MEDB 5501, Module02

2024-08-27

Topics to be covered

- What you will learn
  - Counts and percentages
  - Computing counts and percentages using R
  - Mean and median
  - Percentiles
  - Standard deviation
  - Computing means and standard deviations in R

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

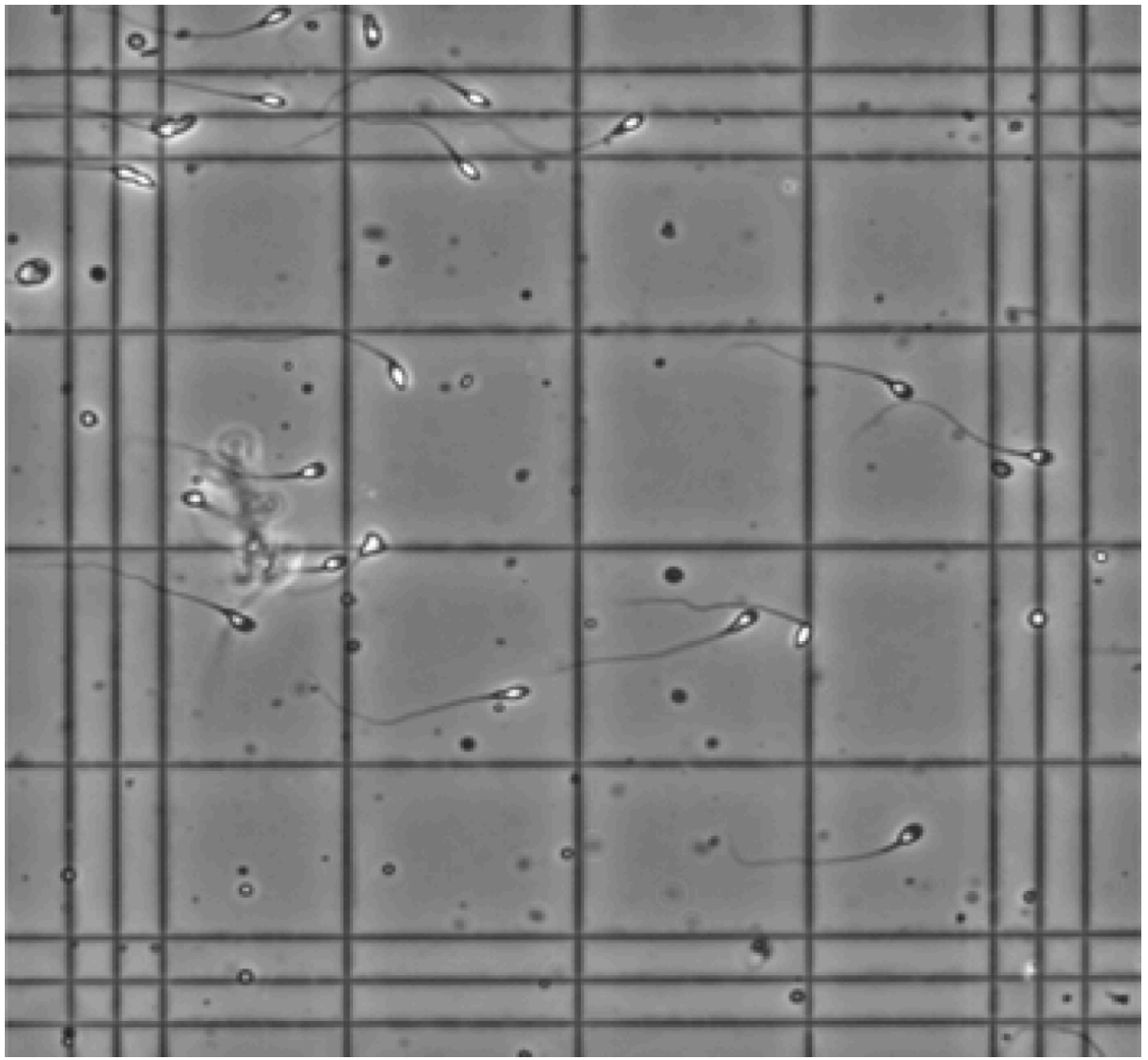


Image of a haemocytometer

Speaker notes

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.

