

Comments for MEDB 5501, Week 3

Count the 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?

A practical counting example

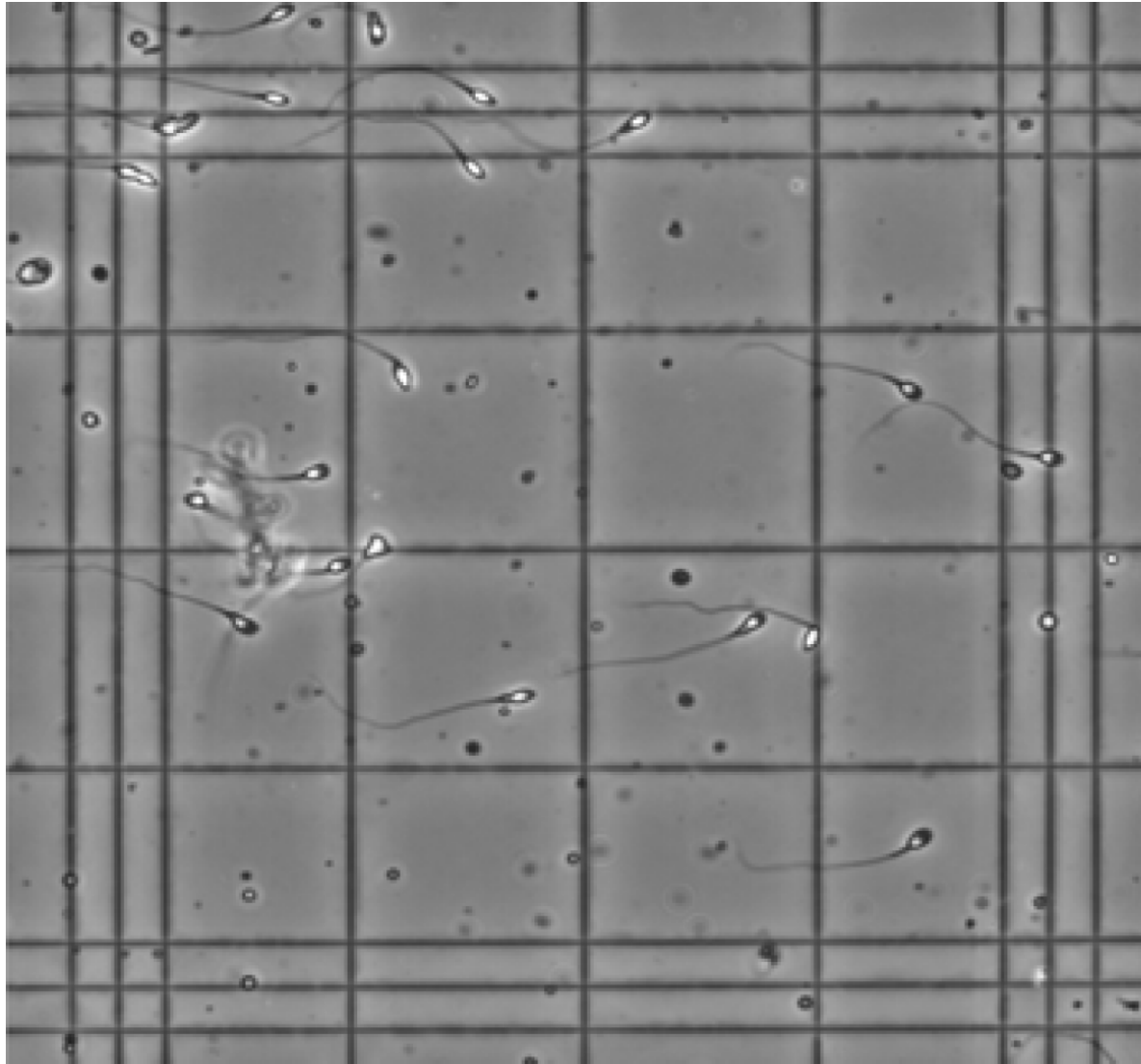


Image of a haemocytometer

Speaker notes

Speaker notes

This image is taken 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.

Measurement error

- Imprecision in a physical measurement
 - Example: GPS location
 - Can be off by up to 8 meters
 - Worse around large buildings
 - Other examples
 - Weight
 - Body temperature
 - Blood glucose

Speaker notes

Speaker notes

Statisticians are the only ones who openly admit the possibility of error. In fact, we obsess about error.

One important source of error is measurement error. This is typically defined with respect to a physical measurement.

I run outdoors for fun and to help me stay fit and lose weight. I've not been doing so hot on the weight loss side, but mostly because I indulge on the diet side of the equation. Anyway, one of the most fun parts of running is tracking the routes that you run and how fast you run those routes. I use two apps, Sportractive and Run Keeper. The Sportractive app shows me where I am at any point during the run using GPS satellites. It can be off by as much as 8 meters (about 26 feet), which causes some variation in how fast the app thinks I am running versus my actual speed. It doesn't make the app useless, but you do have to account for this.

In medicine, there are lots of physical measurements that have measurement error: weight, body temperature, blood glucose levels.

Reducing measurement error

- Calibration
- Consistent environment
- Good equipment
- Quality control
- Training

Speaker notes

Speaker notes

While you can't prevent measurement error, you can reduce it. Measurement that is done with medical equipment requires regular calibration. By the way, don't run all your control samples in the morning, re-calibrate during lunch, and then run all your treatment samples in the afternoon. This sounds obvious, but you'd be surprised how often researchers screw this up.

Consistency is also important. When I weigh myself, I try to do it in the morning before I've had anything to eat. It's usually when I weigh the less, but I'm not doing this to pretend that I am a few pounds lighter. I do it because I get more consistency.

You can get your blood glucose monitored, and it's always best if you can get it monitored after an overnight fast. Don't eat a Snickers bar on the way to your test!

You can measure your body temperature on your forehead, under your arm pit, under your tongue, or at least one other place that I won't mention. Some locations are more consistent (have less measurement error) than others.

Using good equipment can help. Your measurement on a balance beam scale is a bit more accurate than a digital scale. Try to use the exact same piece of equipment for measuring everyone in the study, or try to use the same model from the same manufacturer if you can.

A quality control program with regular assessments using known samples can also help. Monitor your lab daily or weekly with a control chart. If the control chart shows an out of control point, re-run all the samples from the time that the laboratory process was last shown to be in control.

Training of the operators of an medical equipment can also help reduce measurement error.

