Analysis of Psychological Data

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

Ihnwhi Heo (iheo2@ucmerced.edu)
Quantitative Methods, Measurement, and Statistics
Website: https://ihnwhiheo.github.io
Office: https://ucmerced.zoom.us/j/2093557522 (Thursday 3:30 - 5:30 pm)

What are we going to do?

Recap to give you a big picture

Data visualization

Distribution

Food for thought

Some announcements

Lab 3 on February 9 is in person

Check where your lecture room is!

HW 1 (20 points) deadline

February 3 (Thursday) at 11:59 pm on CatCourses

Attendance check

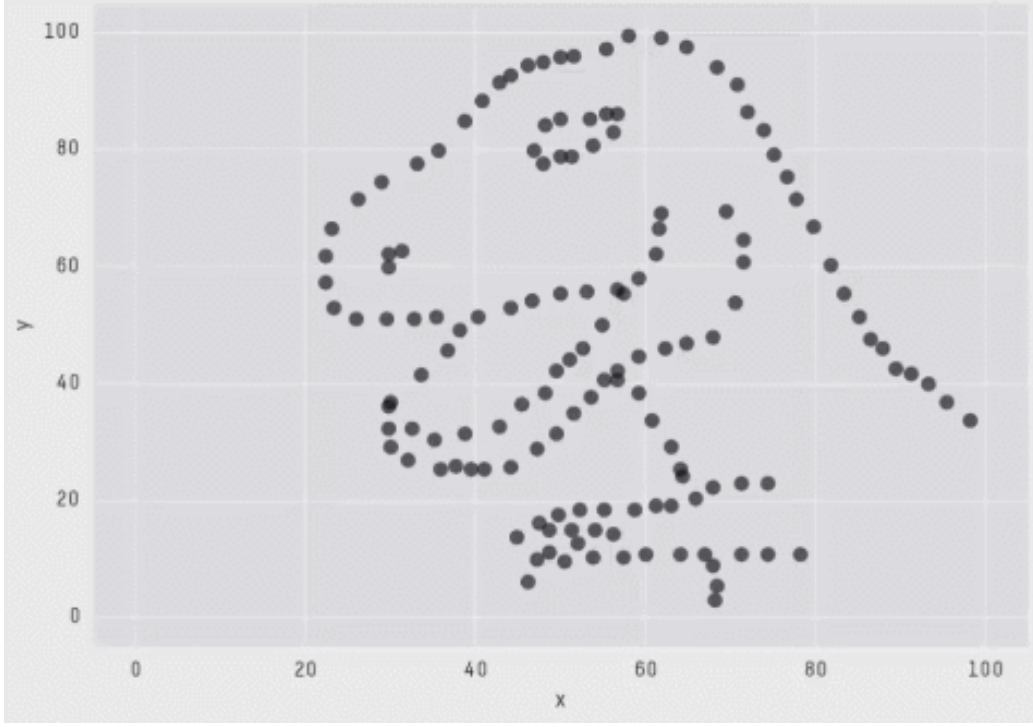
Find one example of data visualization in your life and let me know what it is to my email **iheo2@ucmerced.edu** (simple description is enough) - hint: NEWS? Instagram? Facebook?

Data visualization

You can find everywhere *visualized* data

Data visualization

You want something fun?!



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

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 distribution
  - Cumulative frequency distribution
  - Relative percent distribution
  - Cumulative percent distribution
- Bar chart
- Pie chart
- Histogram

Data visualization

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 distribution	Relative percent distribution
Accumulation	Cumulative frequency distribution	Cumulative percent distribution

Data visualization

Frequency table

Itamar, a data scientist at Marvel Studios, collected responses from 100 people to the question: "Spider-Man: No Way Home was great fun".

