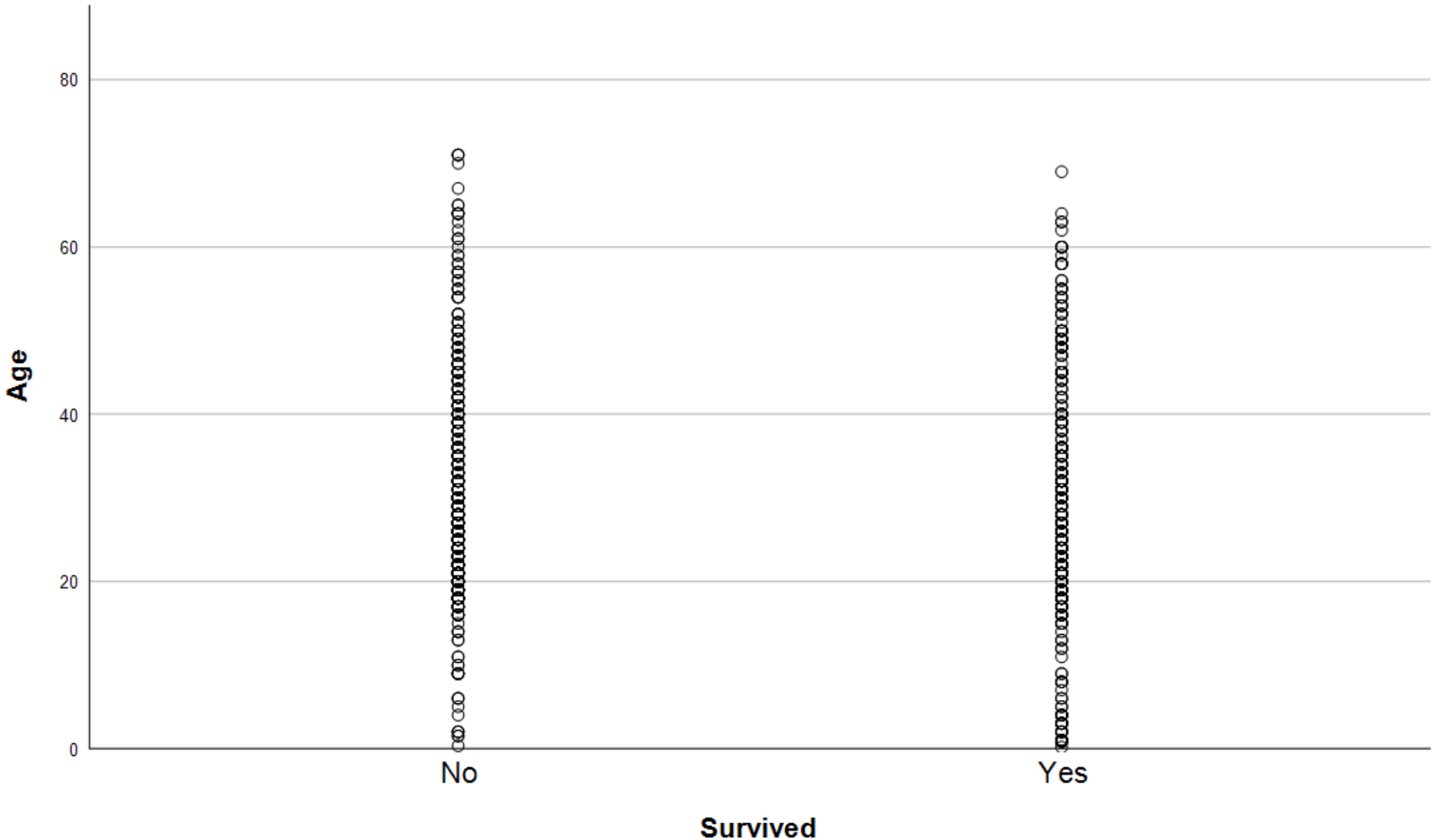


Figure 38: SCatter plot of age vs survived

Speaker notes

Speaker notes

In many cases I like plotting the individual data point. It doesn’t help too much here because you have hundreds of ages.

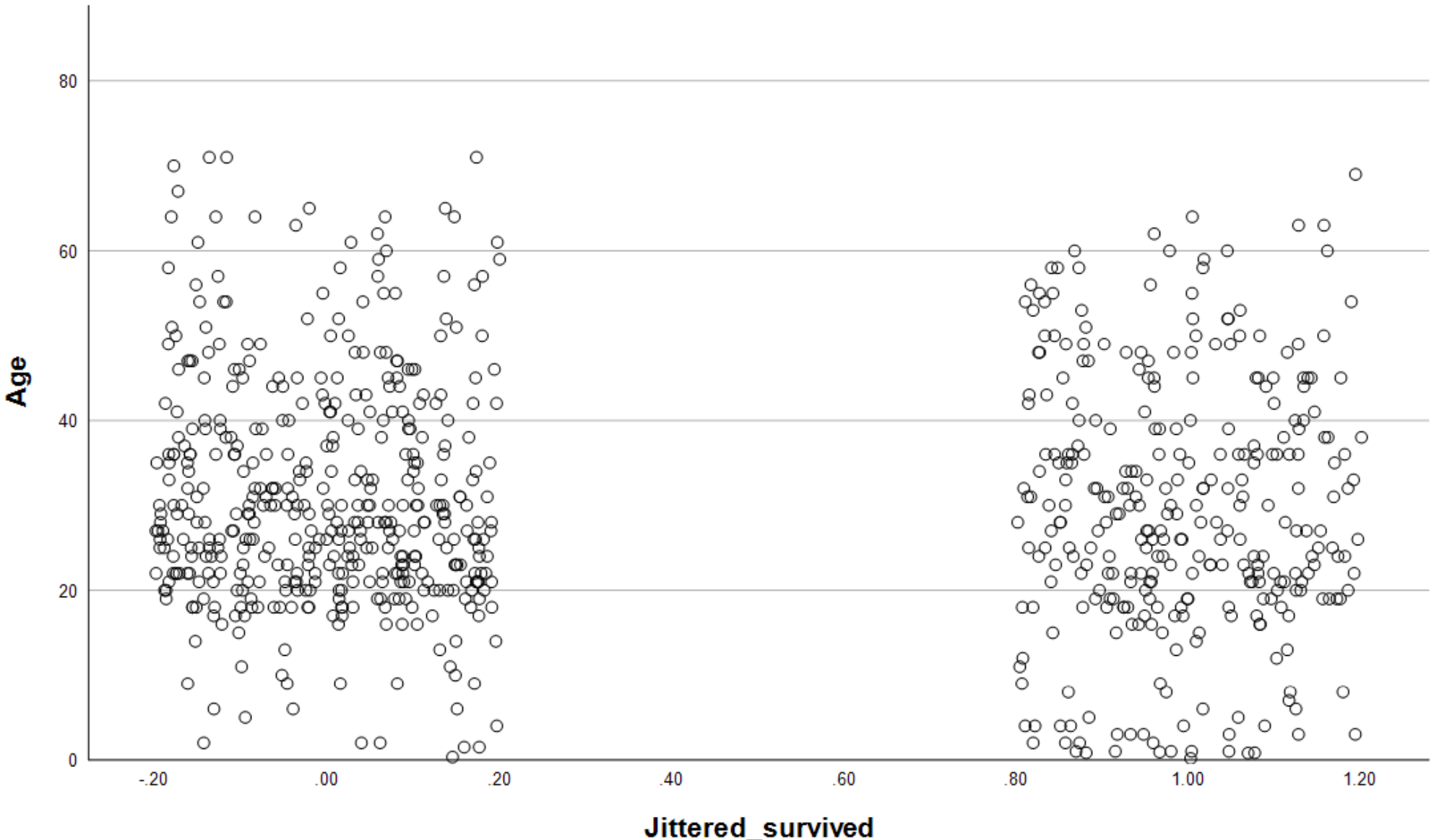
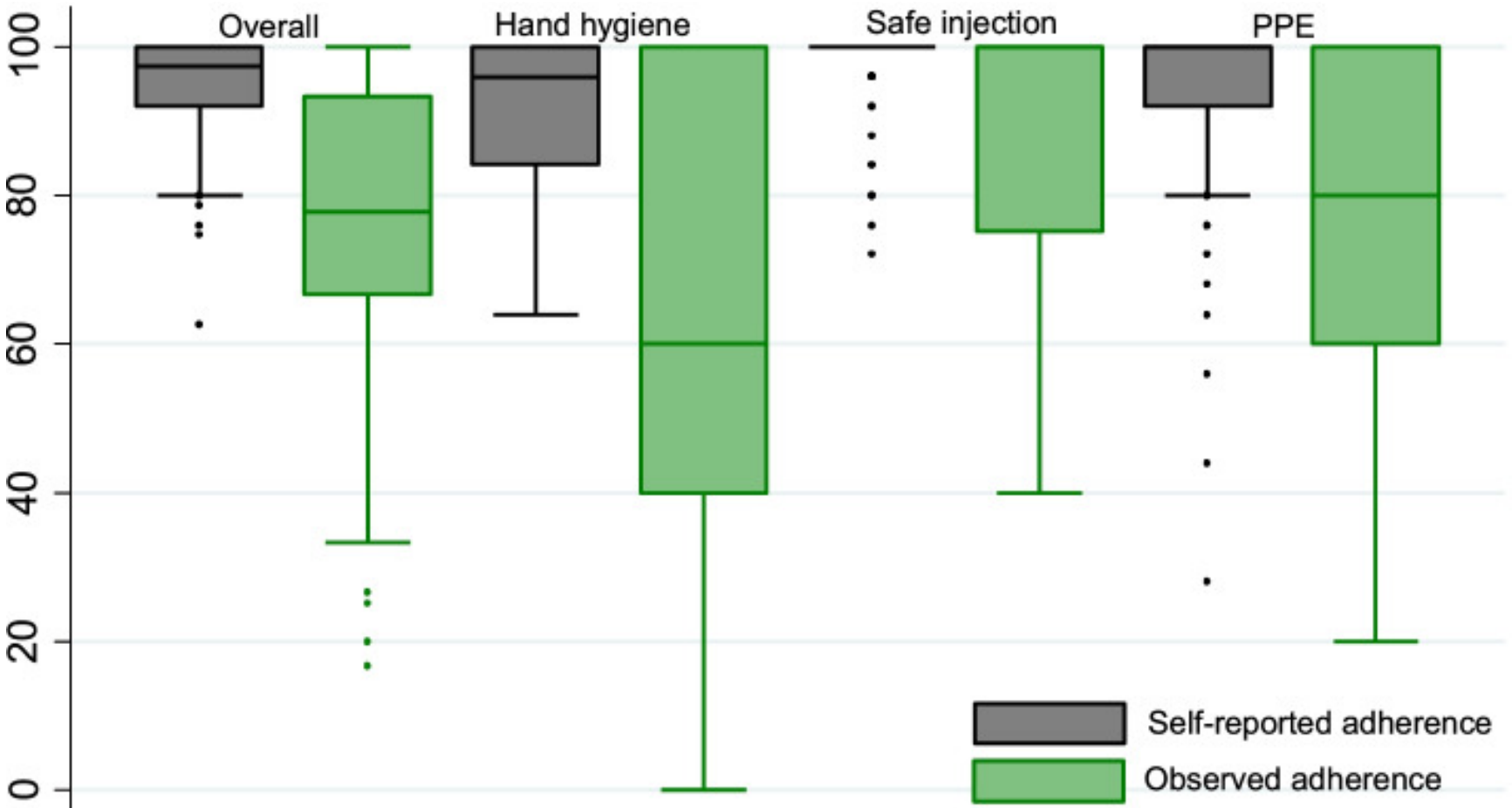


Figure 39: Scatter plot of age vs survived with jittering

Speaker notes

Speaker notes

One method that researchers will use when trying to display hundreds of observations is jittering. Jittering is randomly shifting the data a bit to the left or right. I’m not a big fan of jittering, but sometimes it does help.