Errors of validity

- Mostly used for constructs
- Types of validity
 - Criterion
 - Concurrent
 - Predictive
 - Content/face
 - Many others
- Re-establishing validity

Speaker notes

Speaker notes

Statisticians also worry about errors in validity. These are errors that occur because you are measuring something different than what you think you are measuring.

Assessment of errors in validity is typically reserved for constructs. A construct is an assessment of something that has no direct physical manifestation. Your blood pressure is a physical measurement, but your stress level is not. Now maybe stress induces changes in blood pressure, hormone levels, etc. but stress itself is not a physical measurement like blood pressure is.

Typically, you measure a construct by asking a series of questions that all relate to that construct. There is a scale that measures how easily someone gets disgusted, and it asks questions about cockroaches, unwashed underwear, and ketchup on vanilla ice cream.

There are many types of validity. Criterion validity is comparison of your measurement to a well-accepted criterion. This is often called a gold standard. If you measure your construct and the criterion at the same time, this is called concurrent validity. Often this is comparison of a new construct to an existing and already validated construct. The Yale Single Question Depression Scale (Do you frequently feel sad or depressed?) was compared to the Beck Depression Inventory, a 21 item scale. It did not show a really strong correlation, but might be good enough to serve as an initial screen.

Concurrent validity is when the criterion or gold standard is measured at the same time as the construct. If the criterion is measured later, it is predictive validity. When the use of SAT scores as a measure of student success is validated by comparing it to college graduation rates, that is an example of predictive validity.

Content validity is an examination of individual elements of a construct by a panel of experts. It is a qualitative approach to validity. Closely related is face validity, the use of patients (read non-experts) to examine the elements of a construct. The line between content validity and face validity is very fuzzy.

There are many other types of validity. Don't get lost in all the terminology. Validity, at least the quantitative measures of validity, is almost always some type of correlation. When it is high, you have good validity.

If you are using a construct in a markedly different patient population, with different languages and different cultural norms, you need to re-establish validity, even for measures that have previously demonstrated good validity.

Errors of reliability

- Synonym: repeatability(?)
- Not reproducibility
- Both physical measurements and constructs
- Types of reliability
 - Test-retest
 - Inter-rater
 - Inter-method

Speaker notes

Speaker notes

You might see errors associated with the use of unreliable measurements. Often the term “repeatable” is used interchangeably. Some researchers make a distinction between these terms, but I don’t. I do, however, draw a distinction between reliability and reproducibility. Reproducibility is a demonstration that two different researchers agree when given access to the same dataset and the same software code.

You can assess reliability for both physical measurements and constructs. You demonstrate inter-rater reliability by showing that two evaluators working independently produce close to the same results. This does not work for self-reported outcomes like pain because only you can evaluate yourself.

A measurement taken twice allows you to assess test-retest reliability. The time spacing for the test and the retest is tricky. You want them far enough apart that the assessments are not done from memory, but not too far apart that temporal trends can appear. Some measures are stable over time. IQ, for example, is a measure that does not change overnight. It is presumed to be stable over many years or even many decades. At least until my age, when the deterioration of the brain starts to set in.

Errors due to sampling

- To be covered later
- Easiest to quantify
- Less important in era of big data

Speaker notes

Speaker notes

Although I plan to cover it later in more detail, I have to mention another source of errors. The process of collecting a random sample, even one that is done perfectly, involves error, because a sample is an imperfect representation of the population that the sample is being drawn from.

Sampling error goes down as the size of the sample increases. Unfortunately, other types of error (measurement error, errors in validity, errors in reliability) stay the same, or sometimes get worse as the sample size increases.

You live in a new era, the era of big data, and that lesson is especially critical now. When it is possible to get sample sizes in the millions or even billions, the concept of accounting for sampling error becomes silly. Things like confidence intervals and p-values become meaningless.

We'll still teach about sampling error, because a huge proportion of the data analyses done even today are on data that are not "big".

Short break

What have you learned? + Errors + Measurement + Validity + Reliability + Sampling What's next? + Descriptive statistics + Mean + Median

Speaker notes

Speaker notes

Let's stop and see if you have any questions.



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 throughout this talk.



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

- $\bar{X} = \frac{1}{n} \sum X_i$

- Median

- Sorted values $X_{[1]}, X_{[2]}, \dots, X_{[n]}$

- $X_{[(n+1)/2]}$ if n is odd,

- $(X_{[n/2]} + X_{[n/2+1]})/2$ if n is even

Speaker notes

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

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

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

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

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

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

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

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

Speaker notes

Just average the two middle values.

Criticisms of the mean and median

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

Speaker notes

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 mask 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.



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.



Excerpt from Gould 1985 publication

Speaker notes

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

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.



HHS Public Access

Author manuscript

Value Health. Author manuscript; available in PMC 2020 December 01.

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, MPH², Rena M. Conti, PhD³, Anupam B. Jena, MD, PhD^{4,5}, Dana P. Goldman, PhD^{1,2,5}

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

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

Speaker notes

Later on in the abstract, the authors point out that from the perspective 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.

