Analysis of Psychological Data

Lab 2. Play the Game with Data: Visualization, Distribution, and Central Tendency

Ihnwhi Heo (iheo2@ucmerced.edu)
Quantitative Methods, Measurement, and Statistics
Website: https://ihnwhiheo.github.io
Office: https://ucmerced.zoom.us/j/2093557522 (Friday 1:30 - 3:30 pm)

What are we going to do?

Recap to give you a big picture

Scales of measurement

Data visualization

Distribution

Central tendency

Group activity

Eyes on the variables

Think about the following:

What is the type of the variable GPA? Is it categorical or numerical?

Eyes on the variables

To be categorical or numerical, that's the problem

If measured on a letter grade (e.g., A, B, C), it is categorical (specifically, polynomial)

If measured on a numerical grade (e.g., 4.0, 3.2, 3.7), it is numerical (specifically, continuous)

To fully define and understand a variable, we need to know how that variable is **measured**.

Scaels of measurement

The measurement scale

Type of information provided by the values of a variable

Four scales of measurement

Nominal	**Ordinal**	**Interval**	**Ratio**
Named	Named	Named	Named
	Ordered	Ordered	Ordered
		Equal interval	Equal interval
			True 0 value
Gender	*Letter grade (A~F)*	*Temperature*	*Height/Weight*
Ethnicity	*Political ideology*	*IQ Scores*	*Age*
Color	*Likert scale*	*SAT Scores*	*Income*

Data visualization

Why do we visualize data?

Once we collect data about variables of interest, we want to present important information from the data!

What did we learn?

- Frequency table
  - Simple frequency
  - Cumulative frequency
  - Relative frequency
  - Cumulative relative frequency
- Bar chart
- Pie chart
- Histogram
- Scatterplot

Frequency table

To show the count or the percentage of data points

How can we distinguish different kinds of frequencies

	Count	**Percentage**
Non Accumulation	Simple frequency	Relative frequency
Accumulation	Cumulative frequency	Cumulative relative frequency

Data visualization

Bar chart

We use rectangular bars to represent the count or the proportion of categorical data

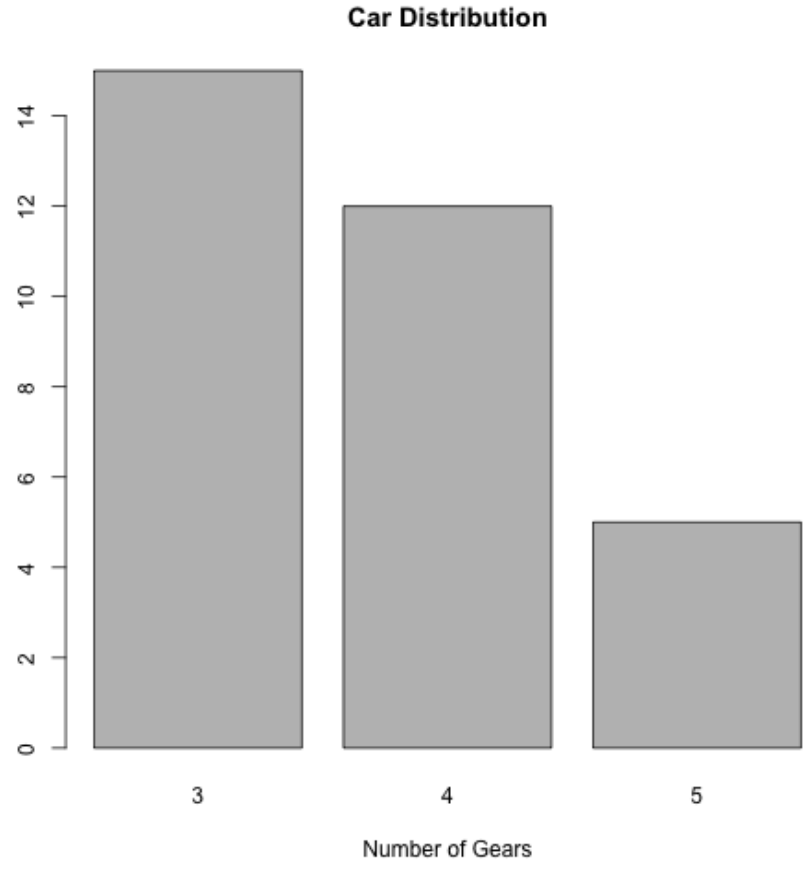
Shall we see some examples?

R code from https://www.statmethods.net/

Data visualization

Can you understand the plot?