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 the 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 that 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 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 interchangably. 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".

Break #1

- What you have learned
  - Counts and percentages
- What’s coming next
  - Computing counts and percentages using R

Data dictionary for sharing, 1

```
---
data_dictionary: sharing.xlsx
source: |
  Saginova, Olga (2020), “Dataset on the
  questionnaire-based survey of sharing
  services users’ motivation”, Mendeley Data,
  V1, doi: 10.17632/c5k8wjrhd9.1
```

Speaker notes

This data comes from a paper on sharing services: sharing cars, sharing bikes, even sharing housing. It has a large number of variables, but for this module, we are only interested in some basic demographics.

Data dictionary for sharing, 2

```
description: |
  From the original source: "The data set
  presents data collected by online survey
  with a questionnaire using Likert scale.
  The survey sample included 184 adults (18+),
  active and potential users of different
  sharing services platforms."
```

Speaker notes

The bulk of the data is a series of Likert scale items asking the degree of agreement or disagreement to various statements about sharing services.

Data dictionary for sharing, 3

```
copyright: |
  CC By 4.0. You can share, copy and modify
  this dataset so long as you give appropriate
  credit, provide a link to the CC BY license,
  and indicate if changes were made, but you
  may not do so in a way that suggests the
  rights holder has endorsed you or your use
  of the dataset. Note that further permission
  may be required for any content within the
  dataset that is identified as belonging to a
  third party.
```

Speaker notes

This is rare and greatly appreciated. You have information about the conditions under which you can use the data. This is a common open source license that allows you to use the dataset as long as you give appropriate credit.

Data dictionary for sharing, 4

```
format:
  proprietary (Excel)
varnames:
  first row of data

missing_value_code:
  not applicable

size:
  rows: 184
  columns: 31
```

Speaker notes

This is an Excel file with lots of columns. You will not be asked to analyze all 31 columns of data, just 3 related to basic demographics.

Data dictionary for sharing, 5

```
age:
  label: How old are you?
  values:
    - "18-25"
    - "26-35"
    - "36-45"
    - "46-60"
    - over 60
```

Speaker notes

This is the first demographic variable.

Data dictionary for sharing, 6

```
gender:
  values:
    - F
    - M
```

Speaker notes

This is the second demographic variable.

Data dictionary for sharing, 7

```
employment_status:
  label: Are you employed?
  values:
    - employed
    - entrepreneur
    - full-time student
    - self-employed
    - temporarily unemployed
    - unemployed
```

Speaker notes

This is the third demographic variable.

simon-5501-02-sharing.qmd, 1

```
---
title: "Counts and percentages"
format:
  html:
    slide-number: true
    embed-resources: true
editor: source
execute:
  echo: true
  message: false
  warning: false
---
```

Speaker notes

Here is a program template that illustrates how to calculate counts and percentages in R. You will make minor additions to this program in your programming assignment.

The first few lines are the documentation header.

simon-5501-01-template.qmd, 2

```
## Data source

This program uses data from a study of sharing services (like sharing an
automobile) and produces counts and percentages for a few demographic
variables. There is a [data dictionary][dd] that provides more details about
the data.

[dd]: https://github.com/pmean/datasets/blob/master/sharing.yaml
```

Speaker notes

Here is a bit of additional documentation.

simon-5501-01-template.qmd, 3

````
## Libraries

Here are the libraries you need for this program.

```{r setup}
library(readxl)
library(tidyverse)
```
````

Speaker notes

Always load the tidyverse library. You also need the readxl library because the data is stored as an Excel file.

simon-5501-01-template.qmd, 4

````
## Reading the data

Here is the code to read the data and show a glimpse. There are 31 columns
total, but I am showing just a few of the columns here.