	Overall (Mean±S.D)	Subcategory (Mean±S.D)		
		Hand hygiene	Safe injection	PPE
Self-report	94.9±6.5	92.3±8.6	98.1±5.1	94.3±10.8
Observation	76.8±19.6	68.2±30.5	90.2±15.7	70.0±24.5

Figure 40: Self reported versus observed adherence

Speaker notes

Comment on these boxplots

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9310710/

Article feedback by type of article

Speaker notes

Speaker notes

Comment on these boxplots

Wikipedia

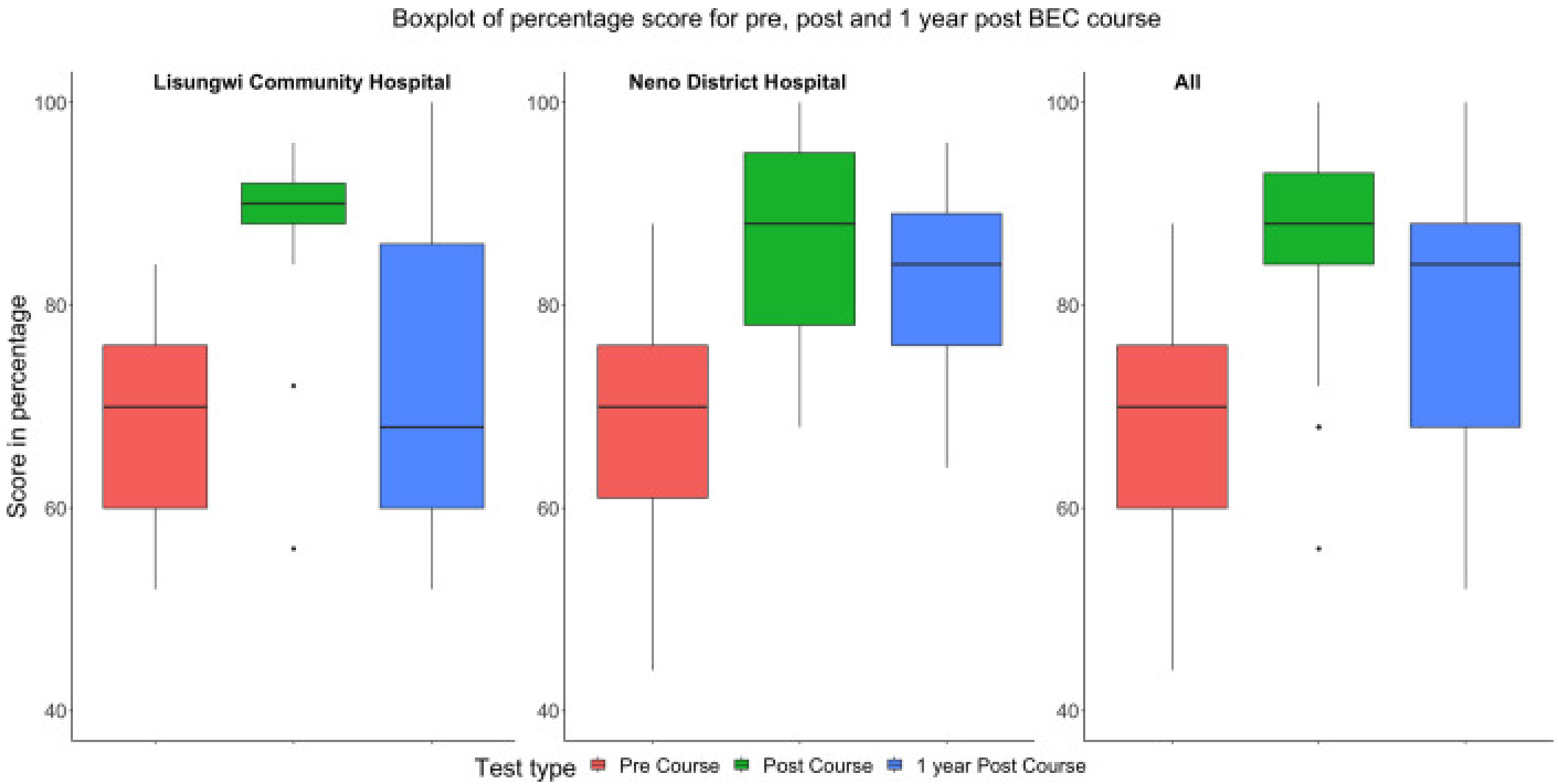


Figure 41: Emergency care course test scores

Speaker notes

Speaker notes

Comment on these boxplots

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9901771/

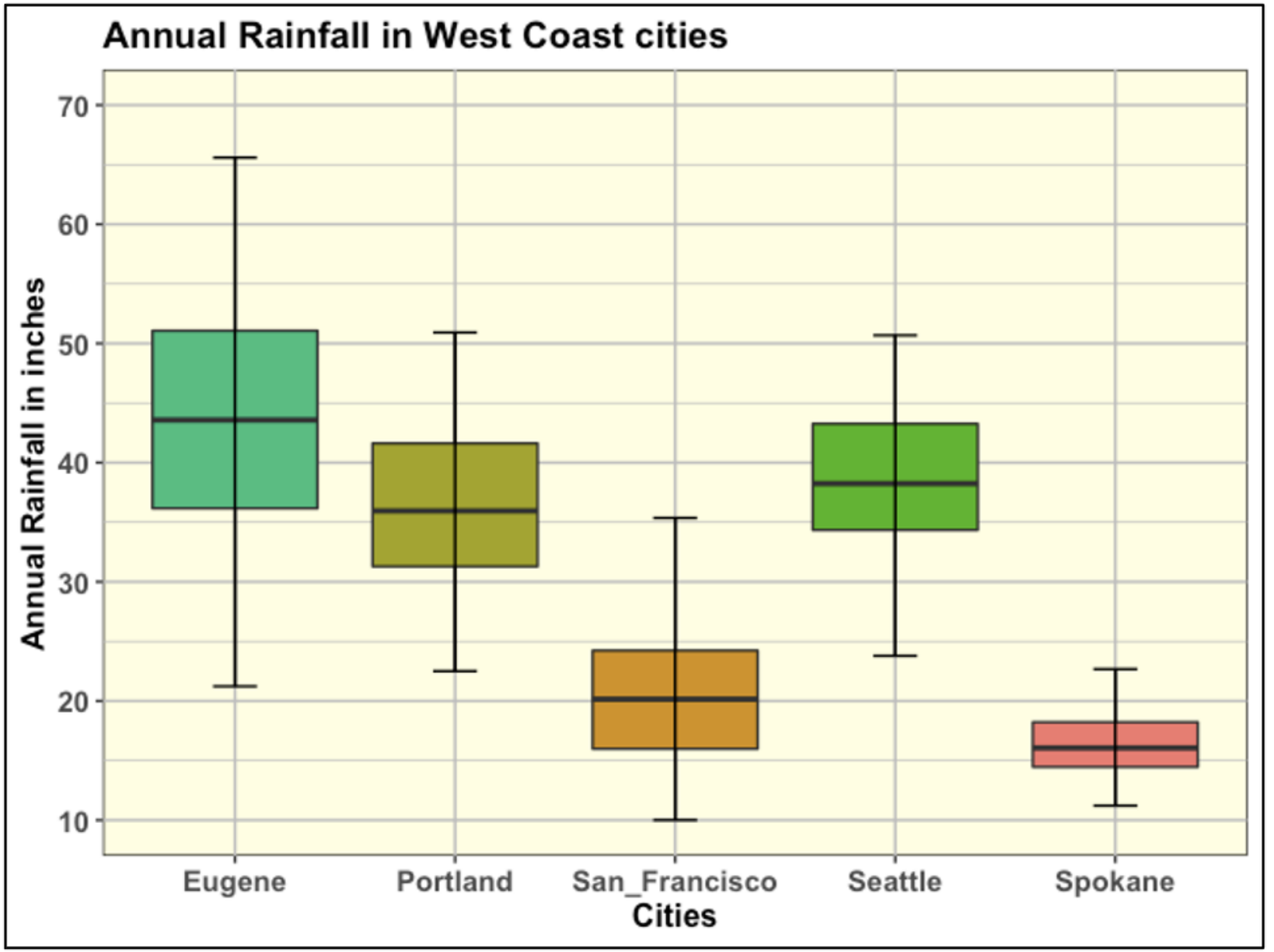


Figure 42: Annual rainfall in selected cities

Speaker notes

Comment on these boxplots

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9310710/

Which visualization to choose?

- Categorical data
  - Use a table
  - Minimize distance of key comparisons
- Continuous data
  - Small datasets: plot all the data
  - Large datasets: boxplots

Speaker notes

I have thrown a bunch of different visualizations at you. Let me try to sort them out.

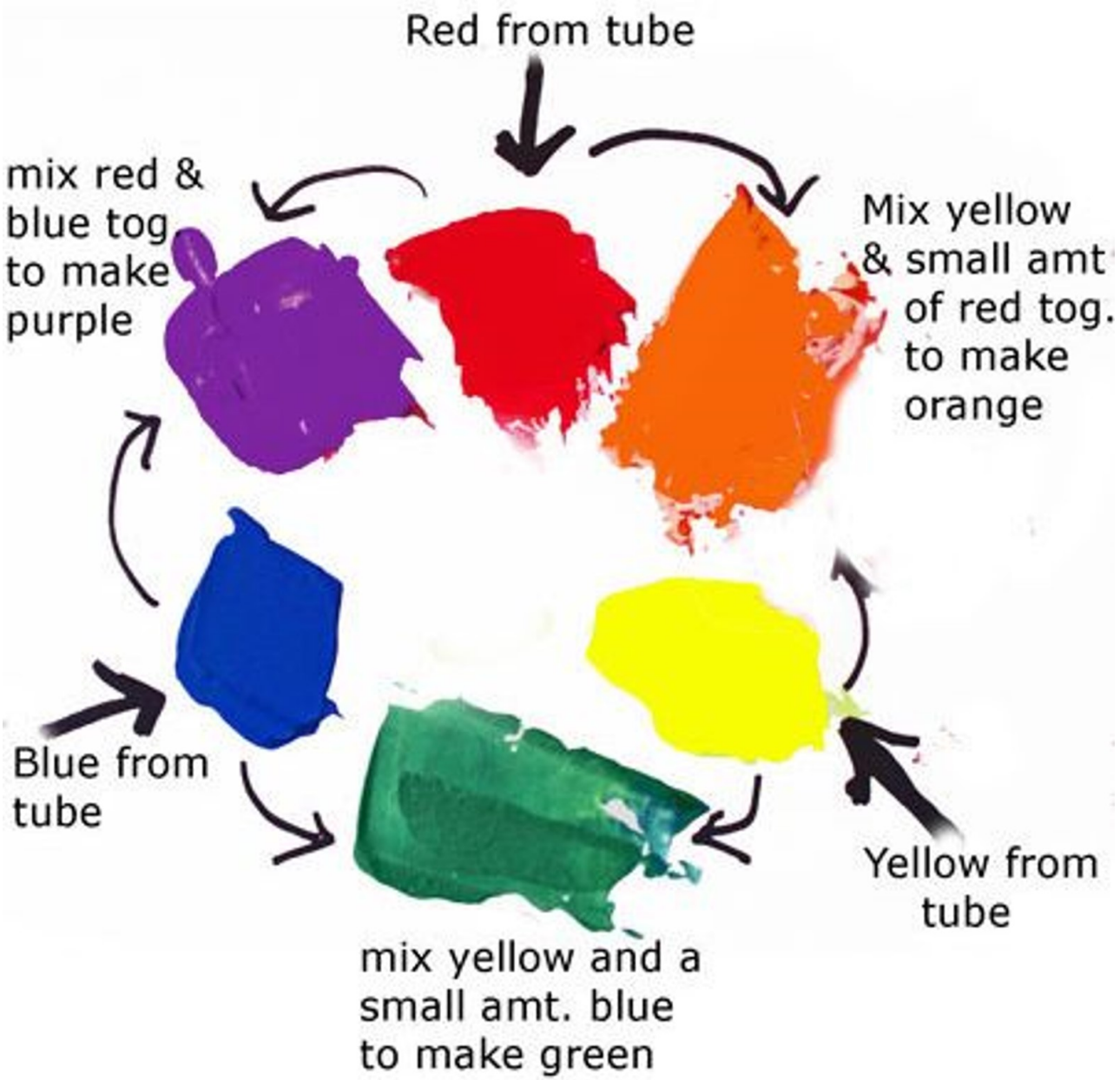


Figure 43: Color combinations

Speaker notes

I want to spend a bit of time talking about color. Color is an important feature of visualization and one that is often poorly handled.