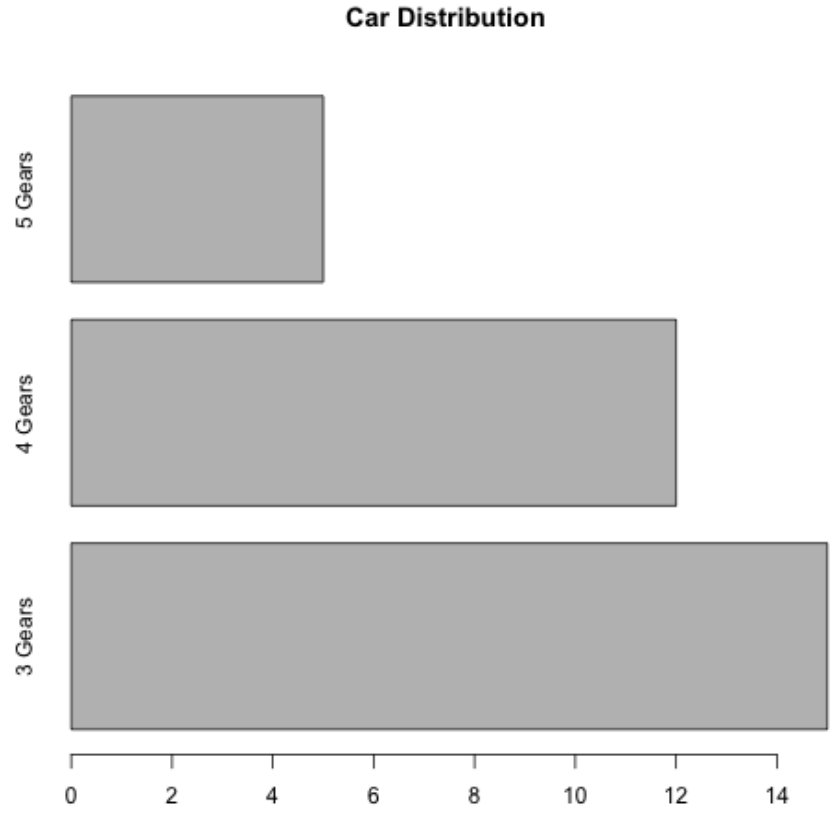
```r
# Simple Bar Plot
counts <- table(mtcars$gear)
barplot(counts, main="Car Distribution",
   xlab="Number of Gears")
```



Data visualization

Horizontal bars are possible.

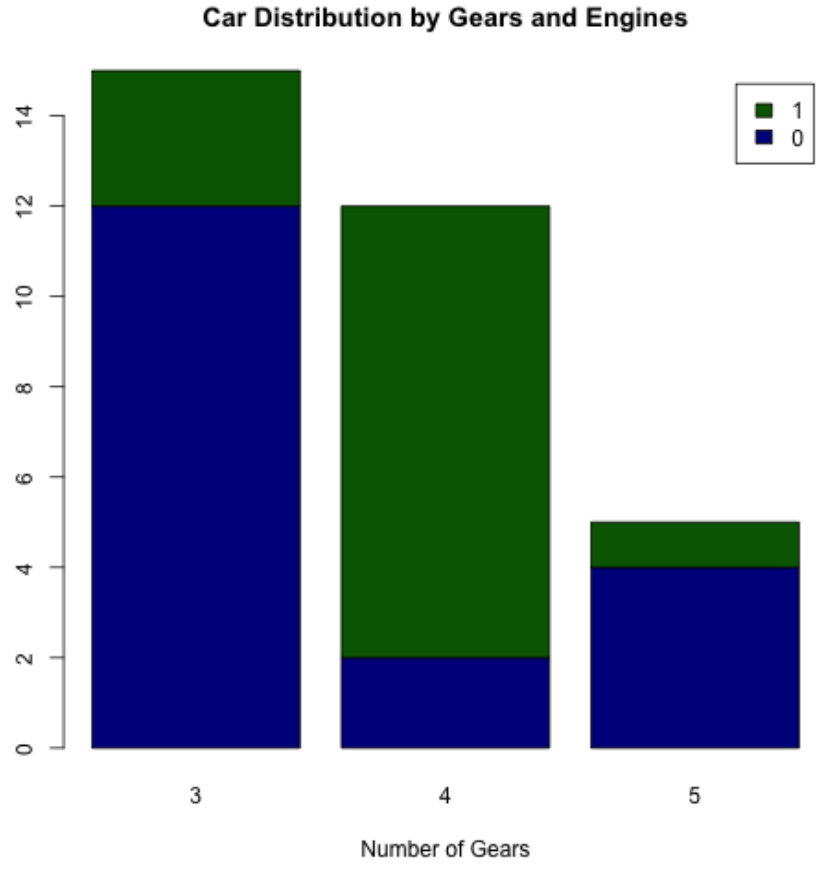
```
# Simple Horizontal Bar Plot with Added Labels
counts <- table(mtcars$gear)
barplot(counts, main="Car Distribution",
        horiz=TRUE,
  names.arg=c("3 Gears", "4 Gears", "5 Gears"))
```



Data visualization

You can even stack the bars!

```
# Stacked Bar Plot with Colors and Legend
counts <- table(mtcars$vs, mtcars$gear)
barplot(counts,
  main="Car Distribution by Gears and Engines",
  xlab="Number of Gears",
  col=c("darkblue","darkgreen"),
  legend = rownames(counts))
```



Data visualization

Pie chart

We use circular pies to represent the count or the proportion of categorical data