```{r read}
fn <- "../data/sharing.xlsx"
sharing <- read_excel(fn)
glimpse(sharing[ , c(1, 5:7)])
```
````

Speaker notes

While you might want to see information about all 31 columns of data, I can’t display them all on a single slide. Column 1 is the id and columns 5, 6, and 7 are demographic variables.

simon-5501-01-template.qmd, 5

````
## Calculate counts and percentages for age group

```{r count-age-groups}
sharing |>
  group_by(age) |>
  summarize(n=n()) |>
  mutate(total=sum(n)) |>
  mutate(pct=100*n/total)
```

The survey respondents were younger than the general population. About half of the survey respondents were 18 to 25 years old. Only 3% were over 60. Six ages were missing.
````

Speaker notes

Here are counts and percentages for the first demographic variable. There are several alternative methods to get counts and percentages. The table function and the count function are both good alternatives.

Note the interpretation. You should always provide an interpretation.

Break #2

- What you have learned
  - Computing counts and percentages using R
- What’s coming next
  - Mean and 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
1	121	11.8	10.1
2	163	8.2	7.2
3	125	7.1	3.8
4	264	14.0	12.0
5	233	10.8	8.3
6	218	10.1	10.5
7	324	14.6	12.1
8	325	14.0	13.7

Speaker notes

Here is the dataset on bacteria counts. Measurements were done in 8 hotel rooms both before and after a rehaul of the air conditioning system.

Calculation of the before mean

$$\frac{1}{8}(11.8 + 8.2 + 7.1 + 14 + 10.8 + 10.1 + 14.6 + 14)$$
$$= \frac{1}{8}(90.6)$$
$$= 11.325$$

The average colony count per cubic foot before remediation, 11.3, is quite large.

Calculation of the after mean

$$\frac{1}{8}(10.1 + 7.2 + 3.8 + 12 + 8.3 + 10.5 + 12.1 + 13.7)$$
$$= \frac{1}{8}(77.7)$$
$$= 9.7125$$

The average colony count per cubic foot after remediation, 9.7, is smaller, but still quite large.

Calculation of the median

- Sort your data from low to high
- Select the middle observation(s)
  - If n is odd
    - Choose the (n+1)/2 observation
  - If n is even(n/2) and (n/2 + 1) if n is even
    - Go halfway between (n/2) and (n/2 + 1) observation

Calculate the before median, 1

Here is the sorted data.

```
  room before
1  125    7.1
2  163    8.2
3  218   10.1
4  233   10.8
5  121   11.8
6  264   14.0
7  325   14.0
8  324   14.6
```

Calculate the before median, 2

Here are the middle two observations

```
  room before middle
1  125    7.1
2  163    8.2
3  218   10.1
4  233   10.8   10.8
5  121   11.8   11.8
6  264   14.0
7  325   14.0
8  324   14.6
```

Calculate the before median, 3

Average the two middle observations

	room	before	middle	median
1	125	7.1		
2	163	8.2		
3	218	10.1		
4	233	10.8	10.8	11.3
5	121	11.8	11.8	
6	264	14.0		
7	325	14.0		
8	324	14.6		

Calculate the after median, 1

Here is the sorted data.

```
  room after
1  125   3.8
2  163   7.2
3  233   8.3
4  121  10.1
5  218  10.5
6  264  12.0
7  324  12.1
8  325  13.7
```

Calculate the after median, 2

Here are the middle two observations

```
  room after middle
1  125   3.8
2  163   7.2
3  233   8.3
4  121  10.1   10.1
5  218  10.5   10.5
6  264  12.0
7  324  12.1
8  325  13.7
```

Calculate the after median, 3

Average the two middle observations

```
  room after middle median
1  125   3.8
2  163   7.2
3  233   8.3
4  121  10.1   10.1   10.3
5  218  10.5   10.5
6  264  12.0
7  324  12.1
8  325  13.7
```

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.

