Clinical statistics for non-statisticians: Day one

Steve Simon

## One warning

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

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

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

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

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

## Start with a bad joke

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

The other statistic responds …

*Put your reaction (“Ha ha”, “Groan”, etc.) in the chat box.*

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

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

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

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

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

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

Okay, I’m waiting for reactions.

## Introduction

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

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

Here are three numbers about me.

## Your turn

I want each of you to share an interesting number about yourself.

## A bit more about myself

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

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

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

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

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

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

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

## Outline of the three day course

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

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

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

## Day one topics

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

Today, you will learn about numerical summaries.

## Day one topics (continued)

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

You will also learn a little bit about data visualization.

## Quiz question 1

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

This data table shows counts and …

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

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

## Quiz question 2

The median might be preferred to the mean if

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

Here is a question about the median.

## Quiz question 3

The problem with error bars is that they

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

Here is a question about error bars

## Counting and proportions

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

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

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

## Student exercise

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

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

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

PAUSE HERE.

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

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

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

|  |
| --- |
| Figure 1: Image of a haemocytometer |

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.

|  |
| --- |
| Figure 2: Titanic data: counts of survival by gender |

Here is some count data from an interesting data set. It shows who survived and who did not on the passenger ship, Titanic.

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

They kept records on everyone on the ship: sex, age, and passenger class. There were 462 women on the ship. 308 of them survived, including Kate Winslet. The men did not fare as well. This was in a time when they really believed in the saying “Women and children first”. If this happened today, I’d push past all the ladies and the little kids and jump in that life boat first.

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

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

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

|  |
| --- |
| Figure 3: Titanic data with column percentages |

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

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

|  |
| --- |
| Figure 4: Titanic data with row percentages |

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

!7%! This shows how poorly the men fared on the Titanic. If you were female, you might have died, but more likely than not you did survive. For the men, not such good news. Most of them died. Only a small fraction survived.

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

## Percentages divided by grand total

|  |
| --- |
| Figure 5: Titanic data with cell percentages |

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

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

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

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

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

## My recommendations

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

I have some general guidelines that I use. They don’t always work, but they work most of the time.

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

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

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

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

## Some rationale for these choices

My way

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

Not my way

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

Now, I believe it is important to think carefully about which is your rows and which is your columns. Here’s the layout that I recommend on the left and the layout that I don’t recommend on the right. The key comparison is among survival rates, 67% for females and only 17% for males. When you orient my way with the treatment/exposure (Sex) as rows and the outcome (Survived) as the columns, the numbers 67% and 17% are very close to one another. In the alternate layout the numbers you are most interested in comparing are not as close together.

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

## Break

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

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

## On your own

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

|  |
| --- |
| Figure 6: Titanic passenger class counts |

|  |
| --- |
| Figure 7: Titanic child counts |

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

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

Be sure to interpret these numbers. Come back together again in about 10 minutes.

|  |
| --- |
| Figure 8: Cartoon image of Professor Mean |

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

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

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

|  |
| --- |
| Figure 9: Road with a median strip |

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

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
* Median
  + Sorted values
    - if n is odd,
    - if n is even

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

suppressMessages(  
 suppressWarnings(  
 library(glue)))  
suppressMessages(  
 suppressWarnings(  
 library(magrittr)))  
suppressMessages(  
 suppressWarnings(  
 library(tidyverse)))

room <- c(  
 121,  
 125,  
 163,  
 218,  
 233,  
 264,  
 324,  
 325)

bacteria\_0 <- c(  
 11.8,  
 7.1,  
 8.2,  
 10.1,  
 10.8,  
 14,  
 14.6,  
 14)

bacteria\_1 <- c(  
 10.1,  
 3.8,  
 7.2,  
 10.5,  
 8.3,  
 12,  
 12.1,  
 13.7)

Room Before After Change  
 121 11.8 10.1 -1.7  
 125 7.1 3.8 -3.3  
 163 8.2 7.2 -1.0  
 218 10.1 10.5 0.4  
 233 10.8 8.3 -2.5  
 264 14 12 -2.0  
 324 14.6 12.1 -2.5  
 325 14 13.7 -0.3