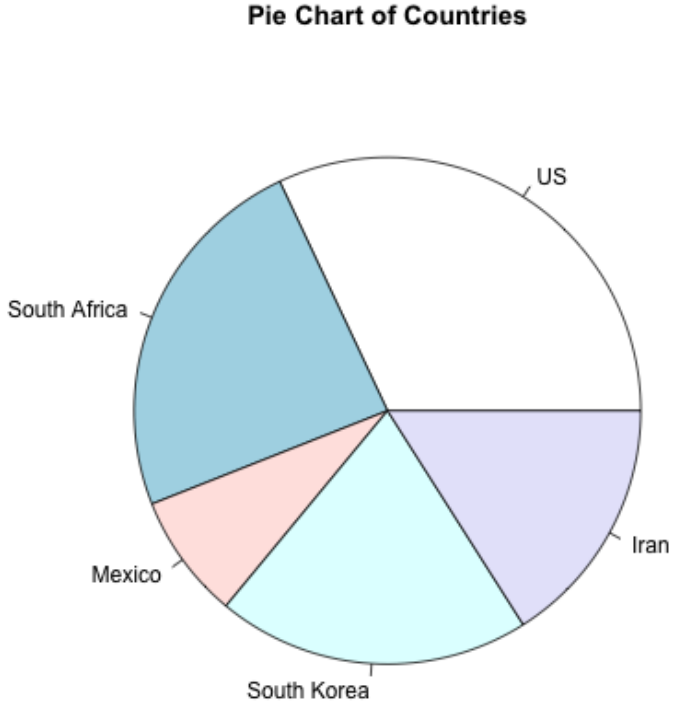
Shall we see some examples?

R code from https://www.statmethods.net/

Data visualization

Can you understand the plot?

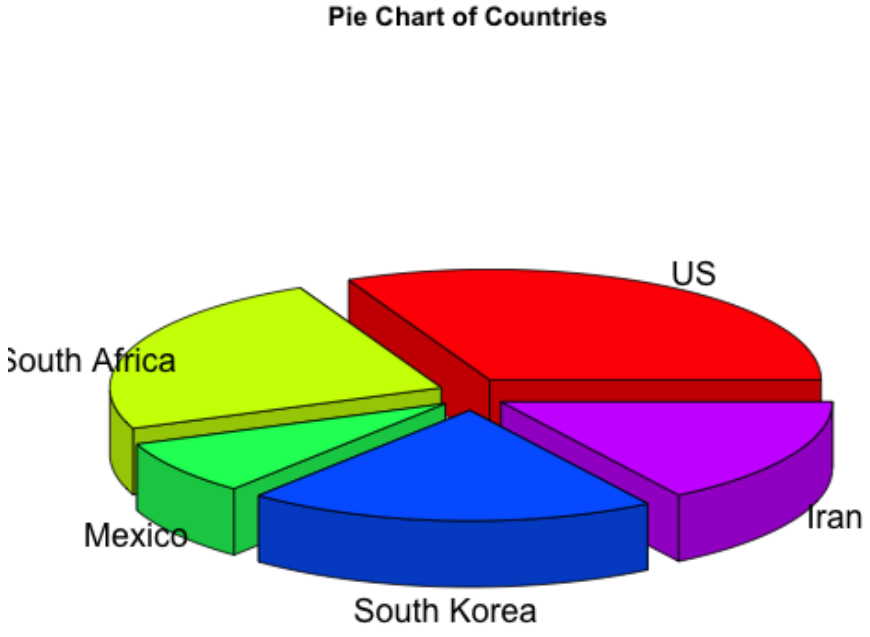
```
# Simple Pie Chart
slices <- c(16, 12, 4, 10, 8)
lbls <-
  c("US", "South Africa", "Mexico",
    "South Korea", "Iran")
pie(slices, labels = lbls,
    main="Pie Chart of Countries")
```



Data visualization

Let me present the 3D plot.

```r
# 3D Exploded Pie Chart
library(plotrix)
slices <- c(16, 12, 4, 10, 8)
lbls <- c("US", "South Africa",
          "Mexico", "South Korea", "Iran")
pie3D(slices,labels=lbls,explode=0.1,
   main="Pie Chart of Countries ")
```



Data visualization

Histogram

We use bins or buckets to represent the frequency distribution of data

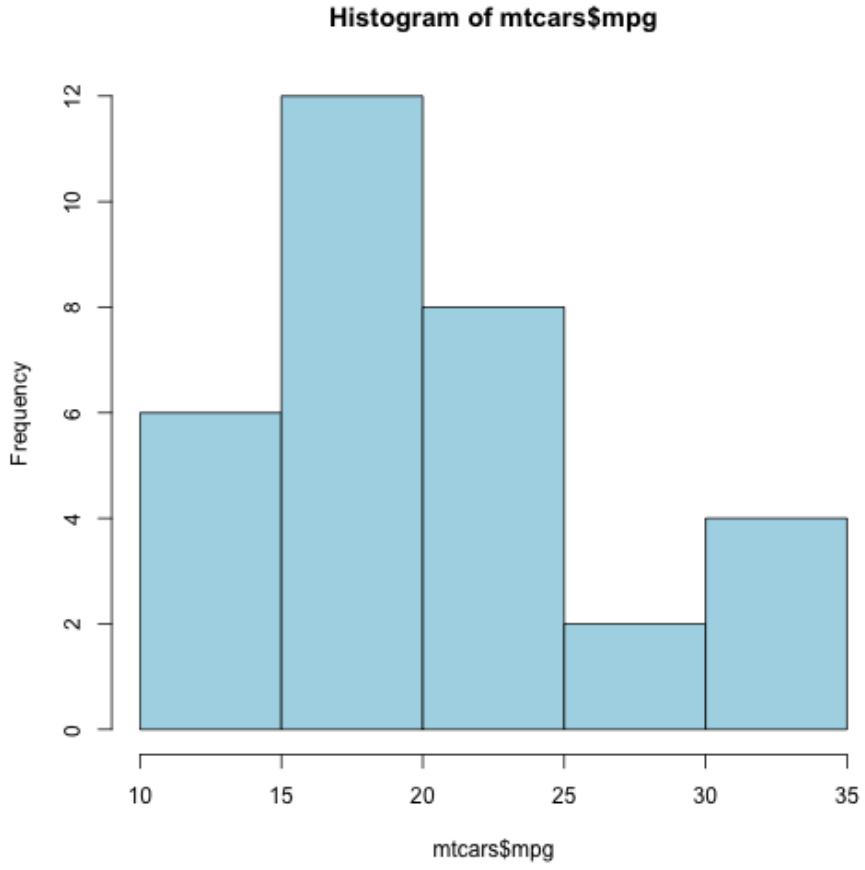
Shall we see some examples?