Response Category	Simple frequency	Relative percent	Cumulative frequency	Cumulative percent
Strongly agree	40	40%	40	40%
Agree	30	30%	70	70%
Disagree	20	20%	90	90%
Strongly disagree	10	10%	100	100%
Total	100	100%		

Data visualization

Bar chart

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

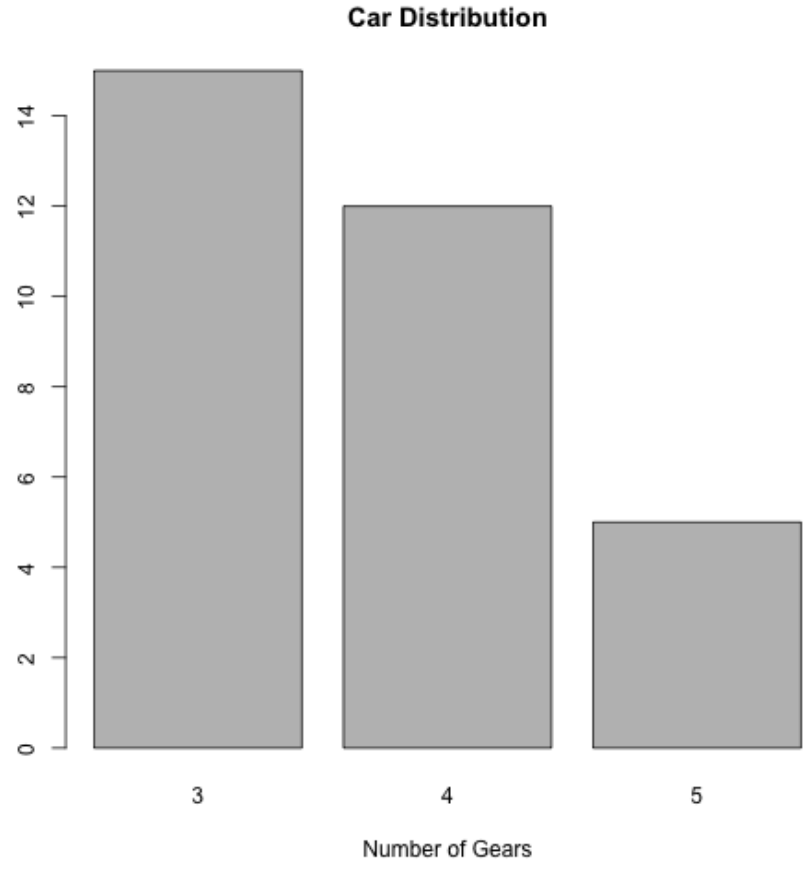
Shall we see some examples?

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

Data visualization

Can you understand the plot?

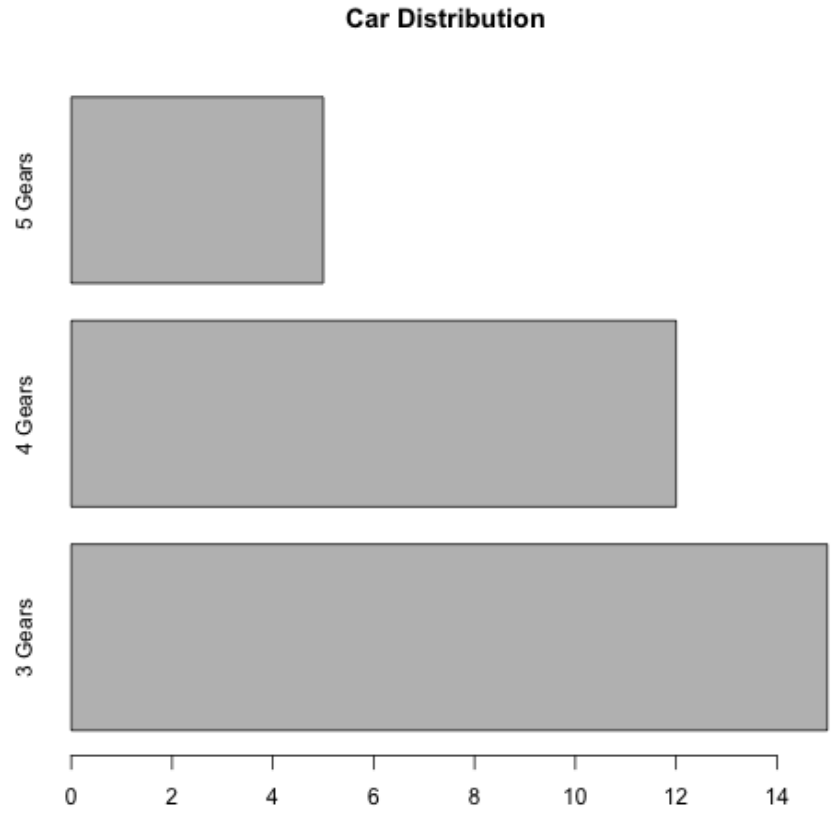
```r
# Simple Bar Plot
counts <- table(mtcars$gear)
barplot(counts, main="Car Distribution",
   xlab="Number of Gears")
```



Data visualization

Horizontal bars are possible.