Let’s start by saying that the way you learned colors as a child is all wrong for computer graphics. The basic system of colors that you use with crayons and paints is that there are three primary colors, red, yellow, and blue and the combinations of any of these two colors produces a secondary color: orange, green, or purple.

The RGB color system

- #rrggbb format
  - #000000 is pure black
  - #FFFFFF is pure white
  - #FF0000 is pure red
  - #00FF00 is pure green
  - #0000FF is pure blue
- You can mix and match to get 16,777,216 colors
  - #800080 is purple, #FF69B4 is pink, #40E0D0 is turquoise

Speaker notes

On the computer, you use the RGB color system. This represents six hexadecimal digits. Hexidecimal means base sixteen and the digits go 0, 1, 2, up to 9, then switches to letters: A for 10, B for 11, C for 12, D for 13, and E for 14 and F for 16. The smallest hexidecimal pair is 00 and the largest is FF.

Three pairs of double 0 represent no red, no green, and no blue. It is the total absence of color: black.

Thre pairs of double FF represents the maximum values for red, green, and blue. This is pure white.

A double FF for red and zeros everywhere else gives pure red.

Now you can mix and match these.

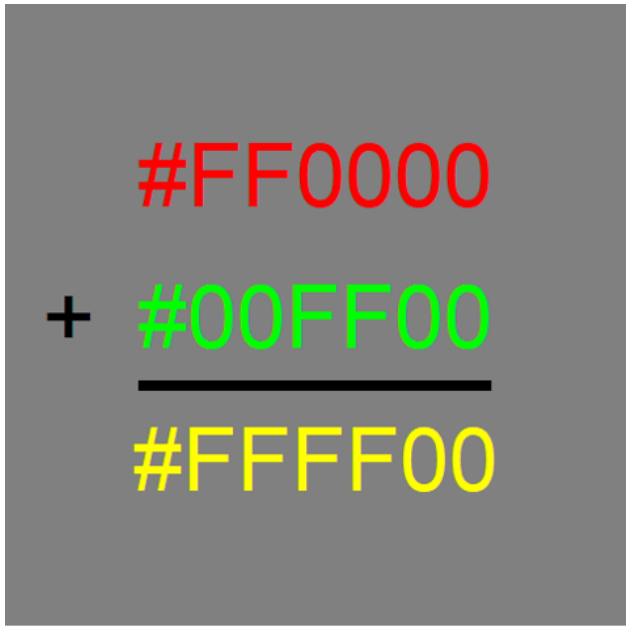


Figure 44: Red plus green equals yellow

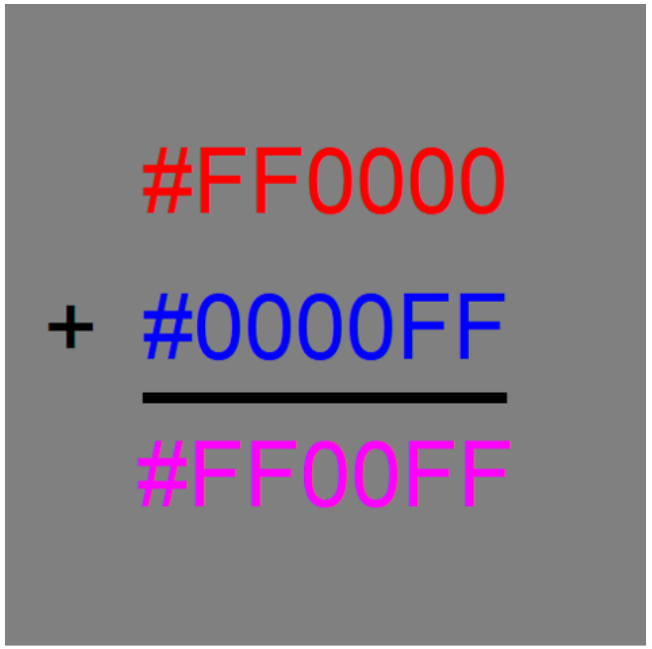


Figure 45: Red plus blue equals magenta

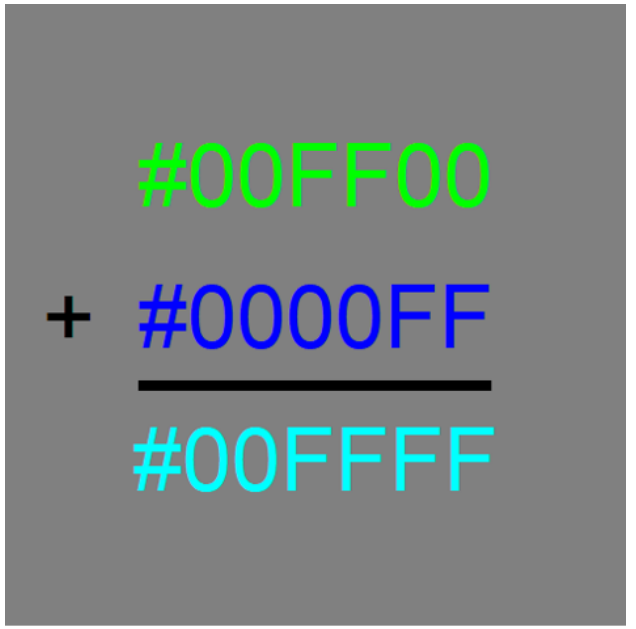


Figure 46: Green plus blue equals cyan

Speaker notes

In the rgb system, colors add up in a way that you never learned in kindergarten. There are new colors like magenta and cyan. The other thing to notice is that blending two colors makes things a bit brighter. This is most obvious in the combination of red and green to produce yellow, but you can see it in the other color combinations as well.

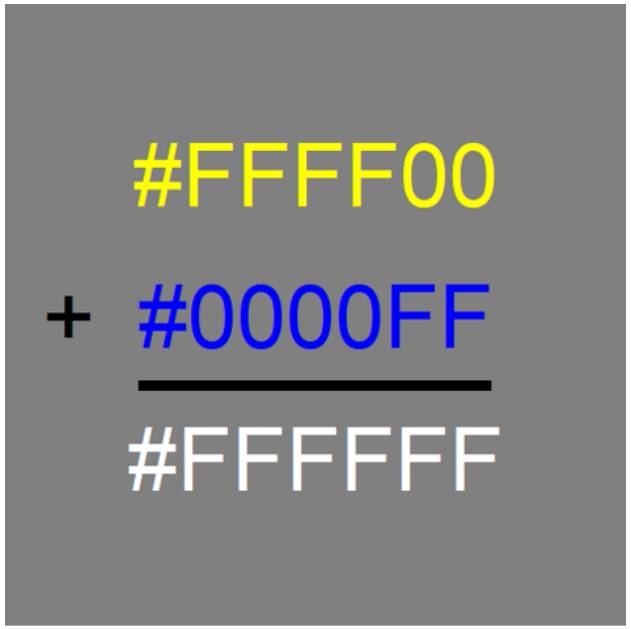


Figure 47: Yellow plus blue equals white

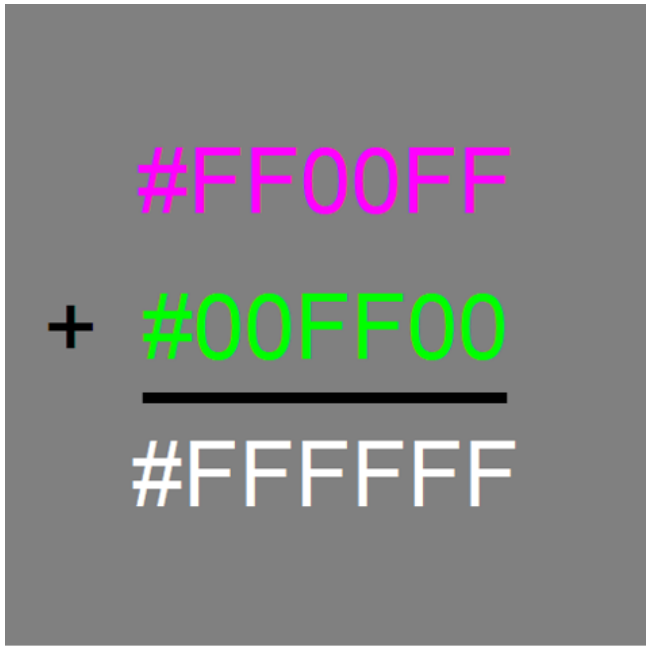


Figure 48: Magenta plus green equals white

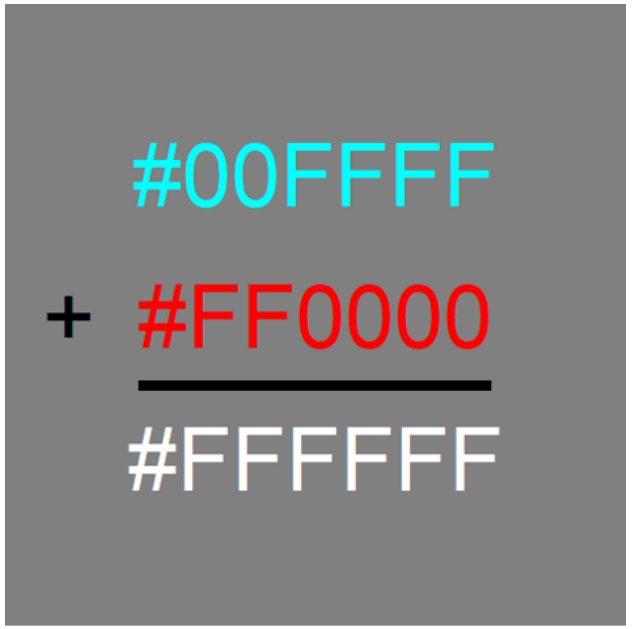


Figure 49: Cyan plus red equals white

Speaker notes

Keep adding together and you get all the way to the brightest color, white. Try this with your kindergarten crayons and you would get a dark muddy brown. The rgb system is additive.