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

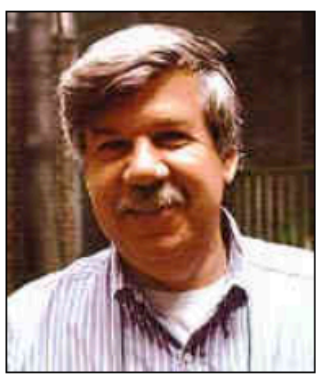
main menu - articles - prostate - stories - sexuality - resources - glossary - search

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.

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]

Chen et al 2019

Speaker notes

Here is an article I found on PubMed 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)

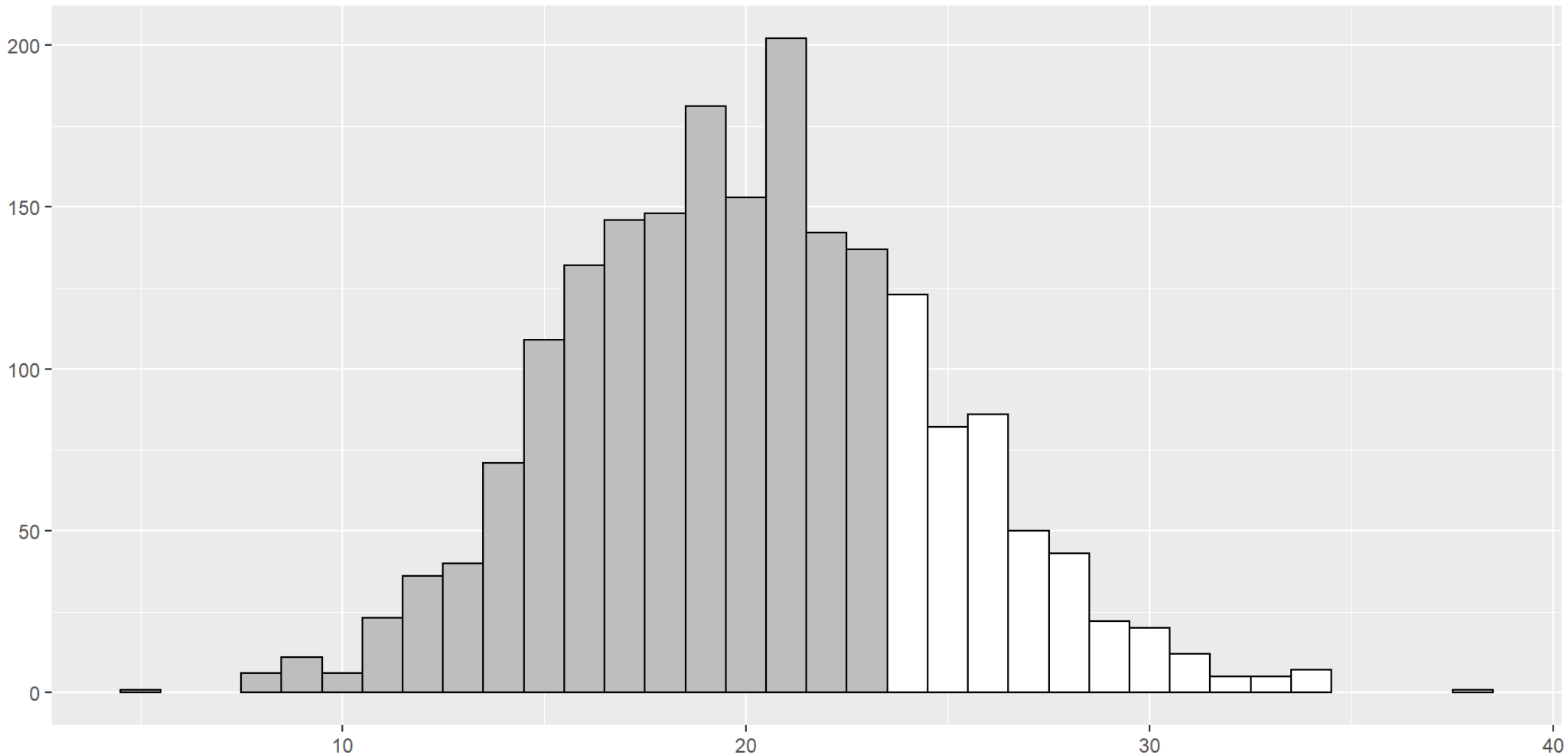
Results: We found that between 1995 and 2012, price increases outstripped median survival gains, a finding consistent with previous literature. **Nevertheless, price per mean life-year gained increased at a considerably slower rate, suggesting that new drugs have been more effective in achieving longer-term 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 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 #3