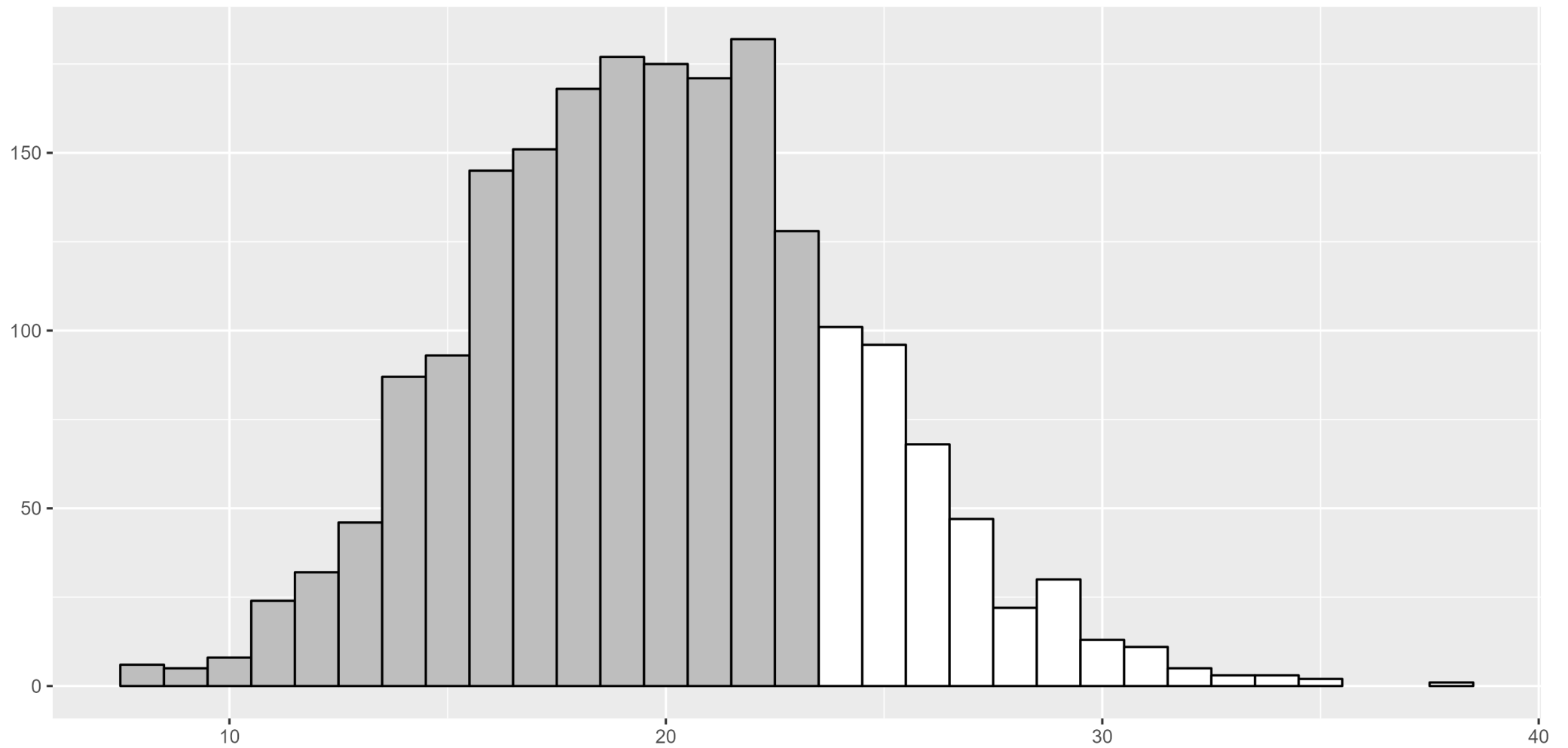
Break

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

Speaker notes

Speaker notes

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



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

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 p th 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, $p(n+1)=2 \rightarrow$ 2nd smallest value
 - For 4th percentile, $p(n+1)=1.6 \rightarrow$ halfway between two smallest values
 - For 2nd percentile, $p(n+1)=0.8 \rightarrow$ smallest value

Speaker notes

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

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

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

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

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

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

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 if 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

Speaker notes

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

Standard deviation

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

At least one alternative formulas.

Speaker notes

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

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

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 concomitant 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 further. 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.

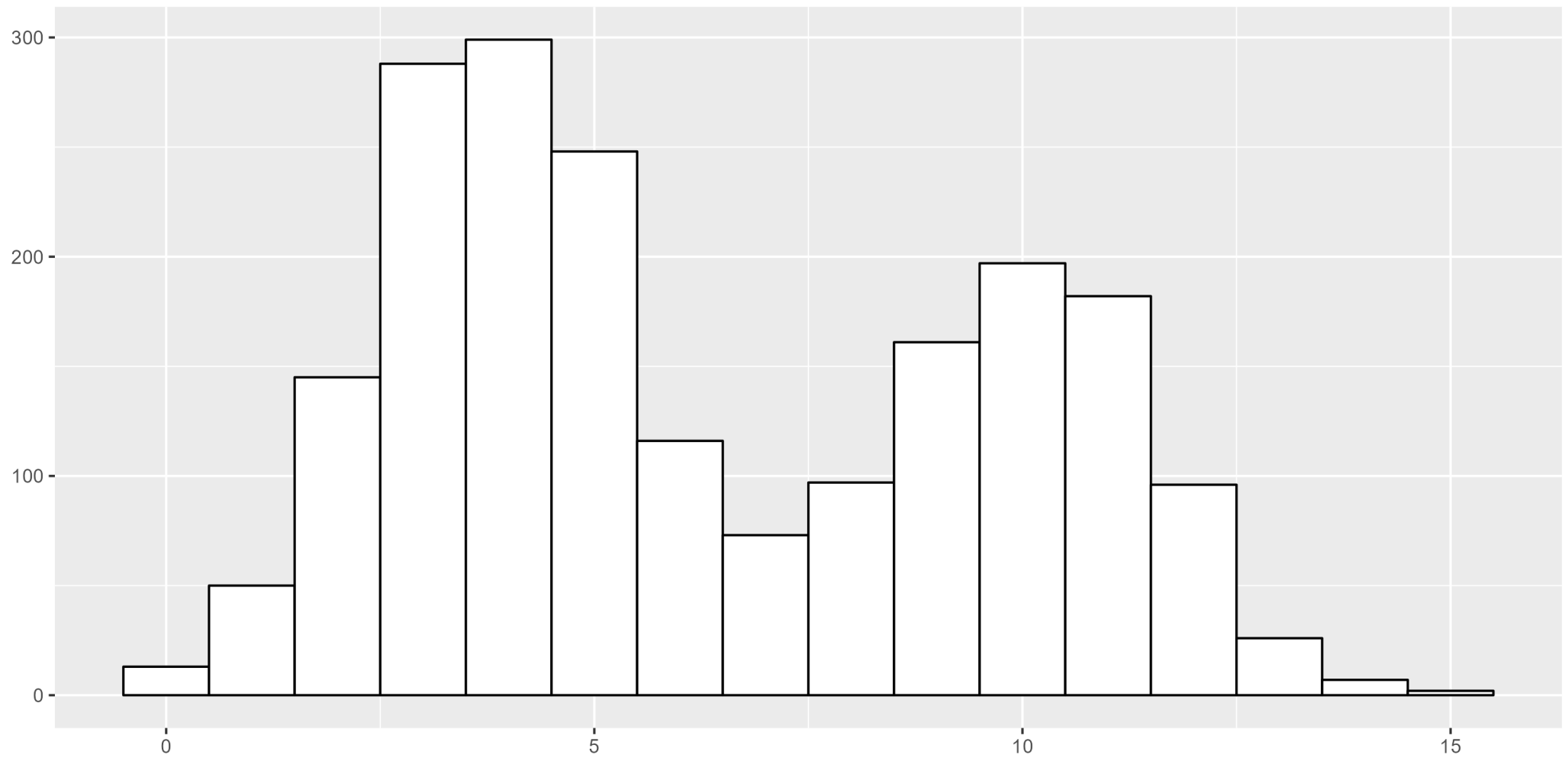
The bell shaped curve

- 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

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.