The color cube

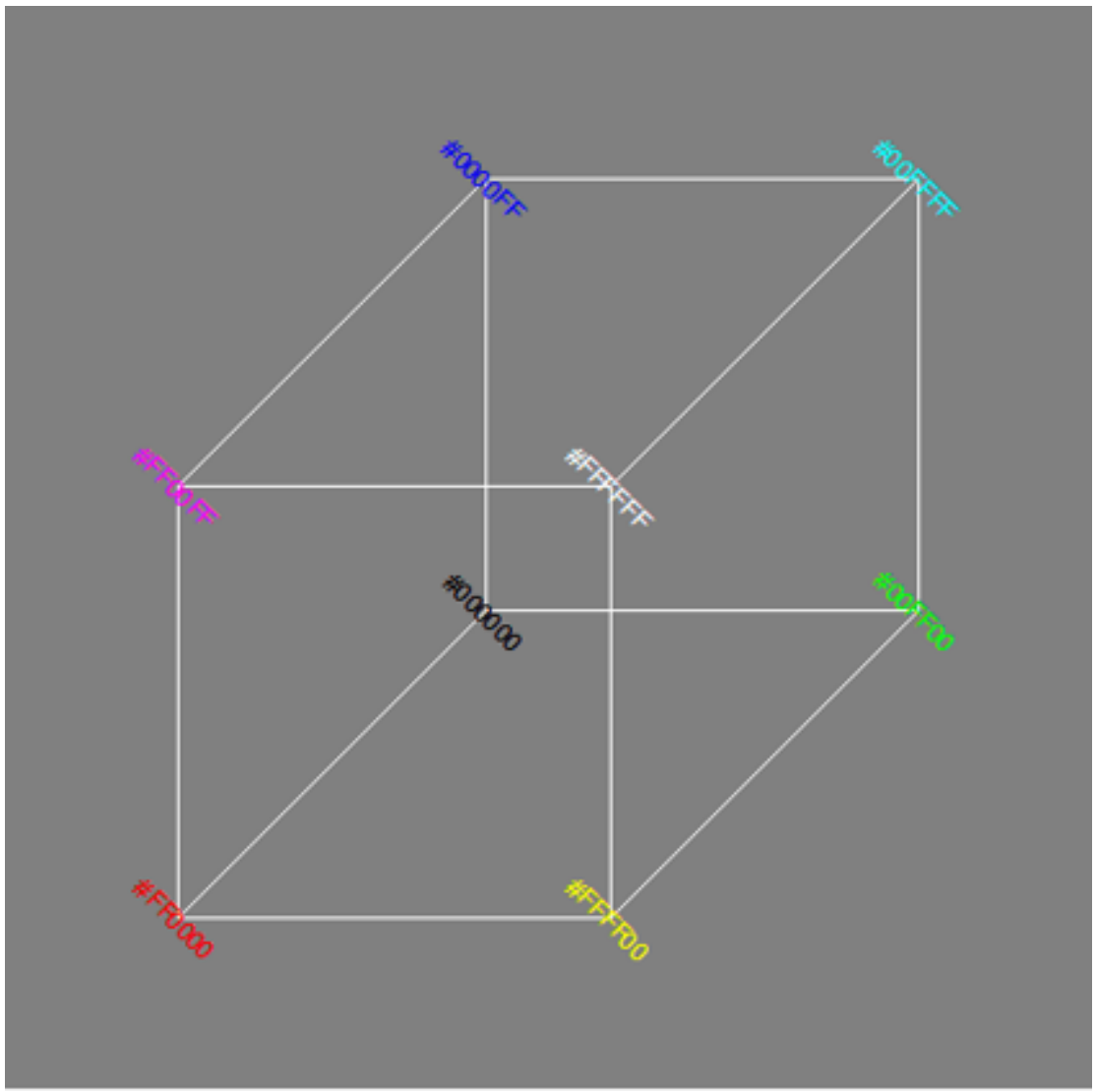


Figure 50: Illustration of the color cube

Speaker notes

Here are the basic rgb colors, along with white and black, arranged in a cube.

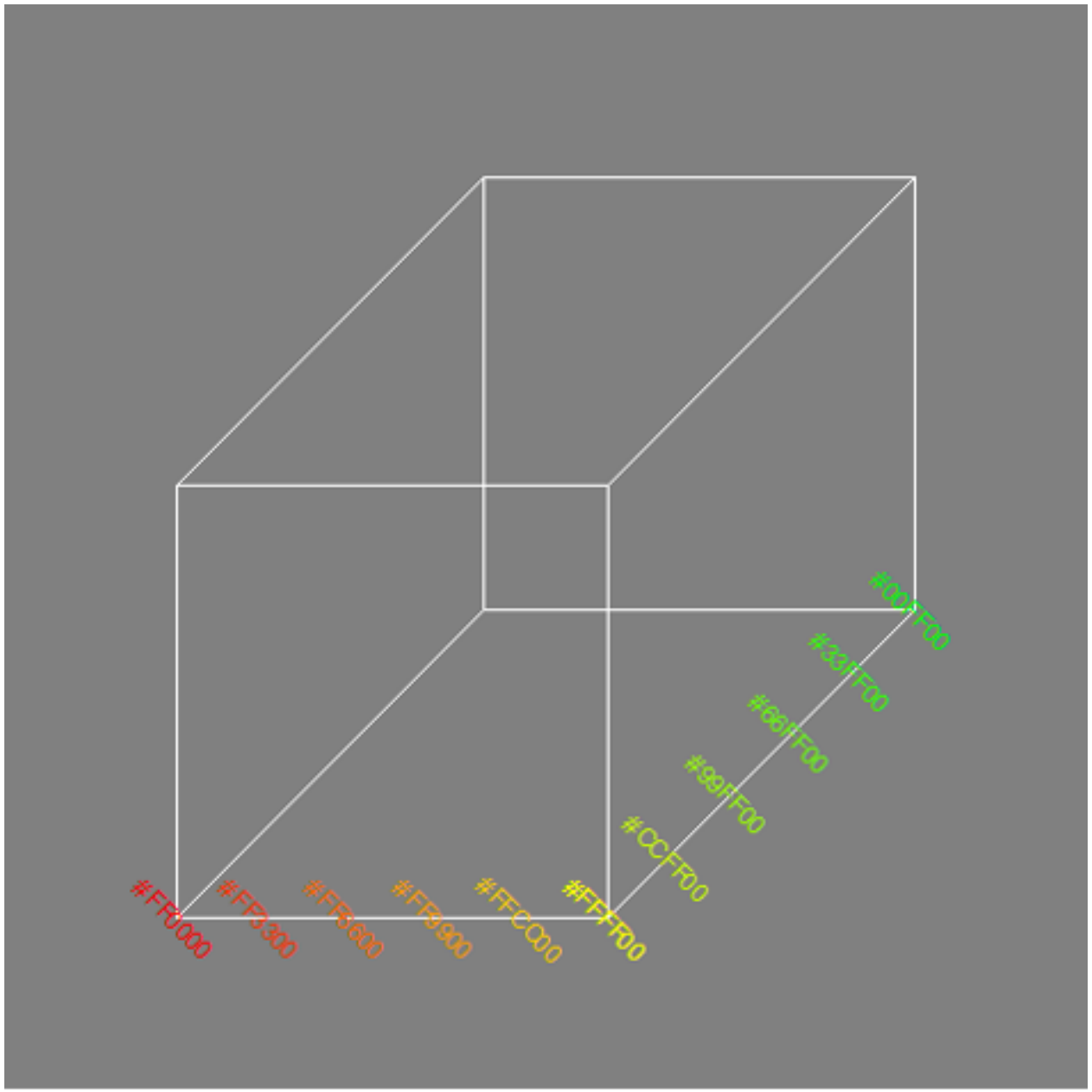


Figure 51: The red to green gradient on the color cube

Speaker notes

Speaker notes

Once you see the colors on a cube, you can figure out all sorts of interesting color gradients. Here’s a commonly used gradient that starts at red, transitions to yellow in the middle, and then ends up at green.

Imagine continuing this route through cyan (the top right corner), blue (the top corner in back), magenta (the top left corner), then back down to red. Visualize this path (it looks like a hexagon from this perspective) turned into a circle.

Rainbow

Speaker notes

This circle is called a rainbow, but it’s not quite accurate. A true rainbow doesn’t have magenta and cyan, and it includes two colors, indigo and violet, that aren’t found on this circle. But so many people call this the rainbow, so that’s what I’ll call it.

A lot of people use the rainbow circle, but it has many problems.

The color cylinder

Figure 52: Color cylinder

Speaker notes

Now extend this circle in two directions. Move the circle lower to create darker colors. The very bottom represents black, the darkest color.

Then draw the circle in towards the center. You get increasingly bright the closer you get to the middle. The very center of the circle is white, the brighest color.

Figure 53: Various foreground and background color combinations

Speaker notes

The first issue with the rainbow is that the colors come across as a bit harsh, especially when juxtaposed. This image shows a red foreground and green background in the top bar, a red foreground and a blue background in the second bar, and so forth. All the foreground and background colors are shown here.

The color combinations seem to vibrate. At times, you might want an intensity like this. But most of the time, I think that this is just too harsh.

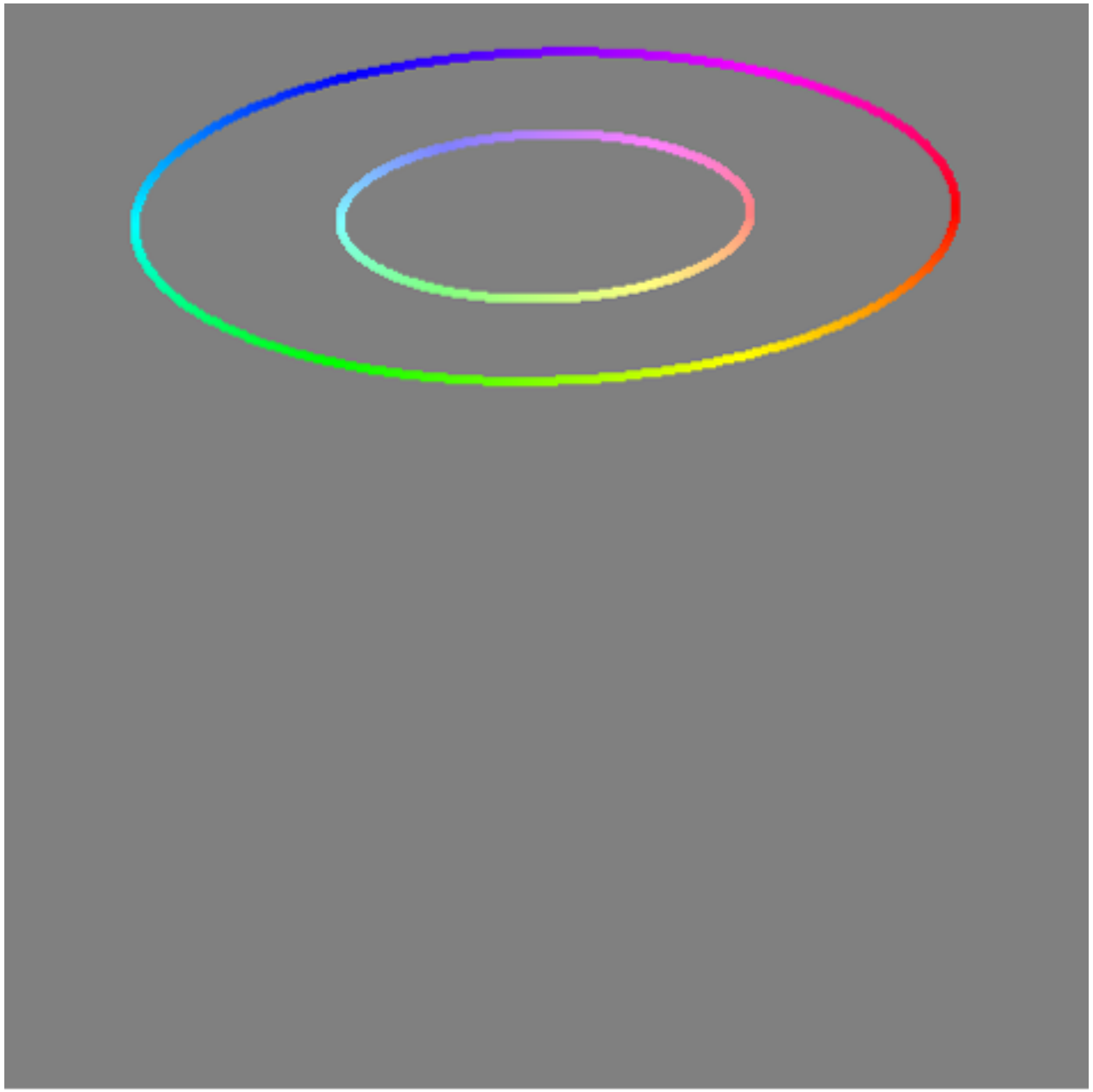


Figure 54: A brighter version of the rainbow

Speaker notes

There are two ways to make the colors less harsh. First, try moving closer to the center of the color cylinder. These produce brighter colors. To my eye, these colors look closer to a pastel version.