- What you have learned
  - Mean and median
- What’s coming next
  - Percentiles


200
150
100
50
0
10
20
30
40

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”.

Calculate before remediation upper quartile, 1

Here is the sorted data.

```
  room before
1  125    7.1
2  163    8.2
3  218   10.1
4  233   10.8
5  121   11.8
6  264   14.0
7  325   14.0
8  324   14.6
```

Calculate before remediation upper quartile, 2

Calculate 0.75*(8+1) = 6.75. Select the 6th and 7th observations

	room	before	pick
1	125	7.1	
2	163	8.2	
3	218	10.1	
4	233	10.8	
5	121	11.8	
6	264	14.0	14
7	325	14.0	14
8	324	14.6	

Calculate before remediation upper quartile, 3

Average the two observations

```
  room before pick   q3
1  125    7.1
2  163    8.2
3  218   10.1
4  233   10.8
5  121   11.8
6  264   14.0   14   14
7  325   14.0   14
8  324   14.6
```

Calculate after remediation upper quartile, 1

Here is the sorted data.

```
  room after
1  125   3.8
2  163   7.2
3  233   8.3
4  121  10.1
5  218  10.5
6  264  12.0
7  324  12.1
8  325  13.7
```

Calculate after remediation upper quartile, 2

Calculate 0.75*(8+1) = 6.75. Select the 6th and 7th observations

	room	after	pick
1	125	3.8	
2	163	7.2	
3	233	8.3	
4	121	10.1	
5	218	10.5	
6	264	12.0	12
7	324	12.1	12.1
8	325	13.7	

Calculate after remediation upper quartile, 3

Average the two observations

```
  room after pick    q3
1  125   3.8
2  163   7.2
3  233   8.3
4  121  10.1
5  218  10.5
6  264  12.0   12 12.05
7  324  12.1 12.1
8  325  13.7
```

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 #4

- What you have learned
  - Percentiles
- What’s coming next
  - Standard deviation

Standard deviation

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

At least one alternative formula.

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.

The two standard deviation rule of thumb

- Approximately 95% of the data lies within two standard deviations of the mean.
  - Except for highly skewed datasets

Speaker notes

There is an empirical observation that for many datasets, most of the data (about 95%), lies within two standard deviations of the mean.

Here is some artificial data to illustrate this rule

```
194 209 211 215 219 221 223 224 227 229
232 233 235 238 239 242 243 250 253 254
254 255 256 257 258 259 261 262 263 268
268 268 269 270 271 272 273 273 274 277
278 281 282 283 287 289 292 293 294 295
296 298 298 300 301 302 302 302 303 304
305 309 310 311 311 313 318 321 322 323
325 326 327 328 332 333 336 342 349 349
351 357 358 358 359 363 368 375 379 379
380 381 388 391 394 397 400 403 418 450
```

The mean is 299.2 and the standard deviation is 54.4. Plus or minus two standard deviations is 190.4 to 408.

Tosato et al 2021, PMID: 34352201, part 1

Here is a practical example of the plus or minus two standard deviation rule.

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 Carfi [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

Speaker notes

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

Tosato 2021, PMID: 34352201, part 2

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

Here is the first half of the abstract.