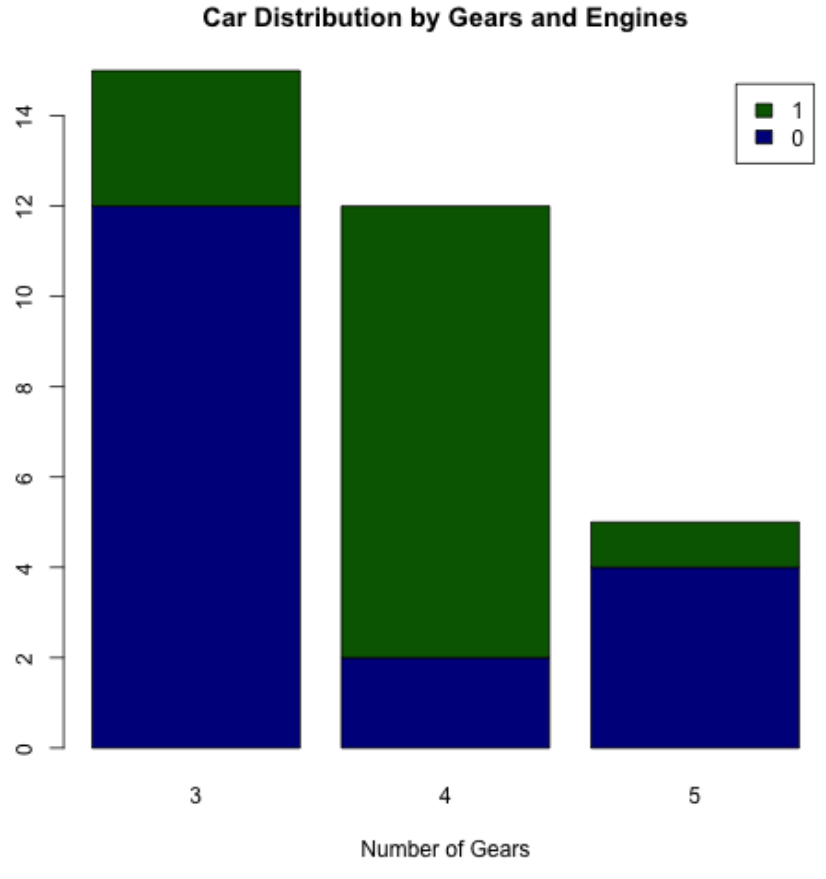
```
# Simple Horizontal Bar Plot with Added Labels
counts <- table(mtcars$gear)
barplot(counts, main="Car Distribution",
        horiz=TRUE,
  names.arg=c("3 Gears", "4 Gears", "5 Gears"))
```



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 qualitative data

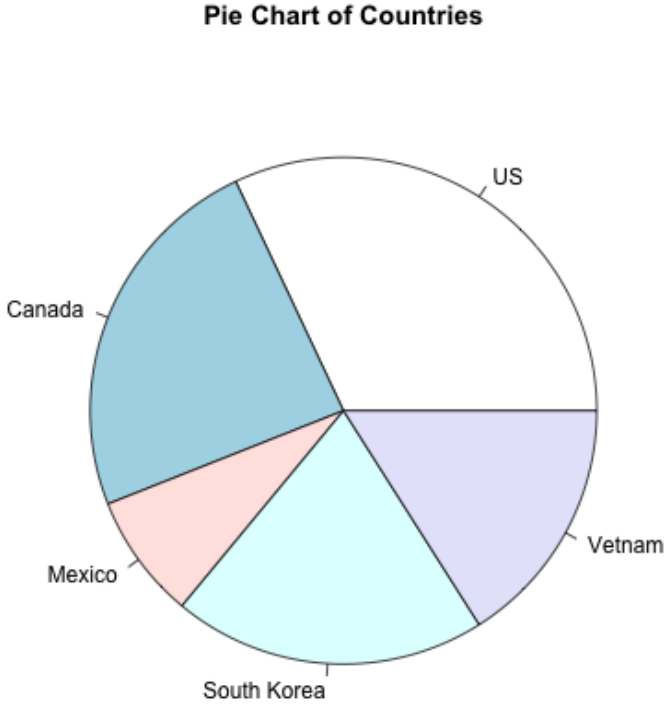
Shall we see some examples?