Darker rainbow

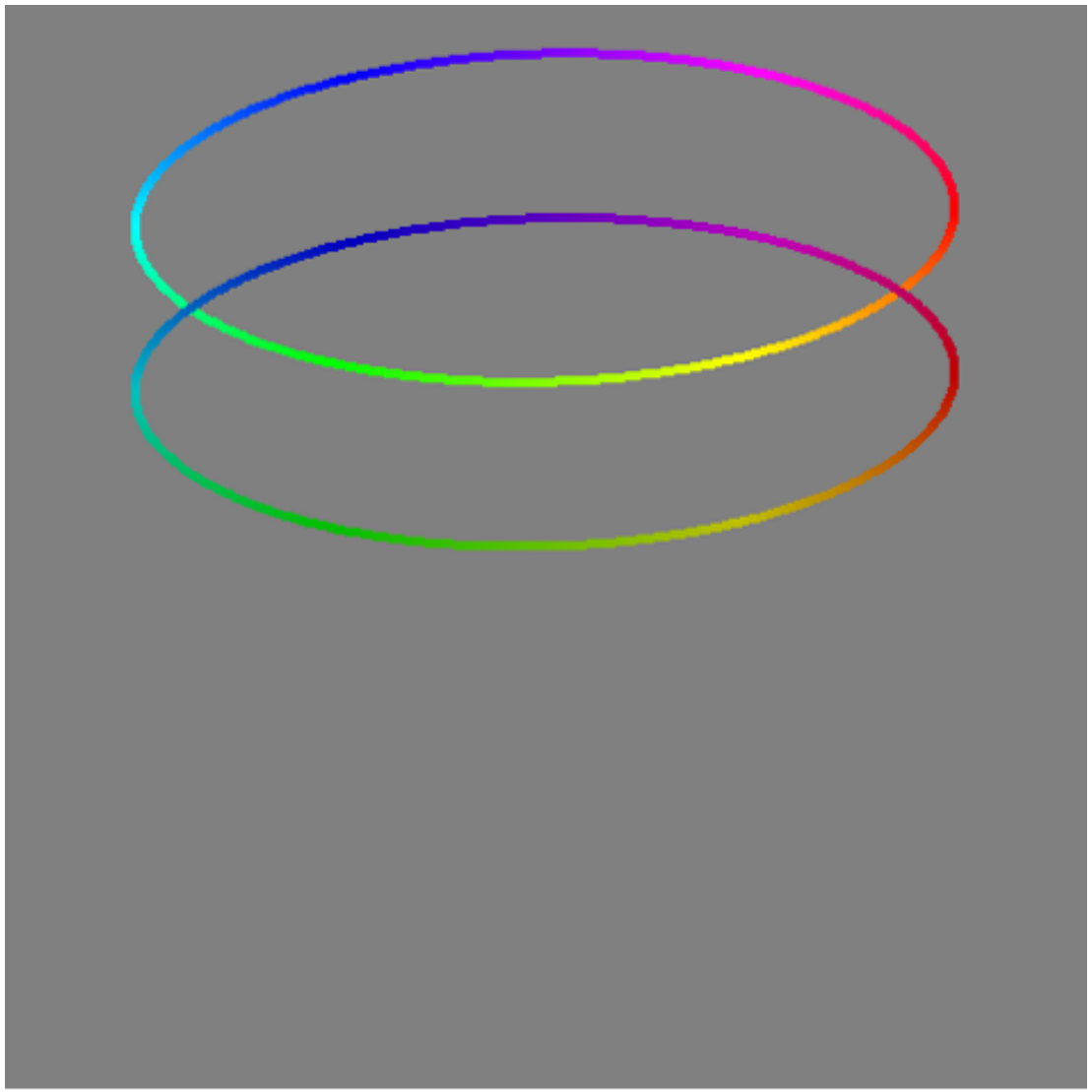


Figure 55: A darker version of the rainbow

Speaker notes

A second way to make the colors less harsh is to move down on the cylinder.

Figure 56: Color combinations using darker foregrounds and lighter backgrounds

Speaker notes

Speaker notes

Here are the combinations of foregrounds and backgrounds where all the foregrounds have been made a bit darker and all the backgrounds have been made a bit lighter.

Most visualization software makes it easy to lighten or darken your colors. They offer alternatives to the pure rainbow colors such as “light green” and “dark red”.

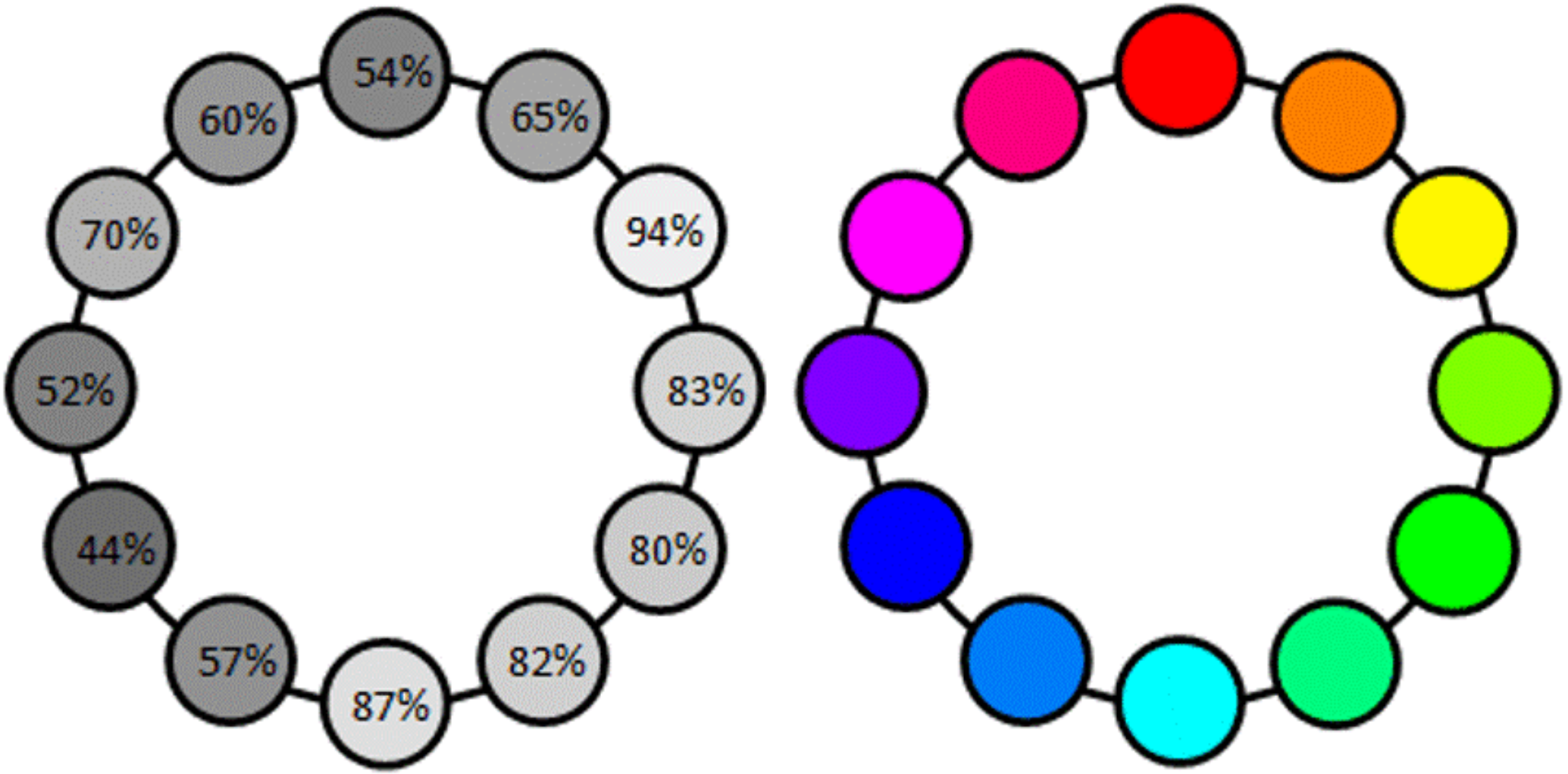


Figure 57: DIffering luminance values of the rainbow

Speaker notes

Another problem with the rainbow is that the colors have different levels of brightness. The technical term is luminance, and this image shows that yellow has the highest luminance and blue has the lowest.

This image comes from the excellent website, workwithcolor.com.

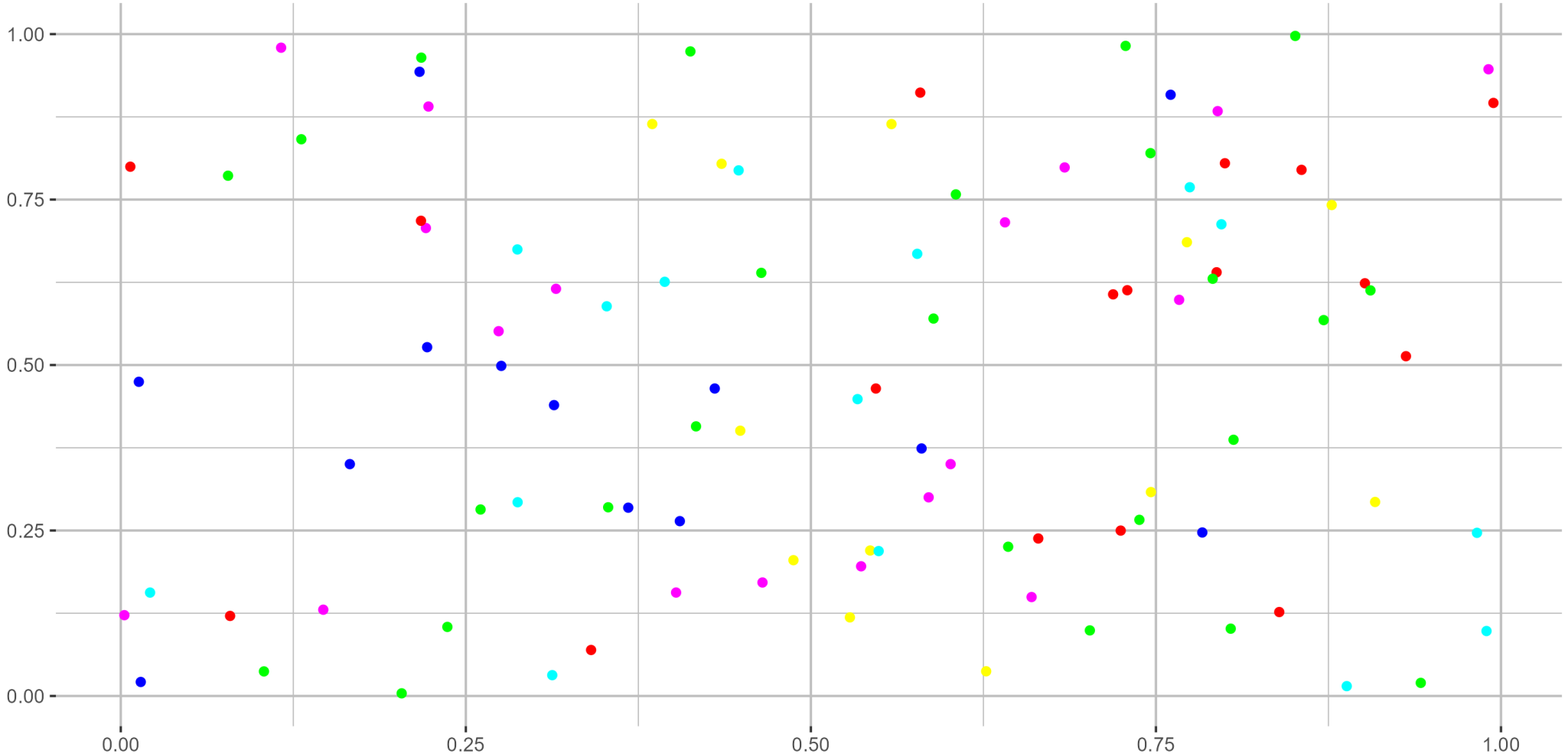


Figure 58: Rainbow colors on a white background

Speaker notes

Because yellow very bright, it has a poor contrast against a white background.