R code from https://www.statmethods.net/

Data visualization

Can you understand the plot?

```r
# Simple Histogram
hist(mtcars$mpg, col="lightblue")
# mpg stands for miles per gallon
```

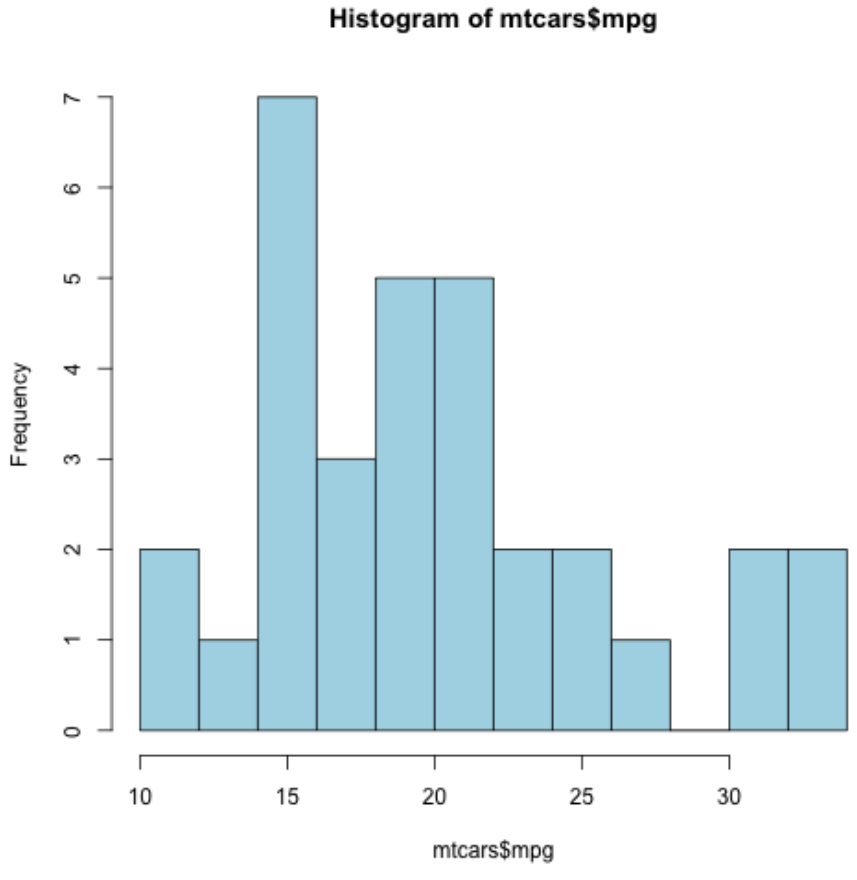


Data visualization

I increased the number of breaks.

```
# Colored Histogram with Different Number of Bins
hist(mtcars$mpg, breaks=12, col="lightblue")
```

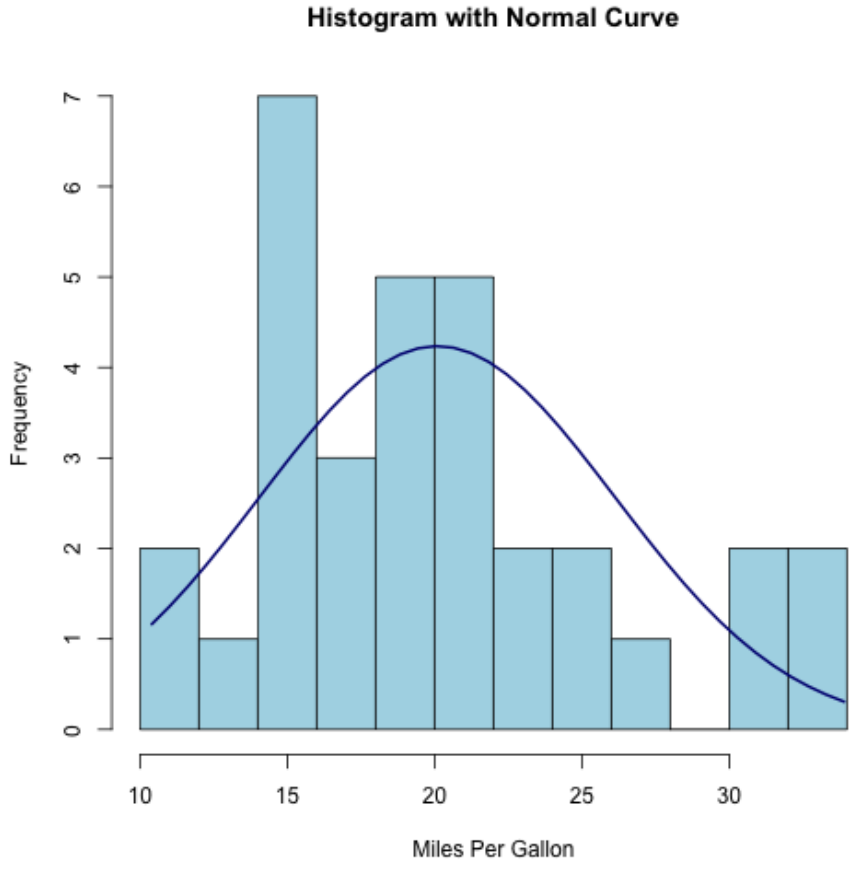
Depending on the number of breaks, we can either gain or lose important information inherent in data.