## Before remediation mean

adds\_0 <- paste0(bacteria\_0, collapse=" + ")   
sum\_0 <- sum(bacteria\_0)  
mean\_0 <- mean(bacteria\_0)

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

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

adds\_1 <- paste0(bacteria\_1, collapse=" + ")   
sum\_1 <- sum(bacteria\_1)  
mean\_1 <- mean(bacteria\_1)

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

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

## Before remediation median (1/4)

space <- function(x) {  
 x1 <- rep("", 15)  
 x1[2\*(1:8)-1] <- x  
 return(x1)  
}  
  
sq <- function(x) {  
 paste0(x, collapse="\n")  
}  
  
step1 <- function(x) {  
 y <- sprintf("%4.1f", x)  
 room %>%  
 paste(y, sep=" ") %>%  
 space  
}

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

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

## Before remediation median (2/4)

step2 <- function(x) {  
 o <- order(x)  
 y <- sprintf("%4.1f", x[o])  
 room[o] %>%  
 paste(y, sep=" ") %>%  
 space  
}

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

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

The data was arranged by toom number, but now it is arranged by bacterial count.

## Before remediation median (3/4)

step3 <- function(x, div=4:5) {  
 o <- order(x)  
 y <- rep("", 8)  
 y[div] <- x[o][div]  
 step2(x) %>%  
 paste(space(y), sep=" ")  
}

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

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)

step4 <- function(x, div=4:5) {  
 o <- order(x)  
 x1 <- x[o]  
 x4 <- step3(x, div)  
 blanks <- paste0(rep(" ", 18), collapse="")  
 x4[2\*div[1]] <- glue(  
 "{blanks}",  
 "({x1[div[1]]} + {x1[div[2]]}) / 2",  
 " = {(x1[div[1]]+x1[div[2]])/2}")  
 return(x4)  
}

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

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

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

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

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

Just average the two middle values.

## Break

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

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

## Criticisms of the mean and median

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

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 predition 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.

## Use of the mean for ordinal data

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

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

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

## An example of ordinal data.

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

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

“Do you agree or disagree with the following statements”

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

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

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

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

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

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

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

## Another example of ordinal data, course grades

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

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

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

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

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

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

|  |
| --- |
| Figure 10: Excerpt from Gould 1985 publication |

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.

## Choosing between the mean and median

* Often a source of controversy
* When do you use the mean?
  + When totals are important
* When do you use the median
  + When outliers/skewness might distort your conclusions
* Often, either is fine

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.

The mean allows extrapolation to totals. This is often important in the analysis of the economic effects of illness.

|  |
| --- |
| Figure 11: Exceprt from Bridge and McKenzie 2001, PMID: 11405531 |

## Bridge 2001, PMID: 11405531 (continued)

The measurement of airway resistance by the interrupter technique (Rint) needs standardization. Should measurements be made be during the expiratory or inspiratory phase of tidal breathing? **In reported studies, the measurement of Rint has been calculated as the median or mean of a small number of values, is there an important difference?**

## Bridge 2001, PMID: 11405531 (continued)

In the present data the mean of a set of values contributing to a measurement was not significantly different from the median. **However, the use of the median has been recommended since it is less affected by possible outlying values such as might be included by fully automated equipment.**

|  |
| --- |
| Figure 12: Chen et al 2019 |

## 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.**

## 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.

## Break

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

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

data.frame(x=rpois(2000, 20)) %>%  
 mutate(cut=quantile(x, probs=0.75)) %>%  
 mutate(y=1+as.numeric(x<=cut)) %>%  
 ggplot(aes(x, fill=factor(y))) +  
 scale\_fill\_manual(values=c("white", "gray")) +  
 theme(legend.position="none") +  
 xlab(" ") +  
 ylab(" ") +  
 geom\_bar(color="black", width=1) -> bar1  
ggsave("../images/illustrate-percentiles.png")

Saving 5 x 4 in image

|  |
| --- |
| Figure 13: Illustration of the 75th percentile |

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

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

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: or lower quartile
  + Median/50th percentiles: or
  + 75th percentile: or upper quartile

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

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

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

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

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

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

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

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

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

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 what I mean is percentiles other than 50%.