Tosato 2021, PMID: 34352201, part 3

- The mean age was 73.1 ± 6.2 years
  - 73 - 2 * 6 = 61
  - 73 + 2 * 6 = 85
- The average time elapsed from hospital discharge was 76.8 ± 20.3 days
  - 77 - 2 * 20 = 37
  - 77 + 2 * 20 = 117

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

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 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.

Break #5

- What you have learned
  - Standard deviation
- What’s coming next
  - Computing means and standard deviations in R

Data dictionary for legionnaires, 1

```
---
data_dictionary: "legionnaire's disease"
format:
  txt: tab-delimited
varnames:
  first row of data
missing_value_code:
  not needed
```

Speaker notes

Here is a dataset you will need for your programming assignment. it is a tab delimited file with variable names in the first row of data.

Data dictionary for legionnaires, 2

```
description: >
  Fictional data on bacteria counts before
  and after air conditioning maintenance.
additional_description:
  https://dasl.datadescription.com/datafile/legionnaires-disease
download_url:
  https://dasl.datadescription.com/download/data/3310
notes: >
  The use of a space in the first variable name might
  cause some minor difficulties during import.
```

Speaker notes

This is fictional data, loosely based on a report about an outbreak of Legionnaire’s disease at a hotel. There’s a warning about the variable names. One of them has a blank in the middle, which can cause some minor difficulties.

Data dictionary for legionnaires, 3

```
source: >
  DASL (Data and Story Library), a repository for various
  data sets useful for teaching.
copyright: >
  Unknown. You should be able to use this data for
  individual educational purposes under the Fair Use
  guidelines of U.S. copyright law.
size:
  rows: 8
  columns: 2
```

Speaker notes

This is a small dataset with eight rows and three columns.

Data dictionary for legionnaires, 4

```
vars:
  Room number:
    label: Hotel room number
  Before:
    label: Bacterial count before maintenance
    unit: colonies per cubic foot


  After:
    label: Bacterial count before maintenance
    unit: colonies per cubic foot
---
```

Speaker notes

The variables are measurements before and after a major overhaul of the air conditioning system. The units are colonies per cubic foot of air. A pump pushes a certain volume of air through a filter and then bacterial colonies are allowed to grow on that filter.

simon-5501-02-legionnaires.qmd, 1

```
---
title: "Univariate statistics for Legionnaires disease"
format:
  html:
    slide-number: true
    embed-resources: true
editor: source
execute:
  echo: true
  message: false
  warning: false
---
```

Speaker notes

The first few lines are the documentation header

simon-5501-02-legionnaires.qmd, 2

```
## Data source

This program uses data from a fictional study of Legionnaires disease and
produces some simple univariate statistics: means, standard deviations, and
percentiles. There is a [data dictionary][dd] that provides more details about
the data.

[dd]: https://github.com/pmean/data/blob/main/files/legionnaires-disease.yaml
```

Speaker notes

Here is some additional documentation.

simon-5501-02-legionnaires.qmd, 3

````
## Libraries

Here are the libraries you need for this program.

```{r setup}
library(tidyverse)
```
````

Speaker notes

Loads the tidyverese library. No other libraries are needed.

simon-5501-02-legionnaires.qmd, 4

````
## Reading the data

Here is the code to read the data and show a glimpse. There are 31 columns
total, but I am showing just a few of the columns here.

```{r read}
fn <- "../data/legionnaires-disease.txt"
ld_raw_data <- read_tsv(fn, col_types="cnn")
glimpse(ld_raw_data)
```
````

Speaker notes

Use the read_tsv function when your data uses tab delimiters.

simon-5501-02-legionnaires.qmd, 5

````
## Rename, 1

Notice how R encloses the first variable name (Room Number) in back-quotes.
This is needed when a variable includes an embedded blank. You should rename
this variable at your first opportunity.