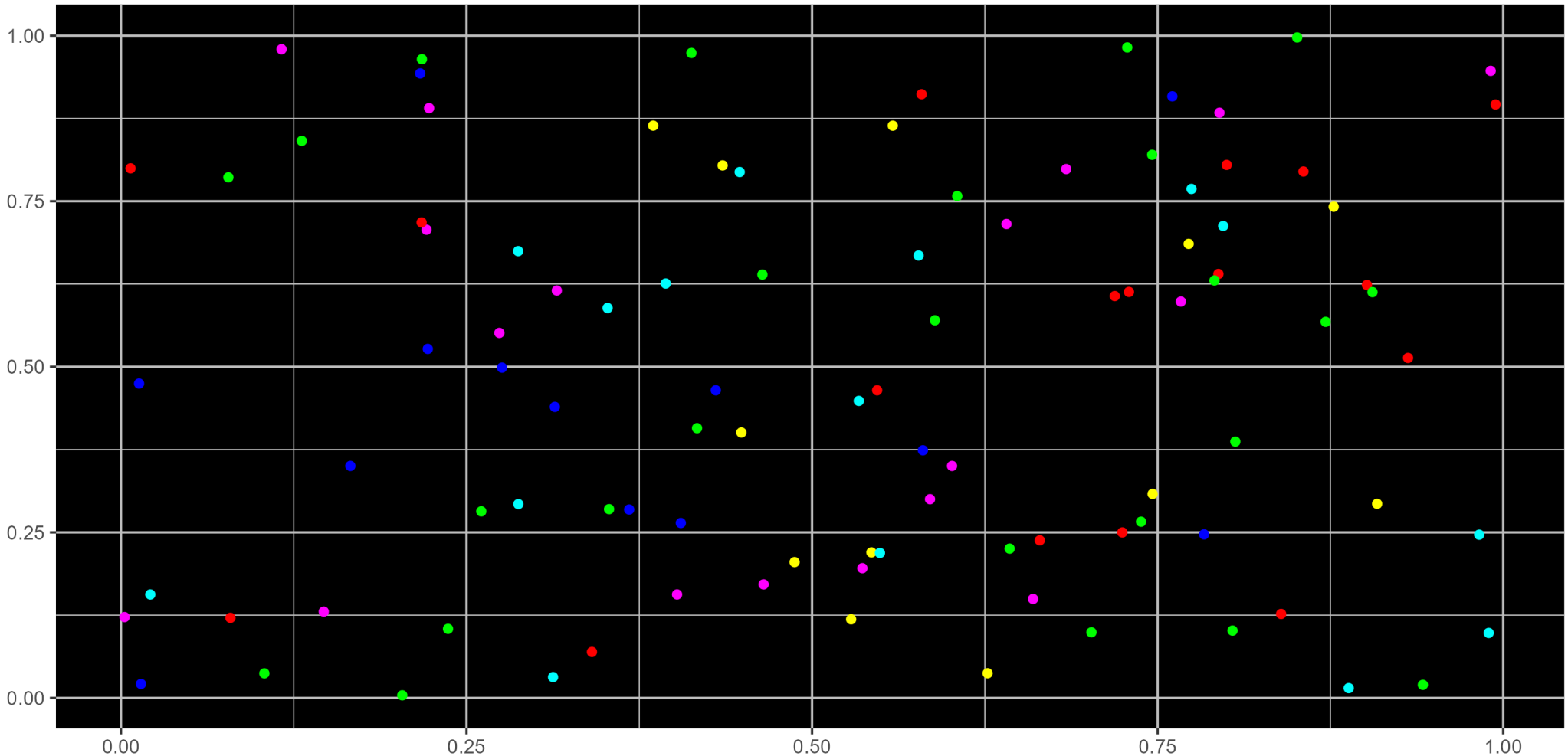


Figure 59: Rainbow colors on a black background

Speaker notes

Because blue is very dark, it has a poor contrast against a black background.

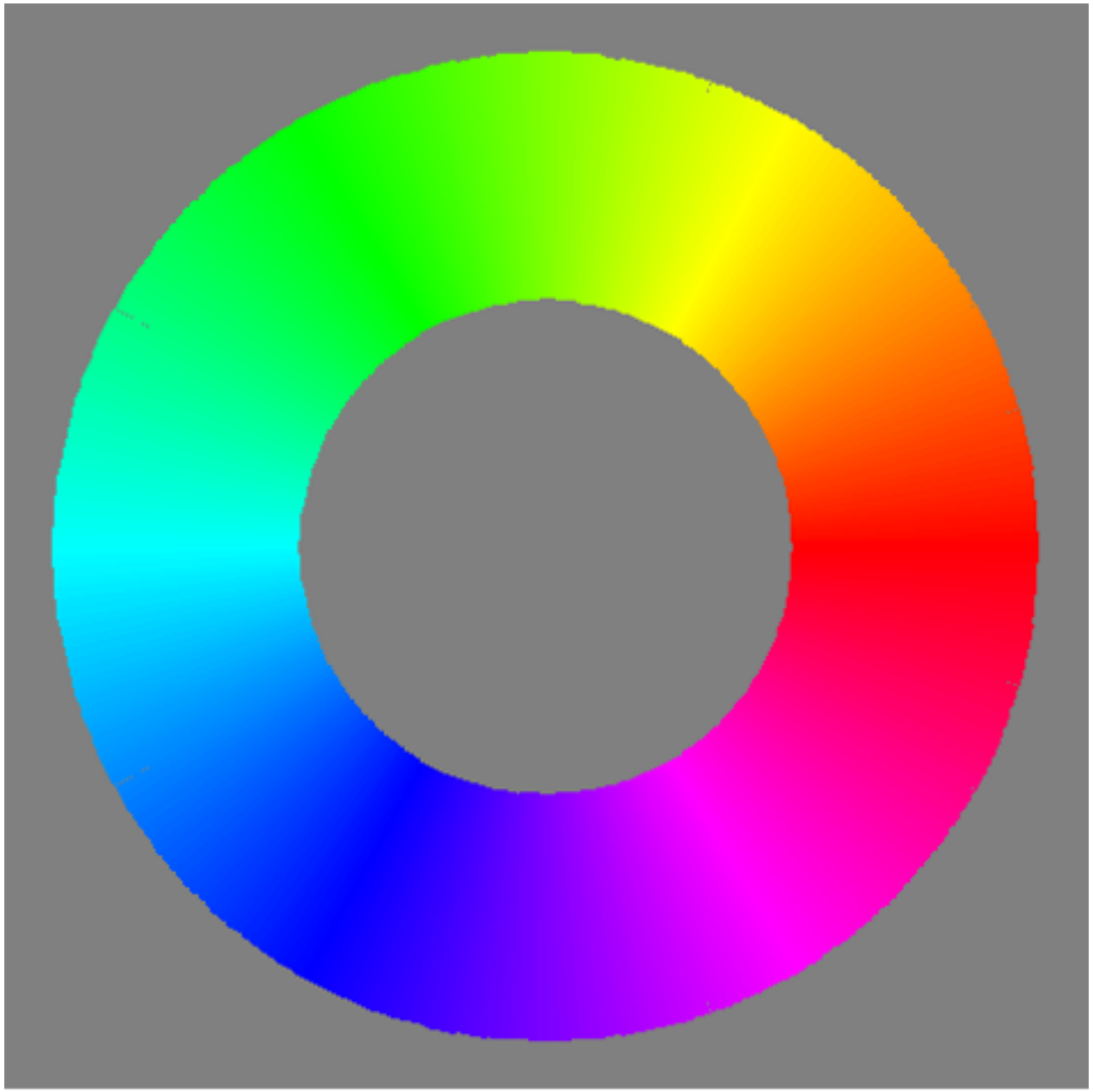


Figure 60: Rainbow colors showing a banding effect

Speaker notes

There is one more problem with the rainbow. It has a banding effect because some of the transitions are sudden and others are more gradual. This producing a banding effect at yellow, magenta, and cyan.

So what colors do I recommend?

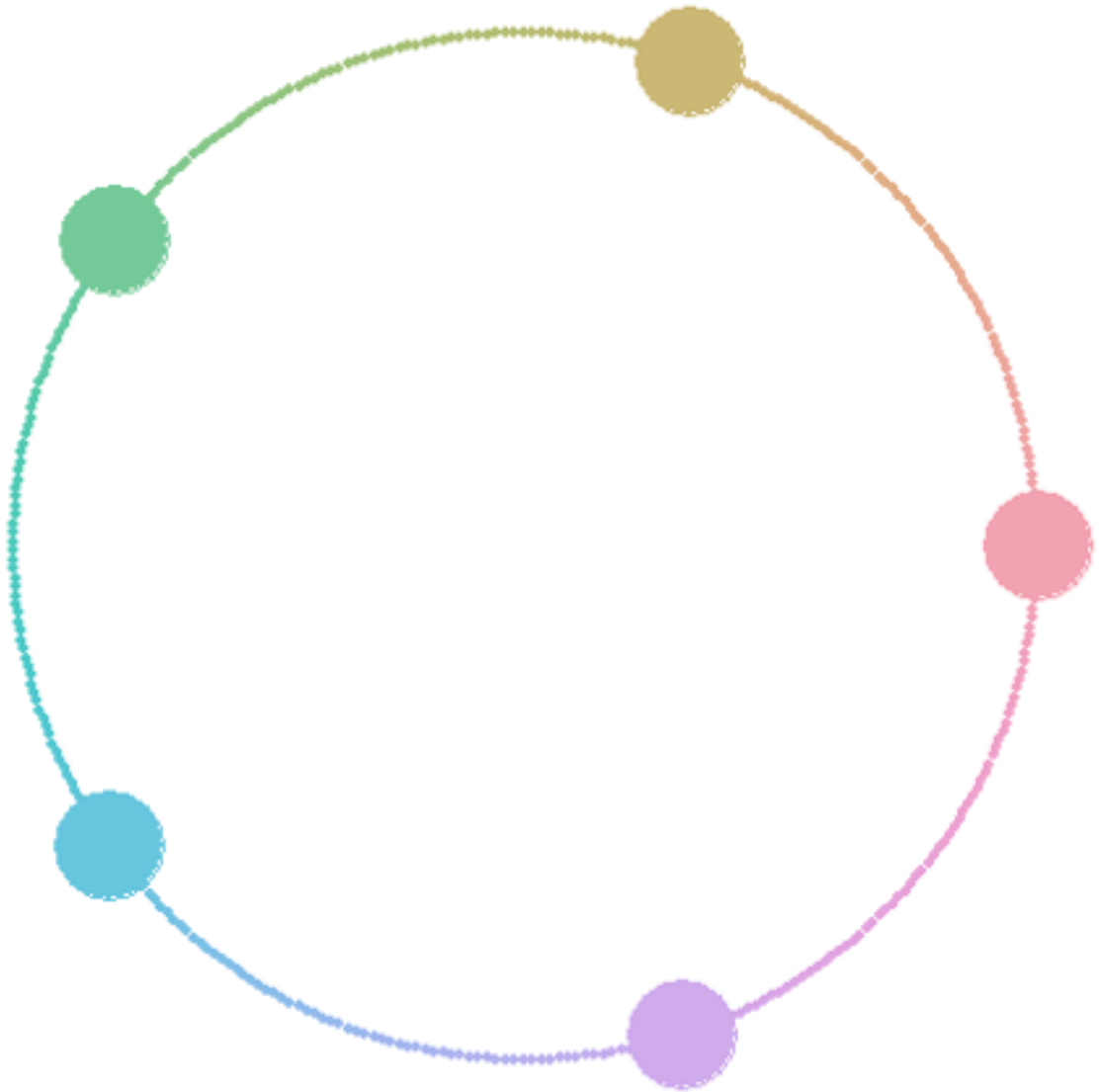
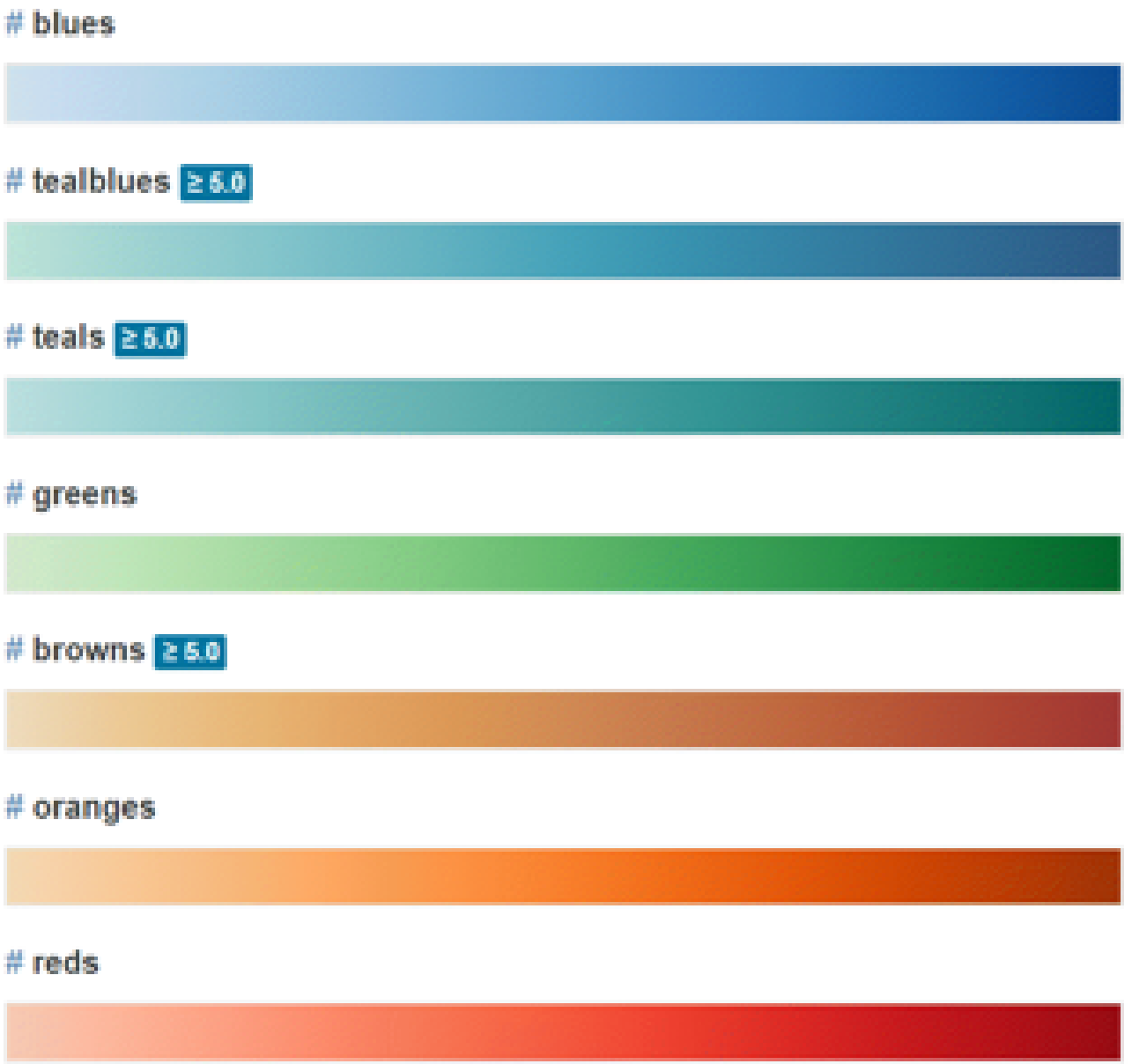