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

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

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

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

## Break

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

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

## Standard deviation

At least one alternative formulas.

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

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

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

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

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

## Standard deviation

At least one alternative formulas.

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 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

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.

x <- c(rnorm(1200, 4, 1.5), rnorm(800, 10, 1.5))  
x <- pmax(0, round(x))  
data.frame(x) %>%  
 ggplot(aes(x)) +  
 theme(legend.position="none") +  
 xlab(" ") +  
 ylab(" ") +  
 geom\_bar(color="black", fill="white", width=1) -> bimodal\_histogram  
ggsave("../images/bimodal-histogram.png")

Saving 5 x 4 in image

|  |
| --- |
| Figure 14: Bimodal histogram, not a bell shaped curve |

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.

data.frame(x=rpois(2000, 5)) %>%  
 ggplot(aes(x)) +  
 theme(legend.position="none") +  
 xlab(" ") +  
 ylab(" ") +  
 geom\_bar(color="black", fill="white", width=1) -> skewed\_histogram  
ggsave("../images/skewed-histogram.png")

Saving 5 x 4 in image

|  |
| --- |
| Figure 15: Skewed histogram, not a bell shaped curve |

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

data.frame(x=round(runif(2000, 0, 15))) %>%  
 ggplot(aes(x)) +  
 theme(legend.position="none") +  
 xlab(" ") +  
 ylab(" ") +  
 geom\_bar(color="black", fill="white", width=1) -> uniform\_histogram  
ggsave("../images/uniform-histogram.png")

Saving 5 x 4 in image

|  |
| --- |
| Figure 16: Uniform histogram, not a bell shaped curve |

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.

x <- rt(1000, 2)  
x <- 15 + c(x, -x) + rnorm(1000)  
x <- pmin(pmax(0, x), 30)  
   
data.frame(x=round(x)) %>%  
 ggplot(aes(x)) +  
 theme(legend.position="none") +  
 xlab(" ") +  
 ylab(" ") +  
 geom\_bar(color="black", fill="white", width=1) -> heavy\_tailed\_histogram  
ggsave("../images/heavy-tailed-histogram.png")

Saving 5 x 4 in image

|  |
| --- |
| Figure 17: Heavy-tailed histogram, not a bell shaped curve |

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

x <- rnorm(2000, 10, 3)  
x <- pmax(0, x)  
   
data.frame(x=round(x)) %>%  
 ggplot(aes(x)) +  
 theme(legend.position="none") +  
 xlab(" ") +  
 ylab(" ") +  
 geom\_bar(color="black", fill="white", width=1) -> normal\_histogram  
ggsave("../images/normal-histogram.png")

Saving 5 x 4 in image

|  |
| --- |
| Figure 18: Bell-shaped histogram, finally! |

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

x <- round(x)  
y <- (x<7) + (x > 13)  
data.frame(x=round(x)) %>%  
 ggplot(aes(x, fill=factor(y))) +  
 theme(legend.position="none") +  
 xlab(" ") +  
 ylab(" ") +  
 scale\_fill\_manual(values=c("gray", "white")) +  
 geom\_bar(color="black", width=1) -> one\_s\_histogram  
ggsave("../images/one-s-histogram.png")

Saving 5 x 4 in image

|  |
| --- |
| Figure 19: Percentage within one standard deviation |

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.

x <- round(x)  
y <- (x<4) + (x > 16)  
data.frame(x=round(x)) %>%  
 ggplot(aes(x, fill=factor(y))) +  
 theme(legend.position="none") +  
 xlab(" ") +  
 ylab(" ") +  
 scale\_fill\_manual(values=c("gray", "white")) +  
 geom\_bar(color="black", width=1) -> two\_s\_histogram  
ggsave("../images/two-s-histogram.png")