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

Data visualization

Can you understand the plot?

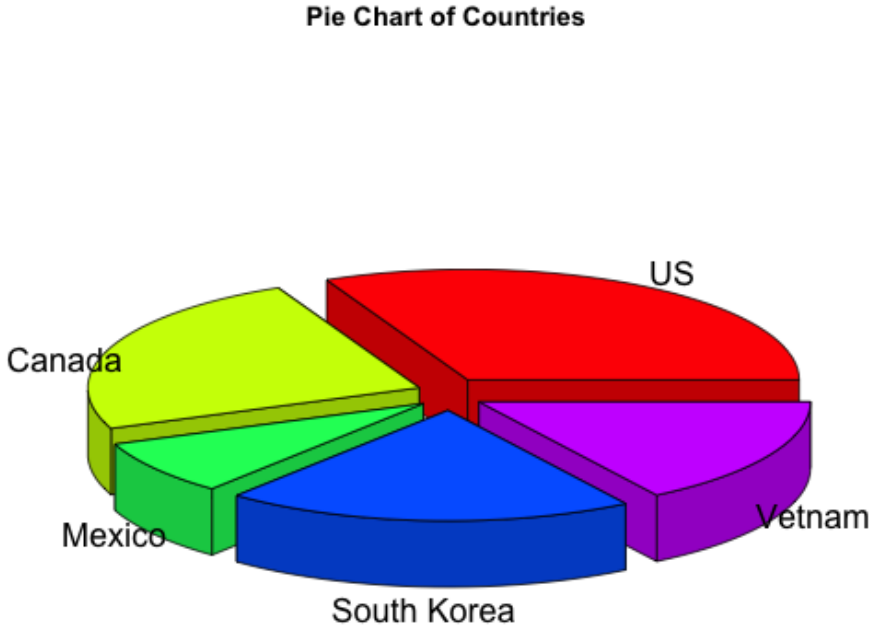
```r
# Simple Pie Chart
slices <- c(16, 12, 4, 10, 8)
lbls <-
  c("US", "Canada", "Mexico",
    "South Korea", "Vetnam")
pie(slices, labels = lbls,
    main="Pie Chart of Countries")
```