Figure 61: One good set of color choices for nominal data

Speaker notes

For nominal data, you want to colors to be about the same brightness, but not too close to each other.



Examples of light to dark gradients

Speaker notes

Speaker notes

For ordinal or continuous data, you have two choices. The first is a gradient from light to dark. Depending on your background color, this will either emphasize the low end of the scale or the high end of the scale.

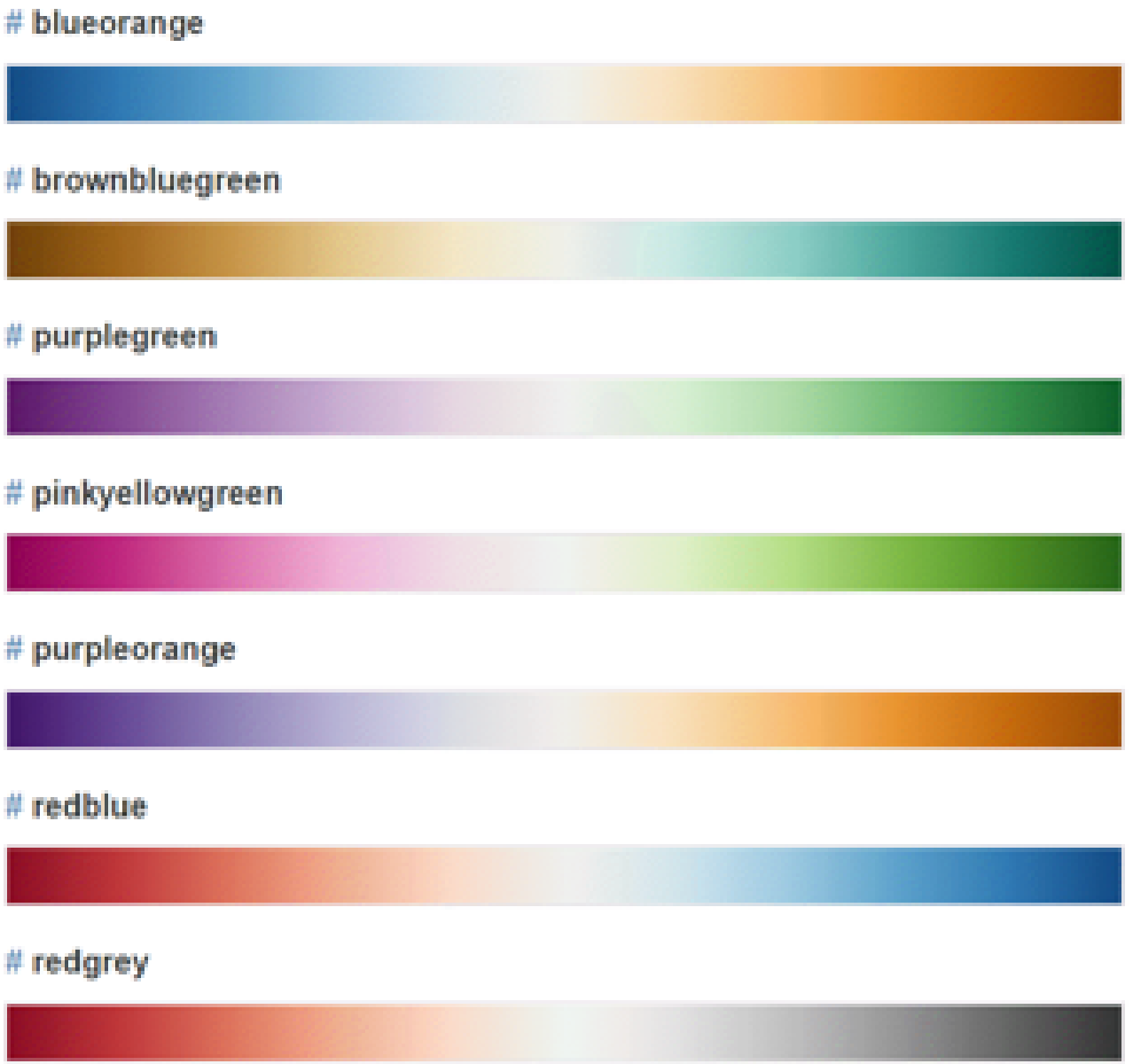


Figure 62: Examples of diverging gradients

Speaker notes

Speaker notes

A second choice is a diverging gradient which has a different dark color at either end and white or a brighter color in the middle. This will emphasize both extremes, presuming that you place the colors against a light background.

Color blindness

- Up to 10% of your audience is color blind
  - Most common: red-green
- Suggestions
  - Use alternate cues (shape, shading)
  - Test your image
  - Find color blind friendly palettes.

Speaker notes

Speaker notes

There are more than a few people who has difficulty distinguishing colors. Color blind people can still distinguish some colors, but others cause problems. The most common problem is red-green color blindness.

You can use alternate visual clues to supplement the codes that colors represent. While I earlier advocated for using similar levels of brightness, some variation in brightness, even a small amount, can help. You can also change the shape of data points along with the color.

There are a number of websites that will simulate what your visualization will look like for various types of color blindness.

You can also find various color combinations that are easier for color blind people to distinguish.

#4 TOO MANY COLORS

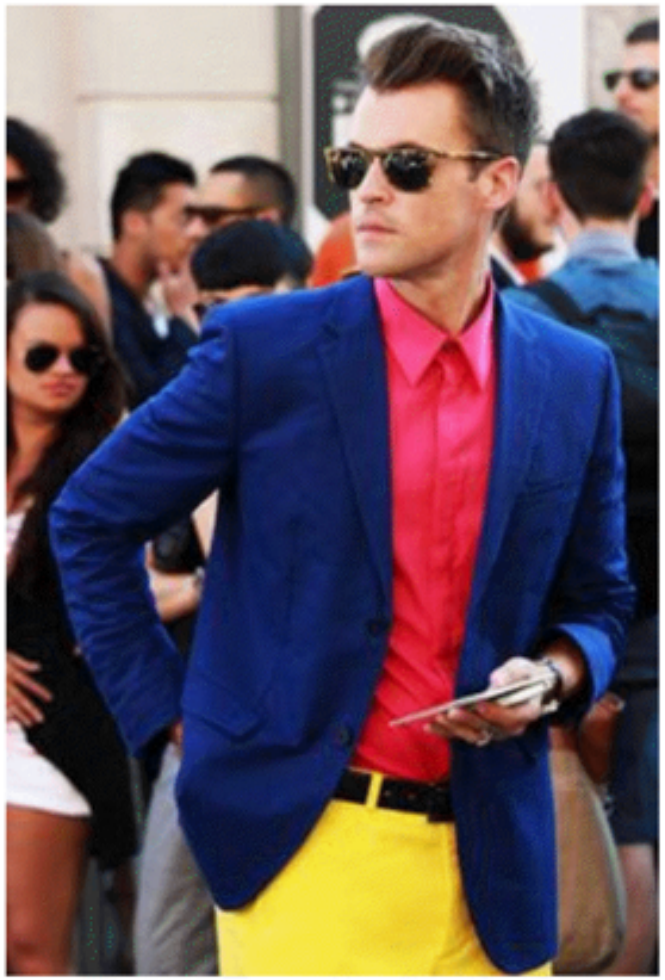


photo source: pinterest.com

Honestly, we find the best application for this quote in fashion: *"simplicity is the ultimate form of sophistication"*. Keeping it simple might be very difficult for several men, and, looking to the picture above, it really is.

Figure 63: Clothing mistake: using too many colors

Speaker notes

I want to encourage you to avoid mixing too many colors. Often the ideal number of colors is two. Here is a graphics image of what people who know fashion call a faux pas: the use of too many colors. I actually think that the colors look good here, but that’s because the model is so good looking. On me, these colors would be atrocious.

ADvertisement with a single red umbrella

Speaker notes

Graphic designers have known for quite a while that a restrained use of colors can be very effective. Here is an image from a YouTube video clip,

The Travelers - Look under the Umberella commercial (1986). Retrieved 2019-09-07 from https://www.youtube.com/watch?v=3zQX66jd_c0

The single red umbrella in a sea of black umbrellas stands out. Your eye can’t help but follow this umbrella as it travels across the screen from left to right. It’s a very powerful image.

A small dollop of color in your visualizations can be far more effective than using a whole bunch of different colors.