Saving 5 x 4 in image

|  |
| --- |
| Figure 20: Percentage within two standard deviations |

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 one standard deviation of the mean. This is only true if the variation follows a bell shaped curve.

x <- round(x)  
y <- (x<1) + (x > 19)  
data.frame(x=round(x)) %>%  
 ggplot(aes(x, fill=factor(y))) +  
 theme(legend.position="none") +  
 xlab(" ") +  
 ylab(" ") +  
 scale\_fill\_manual(values=c("gray", "white")) +  
 geom\_bar(color="black", width=1) -> three\_s\_histogram  
ggsave("../images/three-s-histogram.png")

Saving 5 x 4 in image

|  |
| --- |
| Figure 21: Percentage within three standard deviations |

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 one standard deviation of the mean. 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:

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.

## Diagnosing distributional issues (1/2)

* For all data
  + not midway between and
  + not midway between min and max

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
* For data with an lower and/or upper bound
  + $Q1 = lower bound
  + $Q3 = upper bound
* Don’t overdiagnose, especially with small sample sizes!

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.

|  |
| --- |
| Figure 22: Lin et al 2022, PMID: 36126916 |

|  |
| --- |
| Figure 23: Excerpt from Table 1 of Lin et al 2022: ages |

|  |
| --- |
| Figure 24: Excerpt from Table 1 of Lin et al 2022: CCI |

CCI is short for the Charlson Comorbidity Index.

|  |
| --- |
| Figure 25: Excerpt from Table 1 of Lin et al 2022: PHQ-2 |

PHQ-2 is a quality of life measure.

|  |
| --- |
| Figure 26: Tosato et al 2021, PMID: 34352201 |

## 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.

## 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).

## Ielapi 2021, PMID: 34968328

|  |
| --- |
| Figure 27: 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.

## Ielapi 2021, PMID: 34968328 (continued)

Excerpt from Table 1.

Data reported as mean ± standard deviation or median [Q1-Q3]

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

## Break

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

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

|  |
| --- |
| Entries in the electronic health record by job title |

Interpret this bar chart.

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

Interpret this bar chart.

|  |
| --- |
| Figure 29: Fastest computers by country |

Interpret this bar chart.

## Visualization

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

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

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

|  |
| --- |
| Figure 30: Survivors among first class passengers |

|  |
| --- |
| Figure 31: Survivors among second class passengers |

|  |
| --- |
| Figure 32: Survivors among third class passengers |

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

|  |
| --- |
| Figure 33: Bar chart showing proportion of passenger classes among deaths and survivors |

Note

|  |
| --- |
| Bar chart showing proportion of deaths and survivors among passenger classes |

Note

|  |
| --- |
| Bar chart showing only proportion of survivors among passenger classes |

Note

|  |
| --- |
| Bar chart showing proportion of survivors among passenger classes and sex |

Note

|  |
| --- |
| Bar chart showing proportion of survivors among sex and passenger classes |

Note

|  |
| --- |
| Figure 34: Bar chart showing average age among deaths and survivors |

Note

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

Note

|  |
| --- |
| Figure 35: Boxplot showing ages of deaths and survivors |

Note

|  |
| --- |
| Figure 36: Annotated boxplot |

Note

|  |
| --- |
| Figure 37: Histogram of ages of passengers that died |

|  |
| --- |
| Figure 38: Histogram of ages of passengers that survived |

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

|  |
| --- |
| Figure 39: SCatter plot of age vs survived |

Note

|  |
| --- |
| Figure 40: Scatter plot of age vs survived with jittering |

Note

|  |
| --- |
| Article feedback by type of article |

Comment on these boxplots

Wikipedia

|  |
| --- |
| Figure 41: Emergency care course test scores |

Comment on these boxplots

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

|  |
| --- |
| Figure 42: Self reported versus observed adherence |

Comment on these boxplots

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

|  |
| --- |
| Figure 43: Annual rainfall in selected cities |

Comment on these boxplots

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

## Which visualization to choose?

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

Note

|  |
| --- |
| Figure 44: Color combinations |

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

## The RGB color system

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

On the computer, you use the RGB color system.