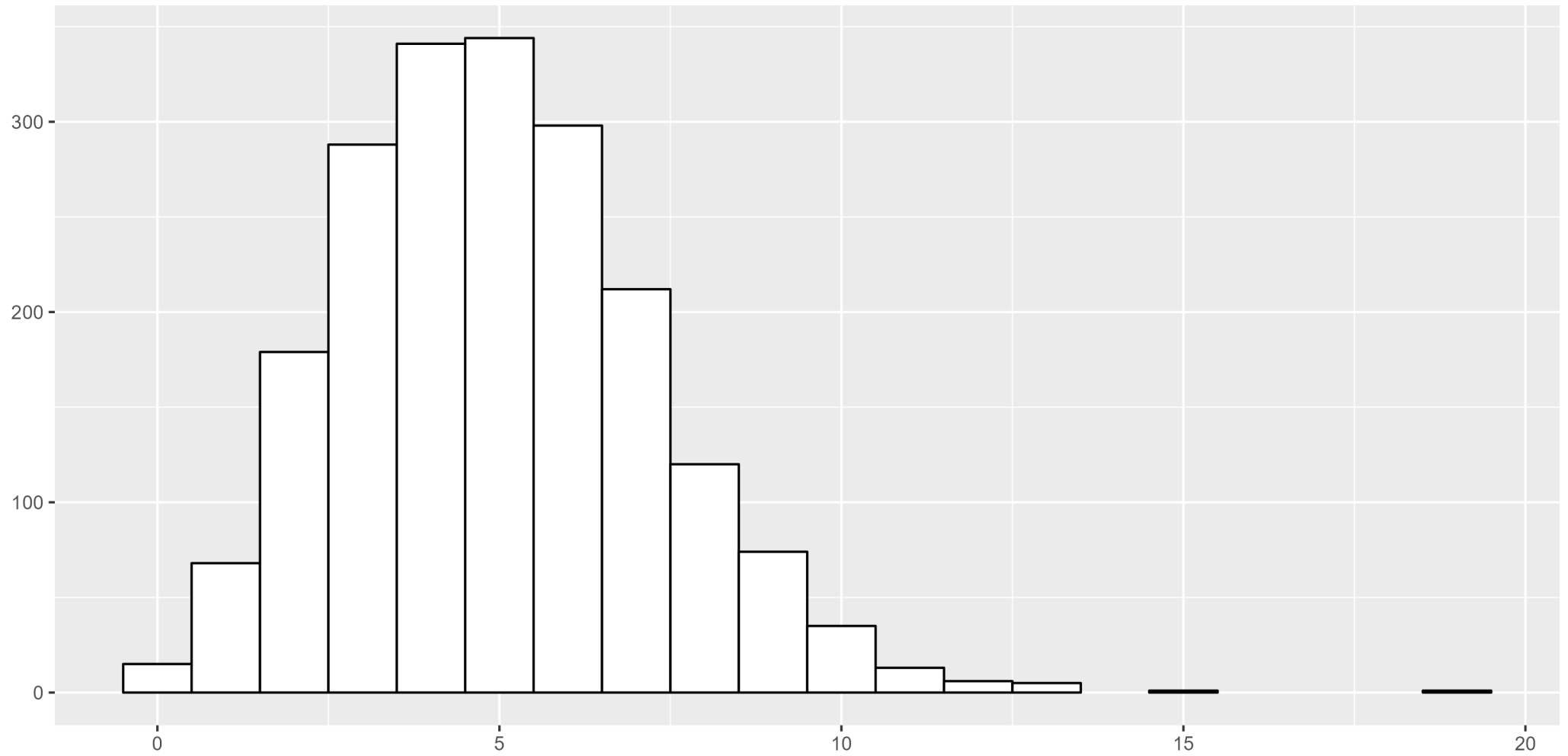


Bimodal histogram, not a bell shaped curve

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.