Data visualization

Let me present the 3D plot.

```
# 3D Exploded Pie Chart
library(plotrix)
slices <- c(16, 12, 4, 10, 8)
lbls <- c("US", "Canada", "Mexico",
    "South Korea", "Vetnam")
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

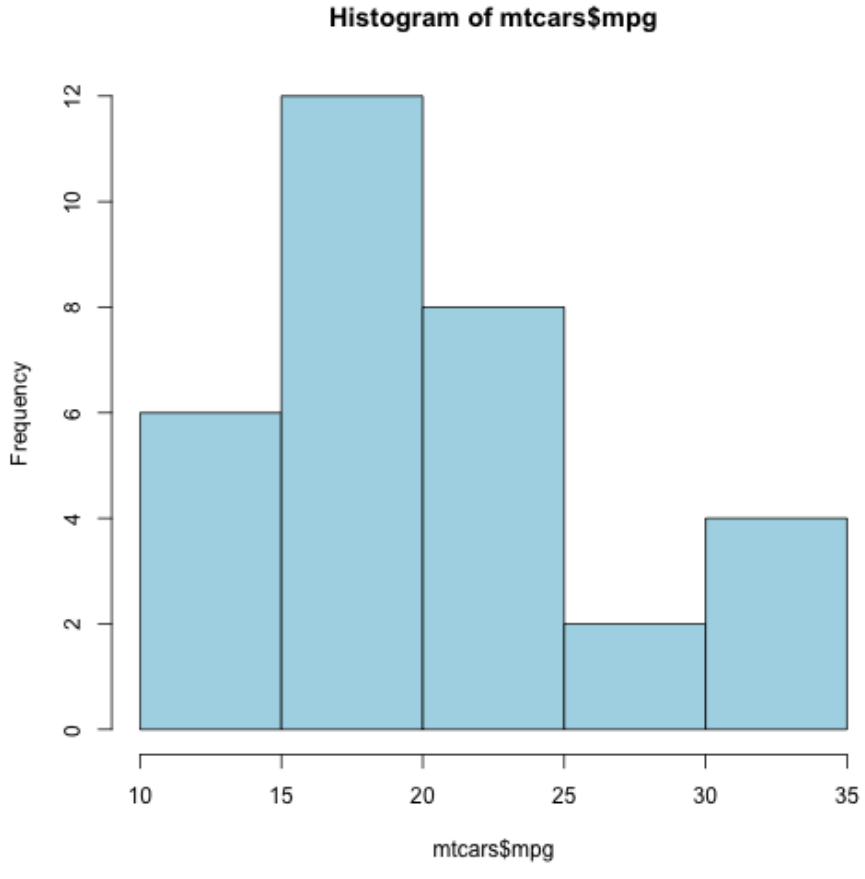
Shall we see some examples?

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
```

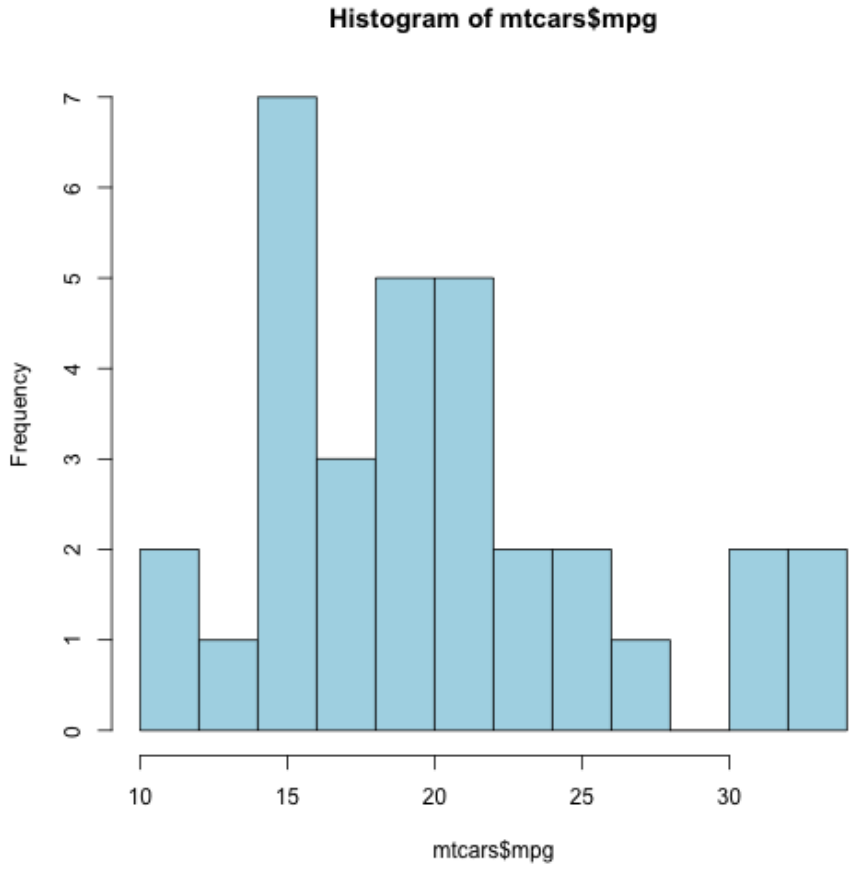


Data visualization

I increased the number of breaks.