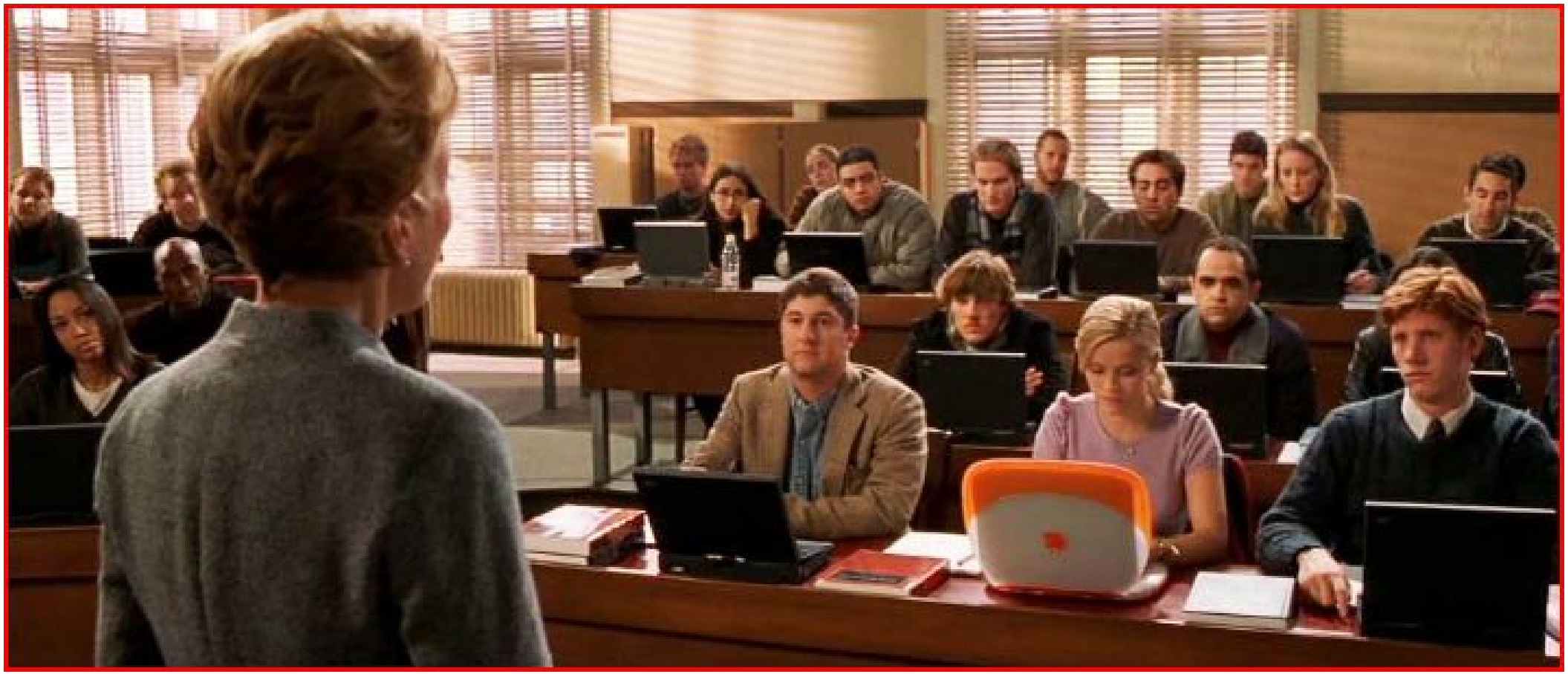


Figure 64: Use of color to highlight a single individual

Speaker notes

Here is a second example, from the movie, Legally Blonde. In this scene, the main character, Elle Woods, played by Reese Witherspoon, shows her individuality by opening up a bright orange and white Macintosh computer. All the other students are using generic black laptops.

This has practical implications for data visualization.

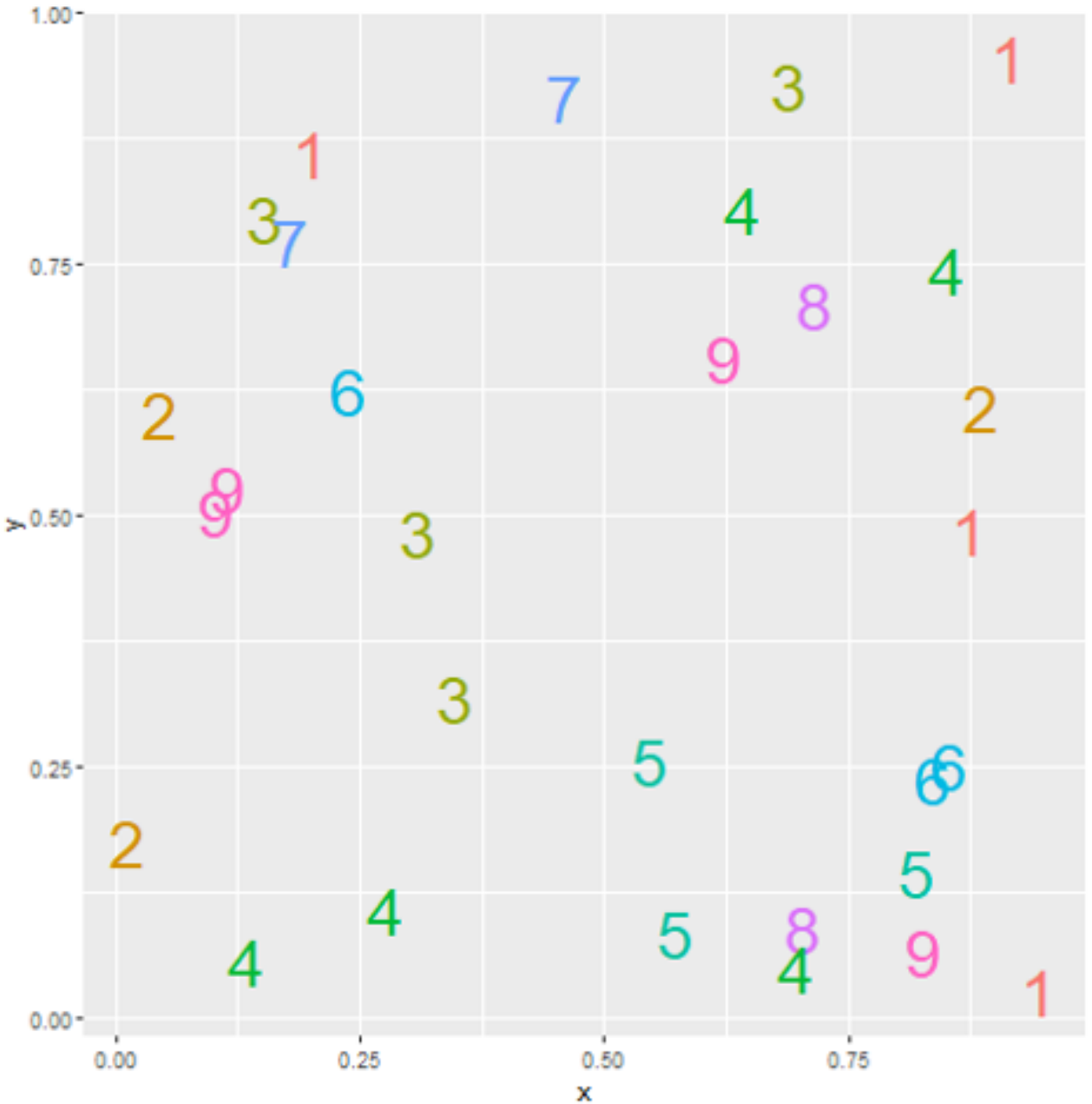


Figure 65: How many “5’s” are in this figure?

Speaker notes

Speaker notes

Here’s a simple exercise, count the number of “5’s” on this graph. Don’t include the “5” that appears in the caption.

When you have an answer, type it in the chat box.

PAUSE HERE

Now I did try to help by using a different color for each number.

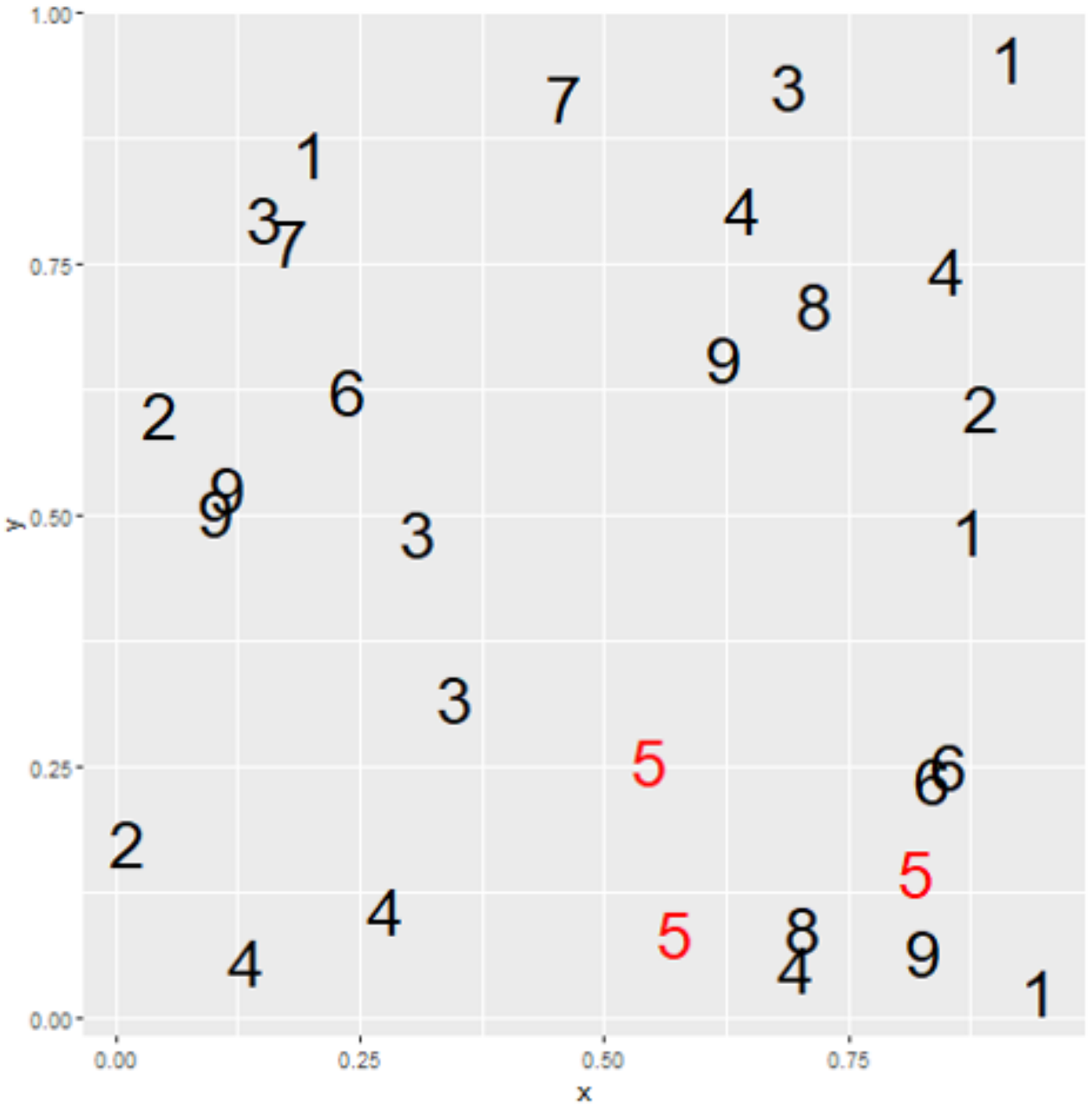


Figure 66: Repeat question. How many “5’s” are in this figure?

Speaker notes

Speaker notes

Okay, now repeat this exercise. How many “5’s” do you count? Notice how much faster it is when there is are two colors instead of nine.

Repeat quiz question 1

```
          No             Yes           Total
Female   154 (33.3%)    308 (66.7%)    462 (100%)
Male     709 (83.3%)    142 (16.7%)    851 (100%)
Total    863 (65.7%)    450 (34.3%)   1313 (100%)
```

This data table shows counts and …

1. cell percents
2. column percents
3. row percents
4. I do not know the answer

Speaker notes

Here is the question I asked earlier about percentages.

Repeat quiz question 2

The median might be preferred to the mean if

1. a single extreme value distorts the mean
2. the data follows a bell shaped curve
3. there is very little variation in the data
4. you have a biased sample
5. I do not know the answer

Speaker notes

Here is a repeat question about the median.

Repeat quiz question 3

The problem with error bars is that they

1. fail to show if the data is skewed
2. have several competing definitions
3. use only two numbers to characterize your data
4. all of the above are correct
5. none of the above are correct
6. I do not know the answer

Speaker notes

Speaker notes

Here is the repeat question about error bars