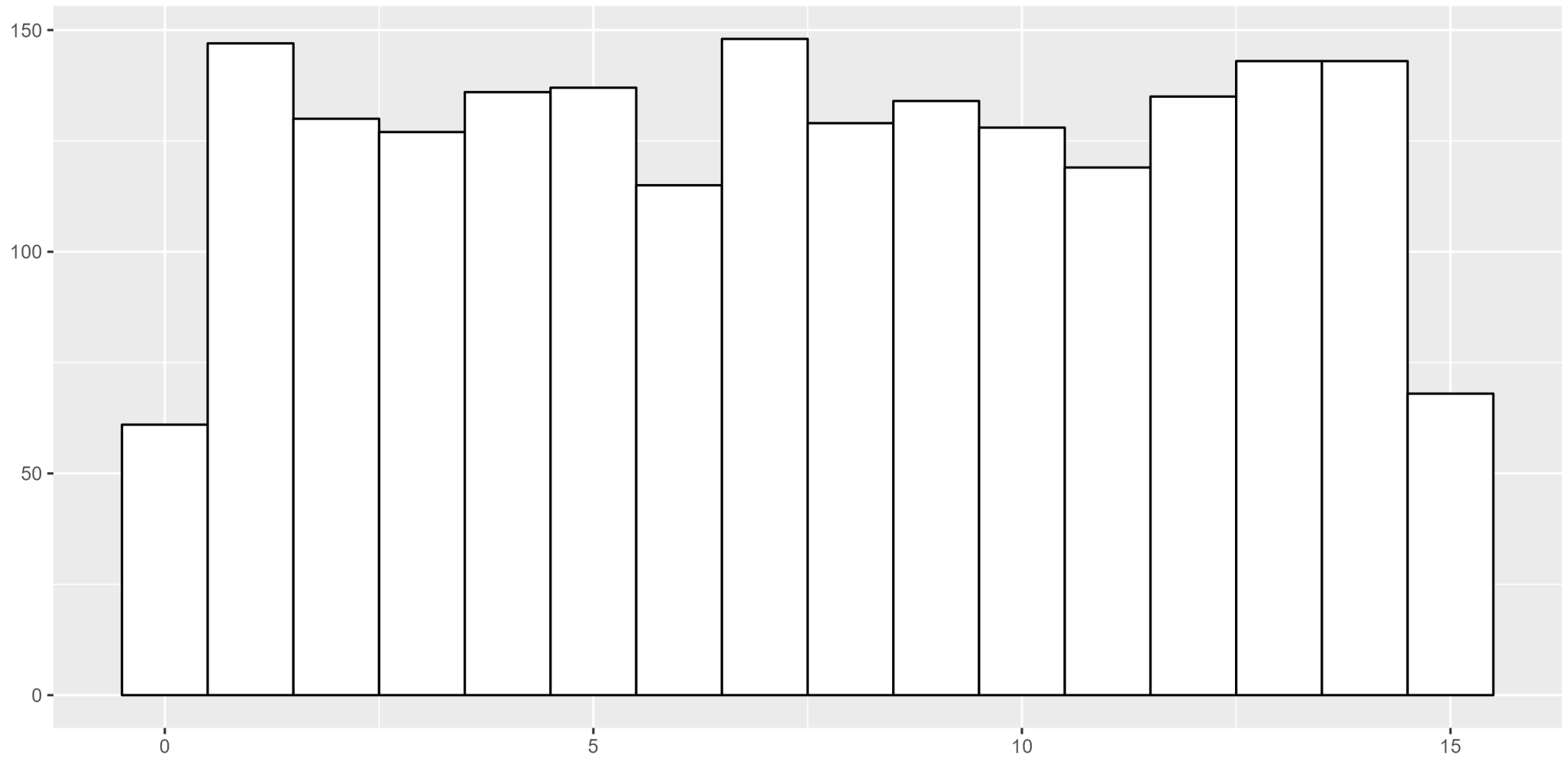


Skewed histogram, not a bell shaped curve

Speaker notes

Speaker notes

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

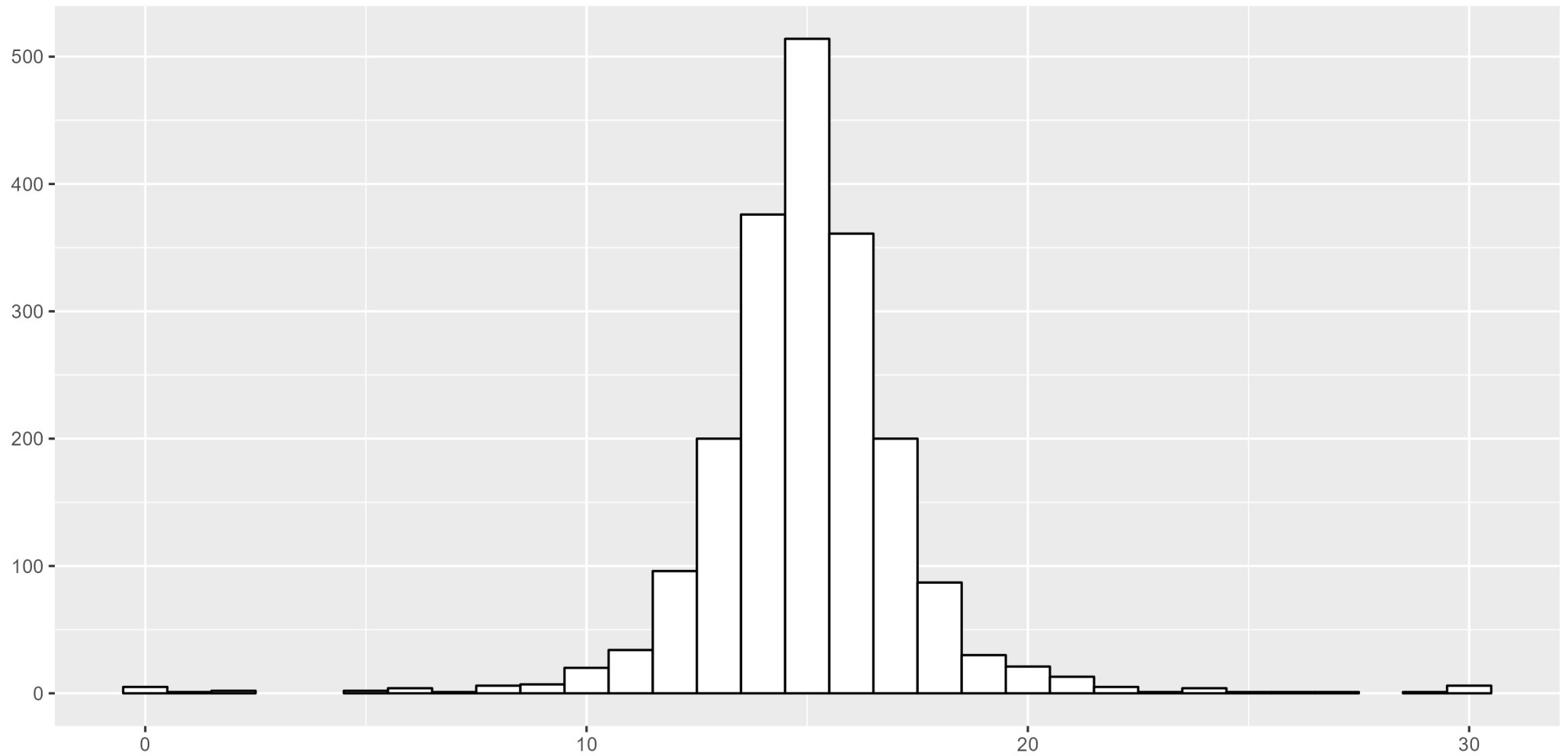


Uniform histogram, not a bell shaped curve

Speaker notes

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.