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

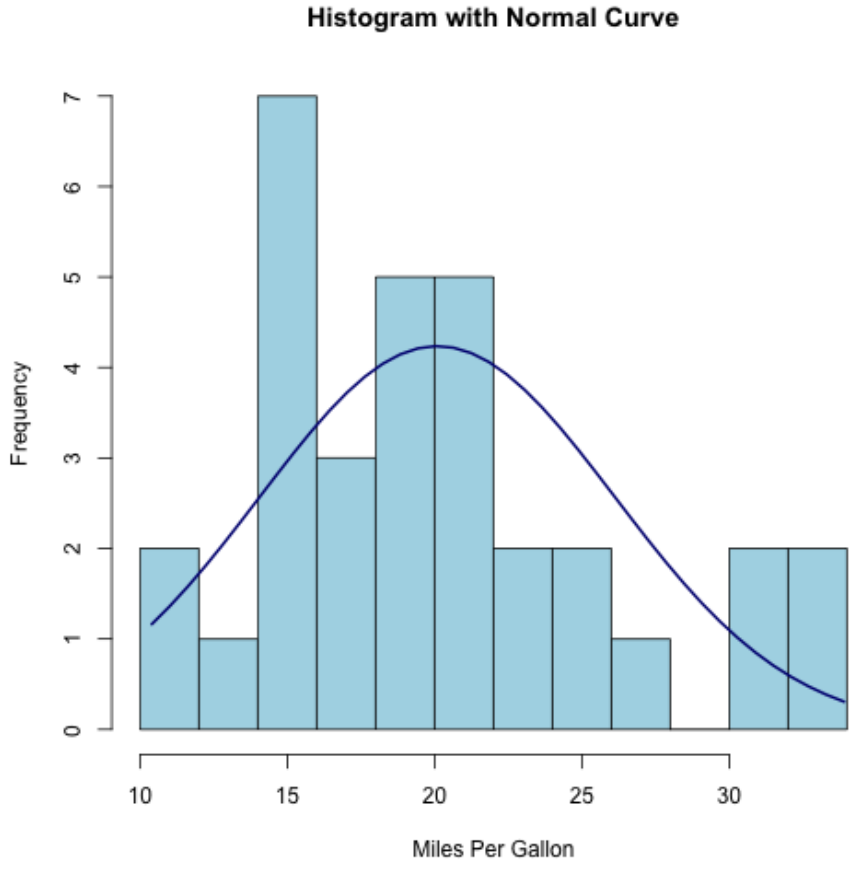
Depending on the number of breaks, we can either gain or lose important information inherent in data.



Data visualization

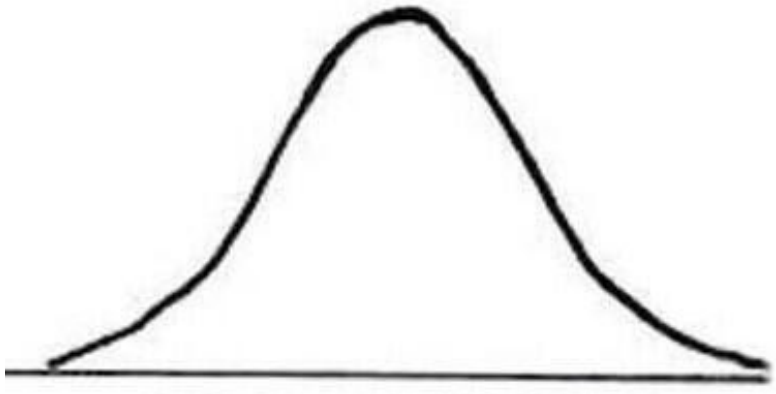
Superimpose the curve that fits the data! → Smoothing