```{r rename-1}
names(ld_raw_data)[1] <- "Room_Number"
glimpse(ld_raw_data)
```
````

Speaker notes

Try to avoid spaces within a variable name. This code changes the space to an underscore.

simon-5501-02-legionnaires.qmd, 6

```
## Rename, 2

I find that many of the mistakes that I make are due to inconsistencies in how
I name variables. Capitalization is one of the biggest problems. So I have
gotten into the habit of converting variable names to all lower case. That way
I don't have to worry about whether it is "Before" or "before". Here is the
code to convert every capital letter to a lowercase letter.

```{r rename-2}
names(ld_raw_data) <- tolower(names(ld_raw_data))
glimpse(ld_raw_data)
```
```

Speaker notes

The tolower fuction replaces every uppercase letter with its lowercase equivalent.

simon-5501-02-legionnaires.qmd, 7

````
## Calculate means and standard deviations before remediation

```{r before-means}
ld_raw_data |>
  summarize(
    before_mn=mean(before),
    before_sd=sd(before))
```

The average colony count per cubic foot before remediation, 11.3, is quite large. The standard deviation, 2.8, represents a moderate amount of variation in this variable.
````

Speaker notes

This code produces a mean and standard deviation for the colony counts before remediation.

simon-5501-02-legionnaires.qmd, 8

````
## Calculate means and standard deviations after remediation

```{r after-means}
ld_raw_data |>
  summarize(
    after_mn=mean(after),
    after_sd=sd(after))
```

The average colony count per cubic foot after remediation, 9.7, is still quite large. The standard deviation, 3.2, represents a moderate amount of variation in this variable and is roughly comparable to the variation before remediation.
````

Speaker notes

This code produces a mean and standard deviation for the colony counts after remediation.

simon-5501-02-legionnaires.qmd, 9

````
## Calculate median and range before intervention

You could also use "median(before)" and  "min(before)" and "max(before)" in
the code below.

```{r before-quantiles}
ld_raw_data |>
  summarize(
    before_median=quantile(before, probs=0.5),
    before_min=quantile(before, probs=0),
    before_max=quantile(before, probs=1))
```

The median colony count before remediation, 11.3, is roughly the same as the
mean. The data ranges from 7.1 to 14.6 colonies per cubic centimeter, a fairly
wide range.
````

Speaker notes

The quantile function calculates the median and other percentiles. Setting probs equal to 0 and 1 produces the minimum and maximum values.

simon-5501-02-legionnaires.qmd, 10

````
## Calculate median and range after intervention

```{r after-quantiles}
ld_raw_data |>
  summarize(
    after_q50=quantile(after, probs=0.5),
    after_min=quantile(after, probs=0),
    after_max=quantile(after, probs=1))
```

The median colony count, 10.3, is slightly lower after remediation. The data range from 3.8 to 13.7 colonies per cubic centimeter and is about as wide as the range before remediation.
````

Speaker notes

Use similar code to get the median, minimum, and maximum for the after remediation measurements.

simon-5501-02-legionnaires.qmd, 11

```
## Additional comments

The names that you choose for the left hand side of the equal sign are
arbitrary. You should choose a descriptive name, but you have lots of options.
A median of the before and after values could be called

-   Before_median, After_median
-   Median0, Median1
-   Second_quartile_A, Second_quartile_B
-   or many other reasonable choices.
```

Speaker notes

It is worth mentioning here that you have lots of options for the choice of names for the various statistics.

simon-5501-02-legionnaires.qmd, 12

````
## Calculate a change score

For data like this with two measurements before and after an intervention, you should compute a change score. The way the computations are done below, a positive value means a reduction in colony counts. Note that any time you make a major change in a dataset, you should save it with a different name. That makes it easier for you to back up if you end up going down a blind alley.

```{r}
ld_raw_data |>
  mutate(change=before-after) -> ld_change_scores
glimpse(ld_change_scores)
```
````

Speaker notes

This is the code for calculating a change score. You will use the change score in your programming assignment for this module.

Summary

- What you have learned
  - Counts and percentages
  - Computing counts and percentages using R
  - Mean and median
  - Percentiles
  - Standard deviation
  - Computing means and standard deviations in R