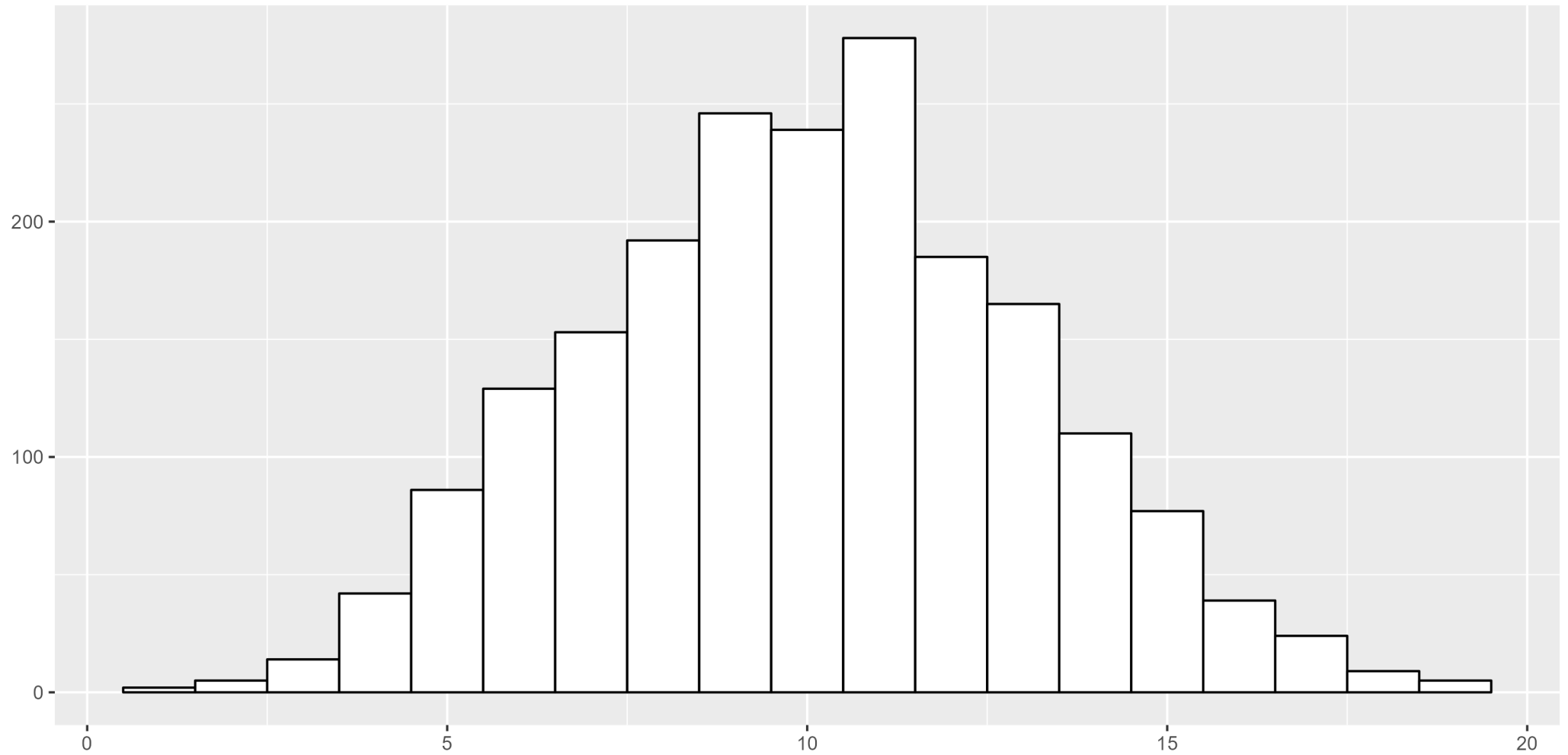


Heavy-tailed histogram, not a bell shaped curve

Speaker notes

Speaker notes

Here's a histogram that shows a symmetric distribution, 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.



Bell-shaped histogram, finally!