Data visualization

Superimpose the curve that fits the data!

```
# Add a Normal Curve (Thanks to Peter Dalgaard)
x <- mtcars$mpg
h<-hist(x, breaks=10, col="lightblue",
   xlab="Miles Per Gallon",
   main="Histogram with Normal Curve")
xfit<-seq(min(x),max(x),length=40)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="navy", lwd=2)
```



Data visualization

Scatterplot

We scatter data points to see the relationship between the variables

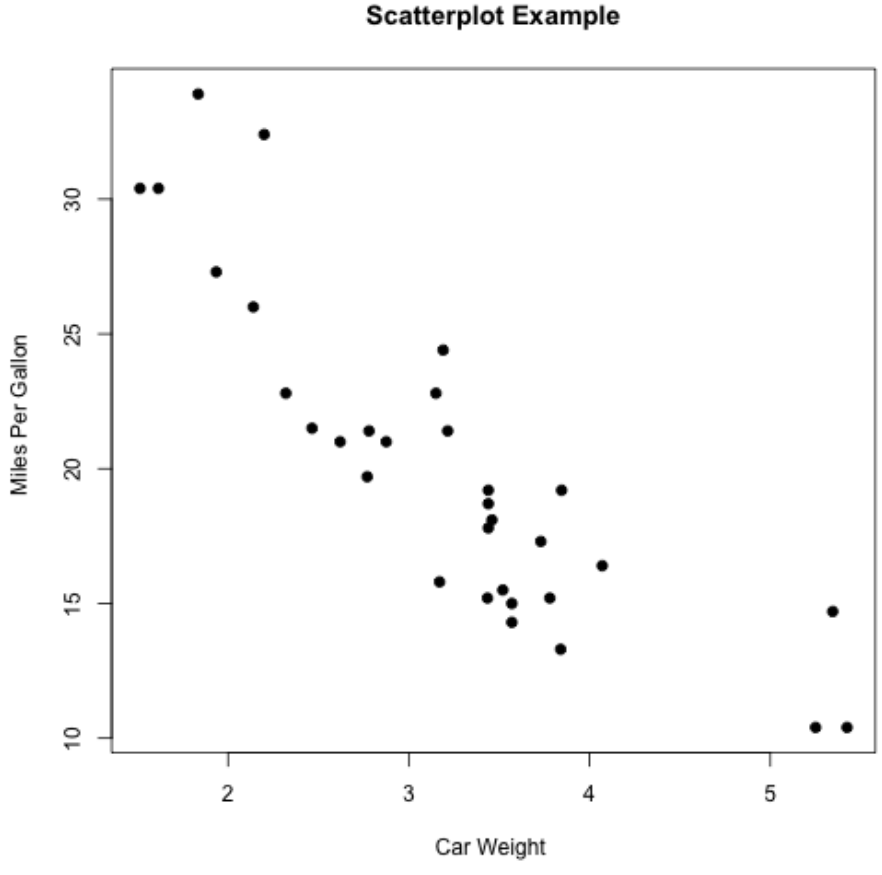
Shall we see some examples?

R code from https://www.statmethods.net/

Data visualization

Can you understand the plot?

```
plot(mtcars$wt, mtcars$mpg,
   main="Scatterplot Example",
   xlab="Car Weight ",
   ylab="Miles Per Gallon ", pch=19)
```