|  |
| --- |
| Figure 45: Red plus green equals yellow |

|  |
| --- |
| Figure 46: Red plus blue equals magenta |

|  |
| --- |
| Figure 47: Green plus blue equals cyan |

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

|  |
| --- |
| Figure 48: Yellow plus blue equals white |

|  |
| --- |
| Figure 49: Magenta plus green equals white |

|  |
| --- |
| Figure 50: Cyan plus red equals white |

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

## The color cube

|  |
| --- |
| Figure 51: Illustration of the color cube |

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

|  |
| --- |
| Figure 52: The red to green gradient on the color cube |

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

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

## Rainbow

|  |
| --- |
|  |

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

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

## The color cylinder

|  |
| --- |
| Figure 53: Color cylinder |

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

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

|  |
| --- |
| Figure 54: Various foreground and background color combinations |

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

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

|  |
| --- |
| Figure 55: A brighter version of the rainbow |

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

## Darker rainbow

|  |
| --- |
| Figure 56: A darker version of the rainbow |

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

|  |
| --- |
| Figure 57: Color combinations using darker foregrounds and lighter backgrounds |

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

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

|  |
| --- |
| Figure 58: DIffering luminance values of the rainbow |

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

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

data.frame(  
 x=runif(100),   
 y=runif(100),   
 z=sample(1:6, 100, replace=TRUE)) %>%  
 ggplot(aes(x, y, color=factor(z))) +  
 scale\_color\_manual(  
 values=c(  
 "red",  
 "green",  
 "blue",  
 "yellow",  
 "magenta",  
 "cyan")) +  
 theme(  
 panel.background =   
 element\_rect(  
 fill = "white",  
 colour = "white",  
 size = 0.5,   
 linetype = "solid"),  
 panel.grid.major =   
 element\_line(  
 size = 0.5,   
 linetype = 'solid',  
 colour = "gray"),   
 panel.grid.minor =   
 element\_line(  
 size = 0.25,   
 linetype = 'solid',  
 colour = "gray")) +  
 theme(legend.position="none") +  
 xlab(" ") +  
 ylab(" ") +  
 geom\_point() -> scatter\_white  
ggsave("../images/scatter-white.png")

Saving 5 x 4 in image

|  |
| --- |
| Figure 59: Rainbow colors on a white background |

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

scatter\_white +  
 theme(  
 panel.background =   
 element\_rect(  
 fill = "black",  
 colour = "black",  
 size = 0.5,   
 linetype = "solid")) +  
 theme(legend.position="none") -> scatter\_black  
ggsave("../images/scatter-black.png")

Saving 5 x 4 in image

|  |
| --- |
| Figure 60: Rainbow colors on a black background |

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

|  |
| --- |
| Figure 61: Rainbow colors showing a banding effect |

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

So what colors do I recommend?

|  |
| --- |
| Figure 62: One good set of color choices for nominal data |

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

|  |
| --- |
| Examples of light to dark gradients |

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

|  |
| --- |
| Figure 63: Examples of diverging gradients |

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

## Color blindness

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

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

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

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

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

|  |
| --- |
| Figure 64: Clothing mistake: using too many colors |

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

|  |
| --- |
| ADvertisement with a single red umbrella |

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

The Travelers - Look under the Umberella commercial (1986). Retrieved 2019-09-07 from https://www.youtube.com/watch?v=3zQX66jd\_c0

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

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

|  |
| --- |
| Figure 65: Use of color to highlight a single individual |

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

This has practical implications for data visualization.

|  |
| --- |
| Figure 66: How many “5’s” are in this figure? |

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

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

PAUSE HERE

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

|  |
| --- |
| Figure 67: Repeat question. How many “5’s” are in this figure? |

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

## Repeat quiz question 1

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

This data table shows counts and …

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

Here is the question I asked earlier about percentages.

## Repeat quiz question 2

The median might be preferred to the mean if

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

Here is a repeat question about the median.

## Repeat quiz question 3

The problem with error bars is that they

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

Here is the repeat question about error bars