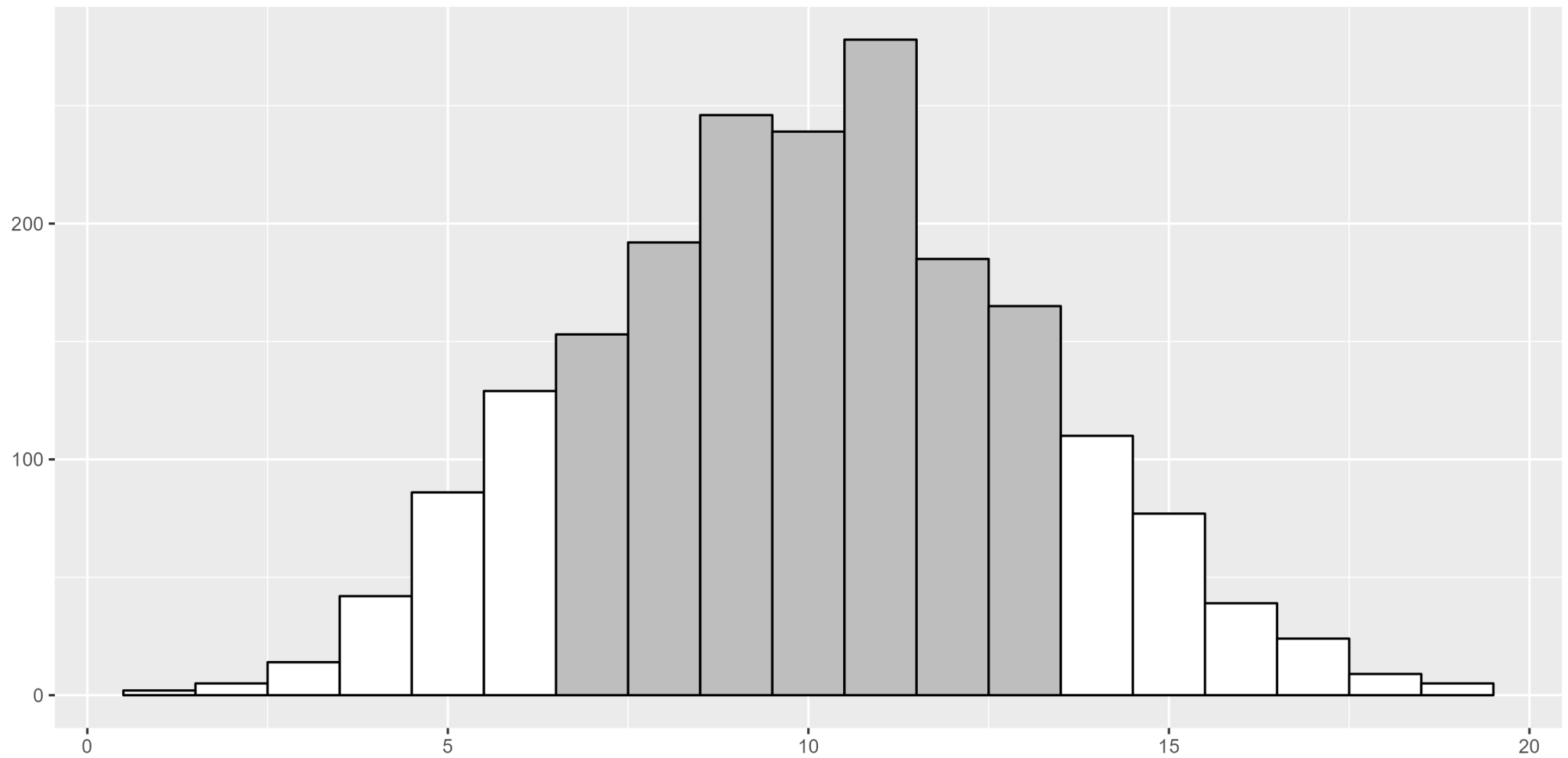
Speaker notes

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

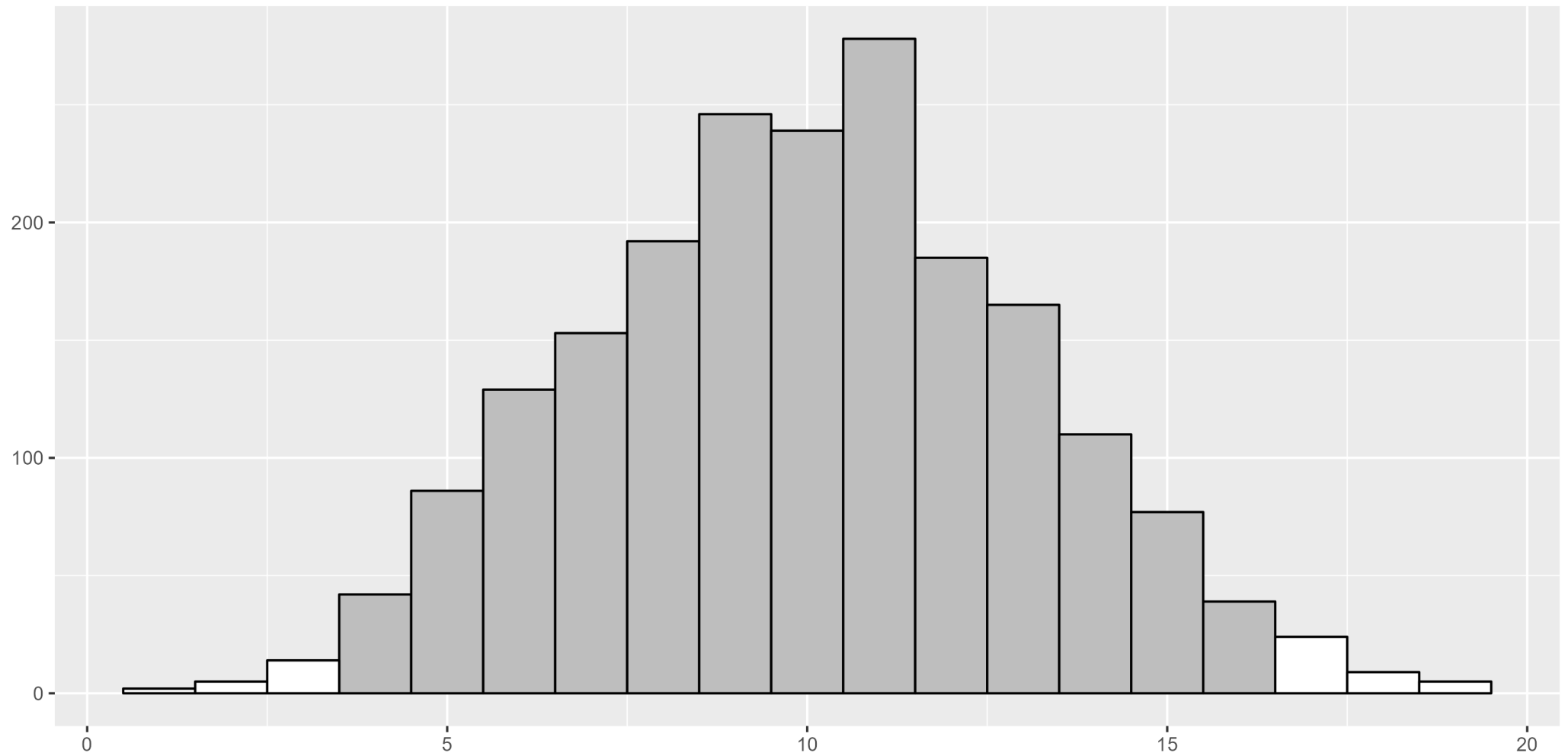


Percentage within one standard deviation

Speaker notes

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.

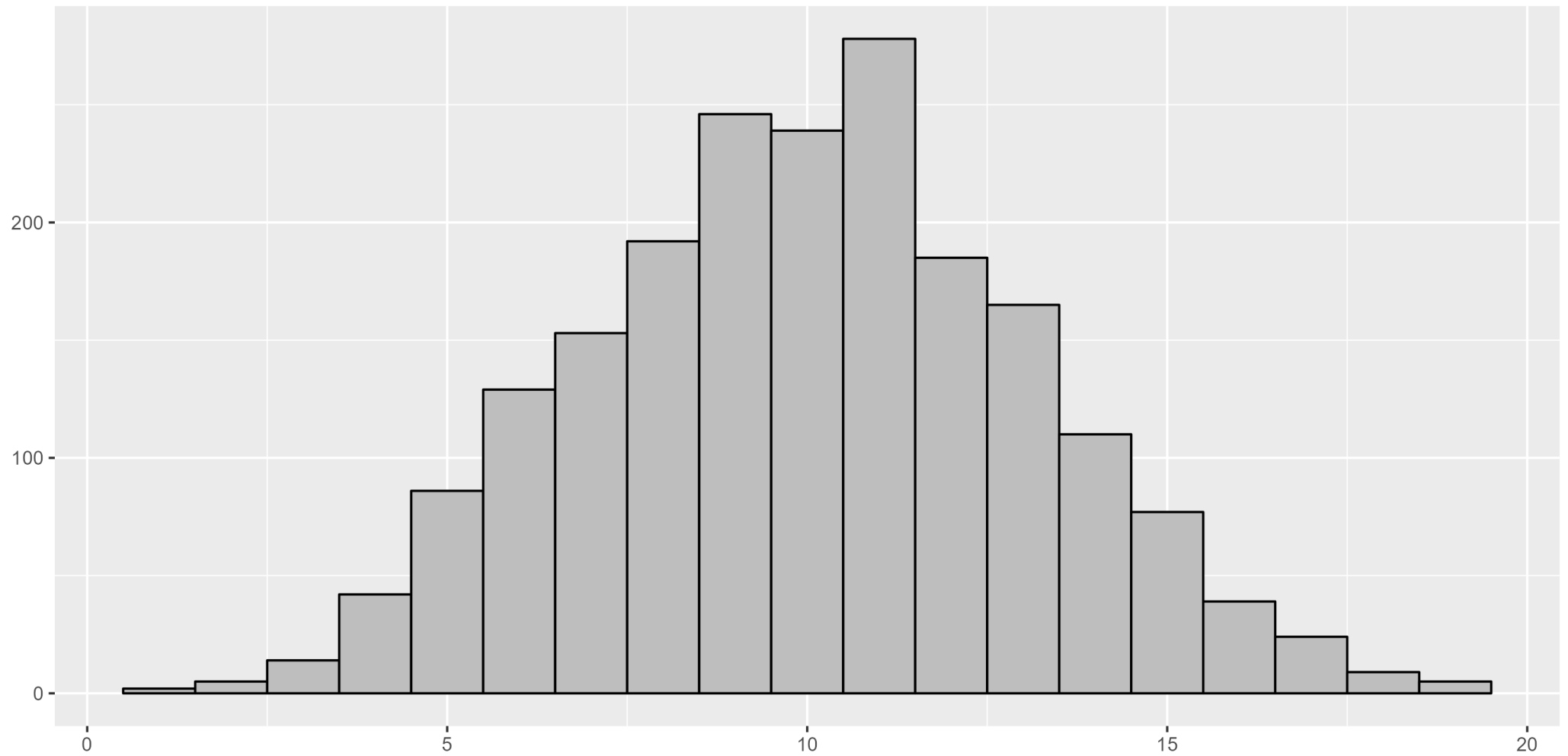


Percentage within two standard deviations

Speaker notes

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.