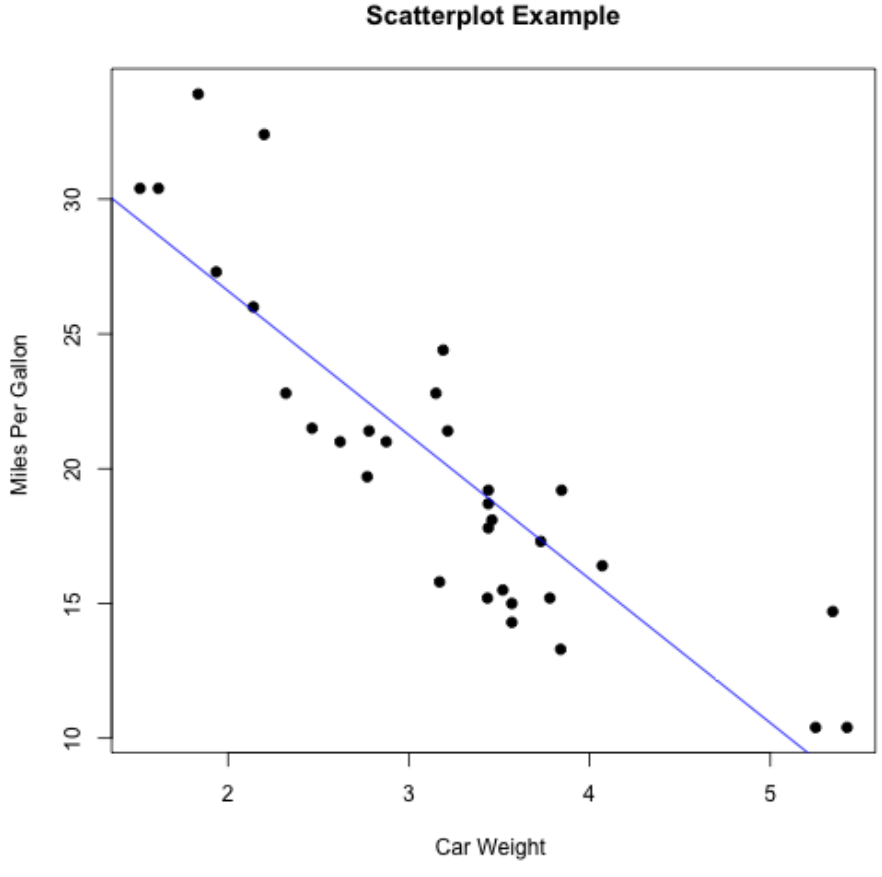


Data visualization

Adding a trend line helps!

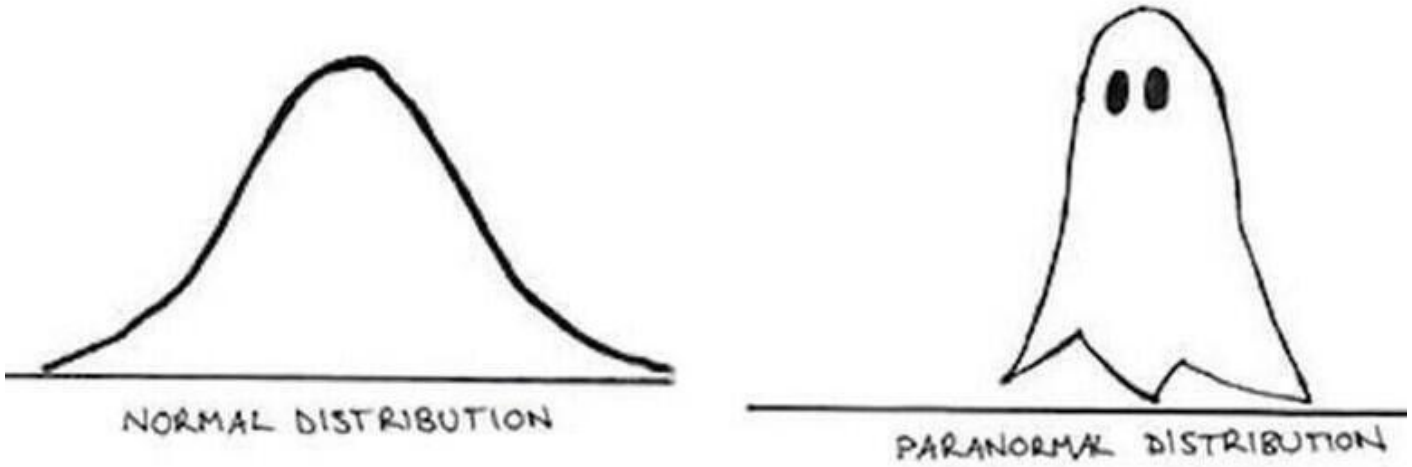
```
plot(mtcars$wt, mtcars$mpg,
   main="Scatterplot Example",
   xlab="Car Weight ",
   ylab="Miles Per Gallon ", pch=19)
abline(lm(mtcars$mpg~mtcars$wt), col="blue")
```

The trend line is actually a regression line.



Distribution

Nerdy stat jokes...



Distribution

Why does distribution matter?

An intuitive way to understand how different values of a variable are spread over

We learned...

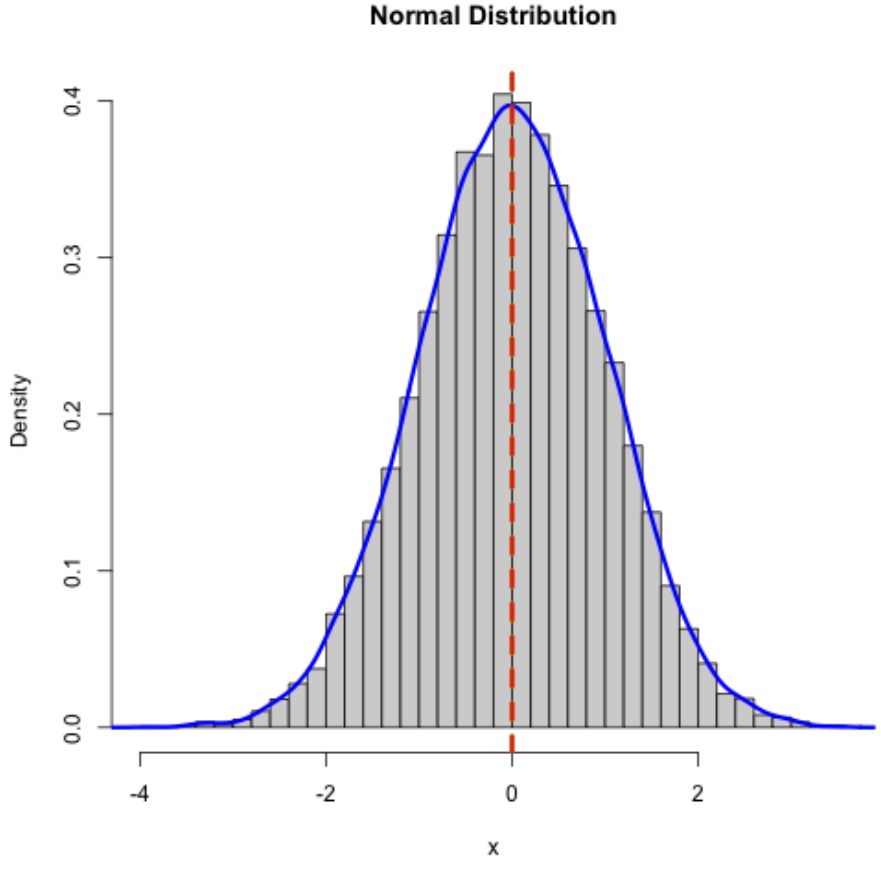
- Normal distribution
- Positive skewness (*aka*. positively skewed)
- Negative skewness (*aka*. negatively skewed)
- ~~Paranormal distribution~~

Distribution

Normal Distribution

```
set.seed(2093557522)
x = rnorm(10000,0,1)
hist(x, main="Normal Distribution", freq=FALSE,
     breaks=50)
lines(density(x), col='blue', lwd=3)
abline(v = c(mean(x),median(x)),
       col=c("green", "red"),
       lty=c(2,2), lwd=c(3, 3))
```

- Trailing off toward both left and right ends
- Any phenomenon that we would **normally** expect to observe
- Bell-shaped & symmetrical around the **center**
- Mean = Median = Mode
- Gaussian distribution (in honor of Carl Friedrich Gauss)

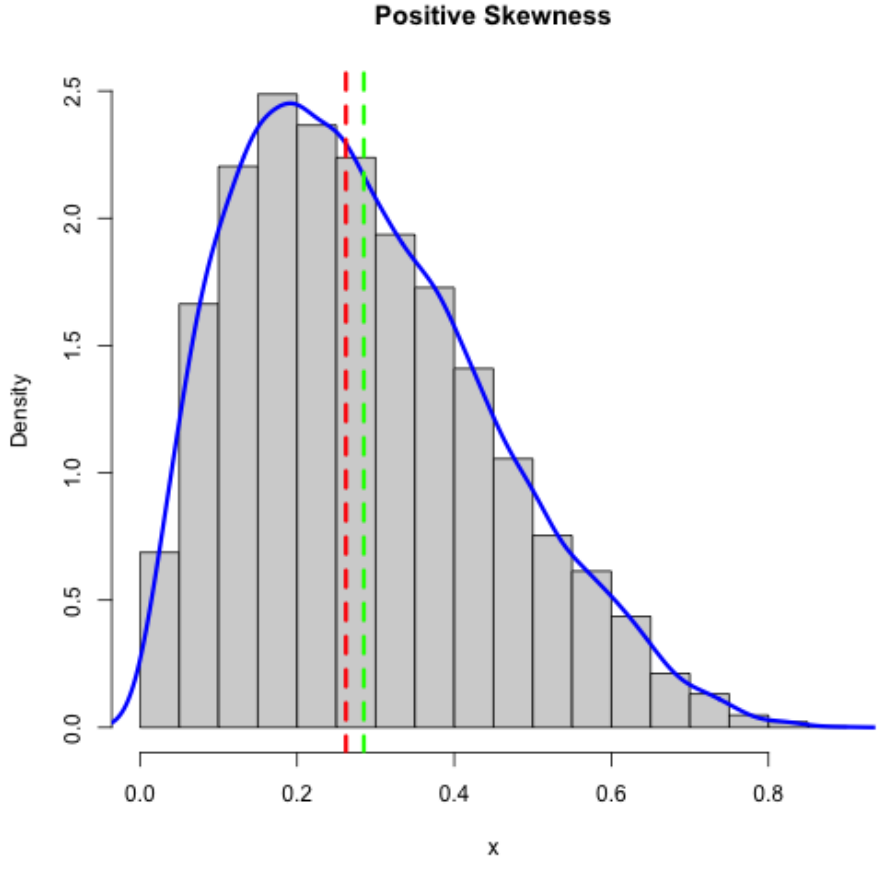


Distribution

Positive Skewness

```
set.seed(322)
x = rbeta(10000,2,5)
hist(x, main="Positive Skewness", freq=FALSE)
lines(density(x), col='blue', lwd=3)
abline(v = c(mean(x),median(x)),
       col=c("green", "red"),
       lty=c(2,2), lwd=c(3, 3))
```