Percentage within three standard deviations

Speaker notes

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. Remember, 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

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

Lin et al 2022, PMID: 36126916

Speaker notes

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

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

Excerpt from Table 1 of Lin et al 2022: CCI

Speaker notes

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

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

Speaker notes


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 Tosato¹, Angelo Carfi¹, Ilaria Martis¹, Cristina Pais¹, Francesca Ciciarello¹,
Elisabetta Rota¹, Marcello Tritto¹, Andrea Salerno¹, Maria Beatrice Zazzara¹,
Anna Maria Martone¹, Annamaria Paglionico¹, Luca Petricca¹, Vincenzo Brandi¹,
Gennaro Capalbo¹, Anna Picca¹, Riccardo Calvani², Emanuele Marzetti³, Francesco Landi³;
Gemelli Against COVID-19 Post-Acute Care Team

Affiliations  expand

PMID: 34352201 PMCID: [PMC8286874](#) DOI: [10.1016/j.jamda.2021.07.003](#)

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

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 Andreucci³, Umberto Marcello Bracale⁴, Davide Costa^{2 5},
Egidio Bevacqua^{2 6}, Andrea Bitonti⁷, Sabrina Mellace⁸, Gianluca Buffone⁹, Stefano Candido¹⁰,
Michele Provenzano⁶, Raffaele Serra^{2 6}

Affiliations  expand

PMID: 34968328 PMCID: [PMC8608071](#) DOI: [10.3390/nursrep11030050](#)

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

Speaker notes

Here's an article about insomnia among nurses.

Ielapi 2021, PMID: 34968328 (continued)

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

Overall (n = 2'355)

Age, years 40.4 \pm 10.3

Months of work 168 [72-300]

Night shifts per month, number 6.3 \pm 1.4

Time to reach workplace, minutes 45 [45-65]

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 or minus two standard deviations for any observations that report the mean and standard deviation.

Ielapi 2021, PMID: 34968328 (continued)

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

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 or minus two standard deviations for any observations that report the mean and standard deviation.

What is a population?

- Population: a group that you wish to generalize your research results to. it is defined in terms of
 - Demography,
 - Geography,
 - Occupation,
 - Time,
 - Care requirements,
 - Diagnosis,
 - Or some combination of the above.

Speaker notes

Speaker notes

A population is a group that you have an interest in. You want to get a better understanding of this group, so you conduct a research study and wish to generalize the results of that study to the population.

In clinical research, a population is almost always a group of people. There are a few exceptions. Sometimes you want to characterize inanimate objects, such as a group of hospitals or a group of medical devices. But let's keep the focus on people for now.

A population of people is defined in terms of certain characteristics. Usually it is a combination of these characteristics.

Example of a population

All infants born in the state of Missouri during the 1995 calendar year who have one or more visits to the Emergency room during their first year of life.

Speaker notes

Speaker notes

Here is an example of a population. It has many of the characteristics described on the previous slide: demography (infants), geography (born in Missouri), time (born in calendar year 1995, during first year of life) and care requirements (one or more ER visits).

Most times the population is so large that it is difficult to get data on all the individuals of that population.

Here, we actually did have access to the data on all 29,637 infants, but most times you would not be so fortunate.

What is a sample?

- Sample: subset of a population.
- Random sample: every person has the same probability of being in the sample.
- Biased sample: Some people have a decreased probability of being in the sample.
 - Always ask “who was left out?”

Speaker notes

Speaker notes

A sample is a subset of a population. Because that population of infants was so large, you decided to collect data on a smaller group, a sample of 100 infants, say.

Statistics, according to one definition is the use of data from samples to make inferences about populations. That may be a bit too narrow a definition, but it does characterize quite a bit of what we statisticians do.

A random sample is a special type of sample. It is chosen in a way to insure that every person in the sample has the same probability of being in the sample.

In contrast a biased sample is one where some people in the population have a decreased chance of being in the sample. Often in a biased sample some people in the population are totally excluded.

An example of a biased sample

- A researcher wants to characterize **illicit drug use in teenagers**. She distributes a questionnaire to students attending a local public high school
- (in the U.S. high school is grades 9-12, which is mostly students from ages 14 to 18.)
- Explain how this sample is biased.
- Who has a decreased or even zero probability of being selected.

Type your ideas in the chat box.

Speaker notes

Speaker notes

Here is a scenario where a researcher selects a biased sample. I should note here that this is an example specific to the United States. In Italy, you might talk about a survey distributed to the scuola secondaria di secondo grado.

STOP AND GET STUDENT RESPONSES

There are a variety of responses here. The sample does not include home schooled students, students in private schools, students with chronic diseases that force frequent school absences, and students who have dropped out.

Fixing a biased sample

- Redfine your population
 - Not all teenagers,
 - but those attending public high schools.

Speaker notes

Speaker notes

Add note

What is a parameter?

- A parameter is a number computed from a sample.
 - Examples
 - Average health care cost associated with the 29,637 children
 - Proportion of these 29,637 children who died in their first year of life.
 - Correlation between gestational age and number of ER visits of these 29,637 children.
 - Designated by Greek letters (μ , π , ρ)

Speaker notes

Speaker notes

Add note

What is a statistic?

- A statistic is a number computed from a sample
 - Examples
 - Average health care cost associated with 100 children.
 - Proportion of these 100 children who died in their first year of life.
 - Correlation between gestational age and number of ER visits of these 100 children.
 - Designated by non-Greek letters (\bar{X} , \hat{p} , r).

Speaker notes

Speaker notes

Add note

What is Statistics?

- Statistics
 - The use of information from a sample (a statistic) to make inferences about a population (a parameter)
 - Often a comparison of two populations

Speaker notes

Speaker notes

Add note

The median is not the message

Q15, Q20