- Trailing off toward the right end
- Mode < Median < Mean
- Score distribution of difficult exams

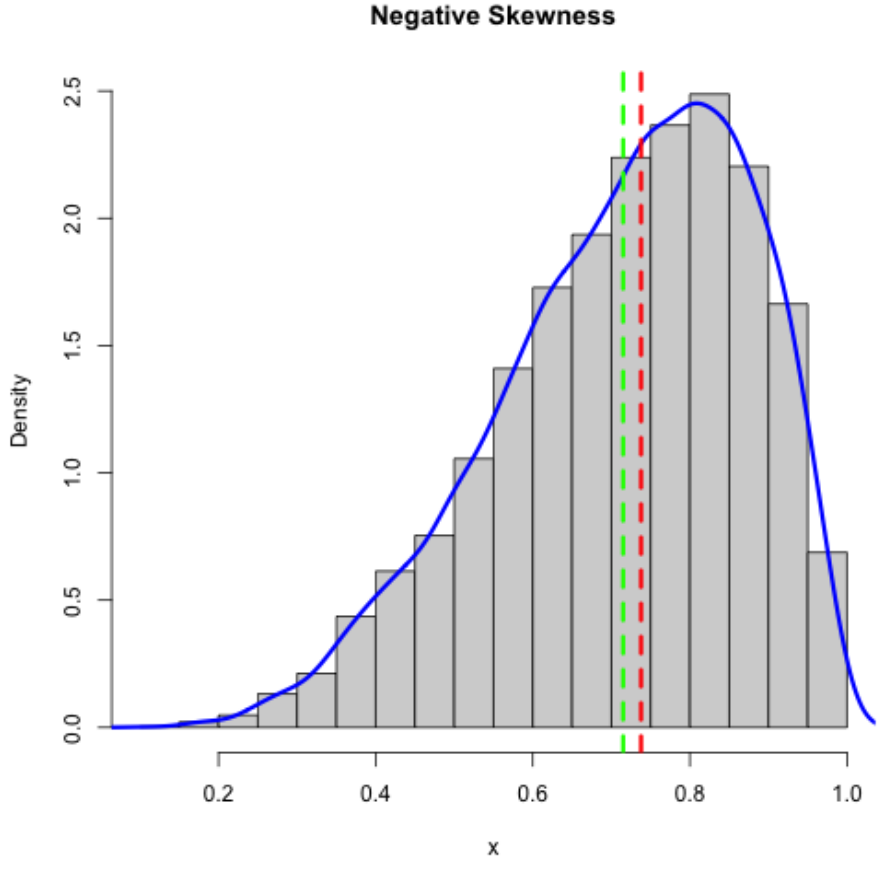


Distribution

Negative Skewness

```
set.seed(322)
x = rbeta(10000,5,2)
hist(x, main="Negative Skewness", freq=FALSE)
lines(density(x), col='blue', lwd=3)
abline(v = c(mean(x),median(x)),
       col=c("green", "red"),
       lty=c(2,2), lwd=c(3, 3))
```

- Trailing off toward the left end
- Mean < Median < Mode
- Score distribution of easy exams



Central tendency

Why do we focus on central tendency?

Reflects where 'values of a variable (i.e., distribution)' are centered

Three measures

- Mean
  - Sample mean = $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
  - Population mean = $\mu = \frac{\sum_{i=1}^{N} X_i}{N}$
- Median, think after arranging values in an ascending numerical order
  - Sample median = a value at the position $\frac{n+1}{2}$
- Mode = value that is the most frequent

Central tendency

Each measure is useful for different situations!

Mean

- Most widely used and the best guess
- Vulnerable to outliers (i.e., extreme values)
- Does not make sense to nominal and ordinal scales (Can you calculate the mean of different genders?)

Median

- Robust (i.e., less vulnerable) to outliers

Mode

- Very useful to nominal and ordinal scales

Group activity

Case 1

Estuardo is interested in how people in California give ratings to the taste of ranch pizza. He collected 20 people in Los Angeles and made a frequency table on the frequency of pizza ratings (1 through 5).

Ratings	Frequency (Freq.)	Cumulative Freq.	Relative Freq.	Cumulative Relative Freq.
1	3			
2	2			
3	5			
4	6			
5	4			

Can you answer below?

- Can you fill in the blanks of cumulative freq., relative freq., and cumulative relative freq.?
- What is the cumulative relative freq. for giving 2 stars or lower?

Group activity

Case 2

Ari has been working as a data scientist at Regal. Since Shang-Chi and the Legend of the Ten Rings is on the screen, she wants to know how people rate the movie. She made a questionnaire on which score you want to give to the movie. The score ranges from 0 to 100 on a 5-point scale. So far, she has collected responses from 19 people, which is summarized in the following table:

Scores	**65**	**70**	**75**	**80**	**85**	**90**	**95**	**100**
Frequency	1	1	1	2	3	2	4	5

Can you answer below?

- How does the distribution look like? Normally distributed, negatively skewed, or positively skewed?
- What are the mean, the median, and the mode of the sample? Feel free to use a calculator, if necessary.
- Can you calculate the population mean?
- Imagine Ari receives one response from a person who is very disappointed with the movie. If that person gives a score of 0, how do the mean, the median, and the mode change?

Before you go home...

Any questions or comments?

Thanks! Have a wonderful weekend!