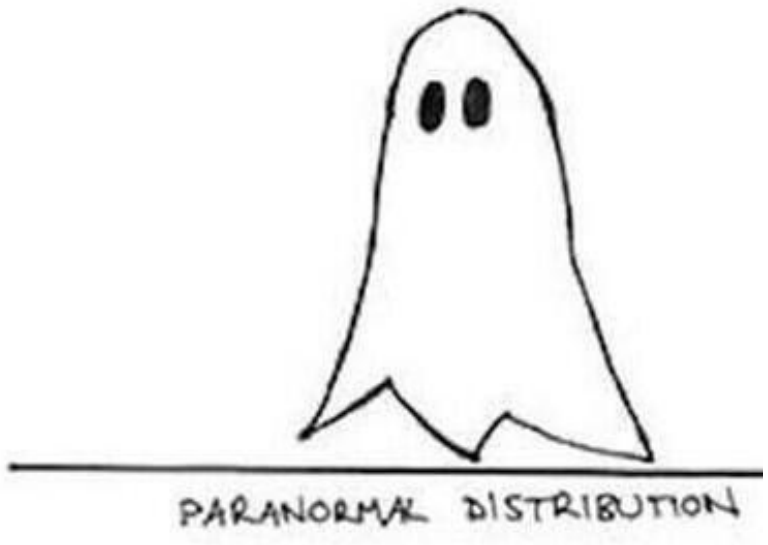
```
# Add a Normal Curve (Thanks to Peter Dalgaard)
x <- mtcars$mpg
h<-hist(x, breaks=10, col="lightblue",
   xlab="Miles Per Gallon",
   main="Histogram with Normal Curve")
xfit<-seq(min(x),max(x),length=40)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit,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



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...

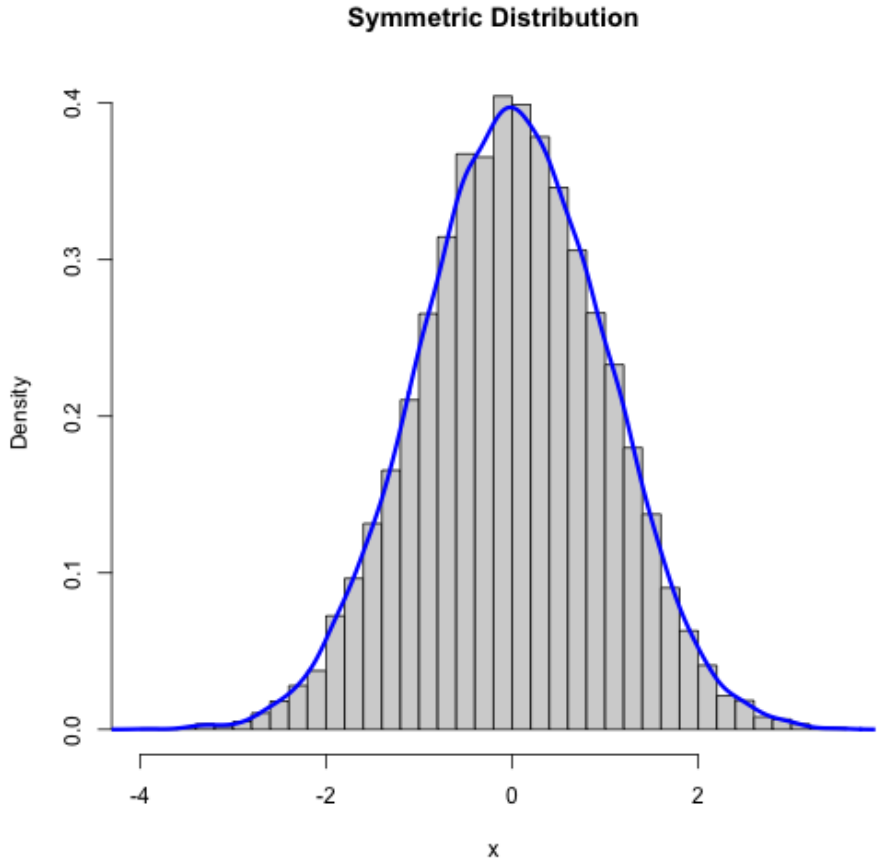
Symmetric distribution
Positive skewness (*aka*. positively skewed distribution)
Negative skewness (*aka*. negatively skewed distribution)
Unimodal distribution
Bimodal distribution
... and ~~Paranormal distribution~~

Distribution

Symmetric Distribution

```
set.seed(2093557522)
x = rnorm(10000,0,1)
hist(x, main="Symmetric Distribution", freq=FALSE,
     breaks=50)
lines(density(x), col='blue', lwd=3)
```

- Symmetrical around the **center**
- Any phenomenon that we would **normally** expect to observe
- Actually... this distribution has its name *normal distribution*

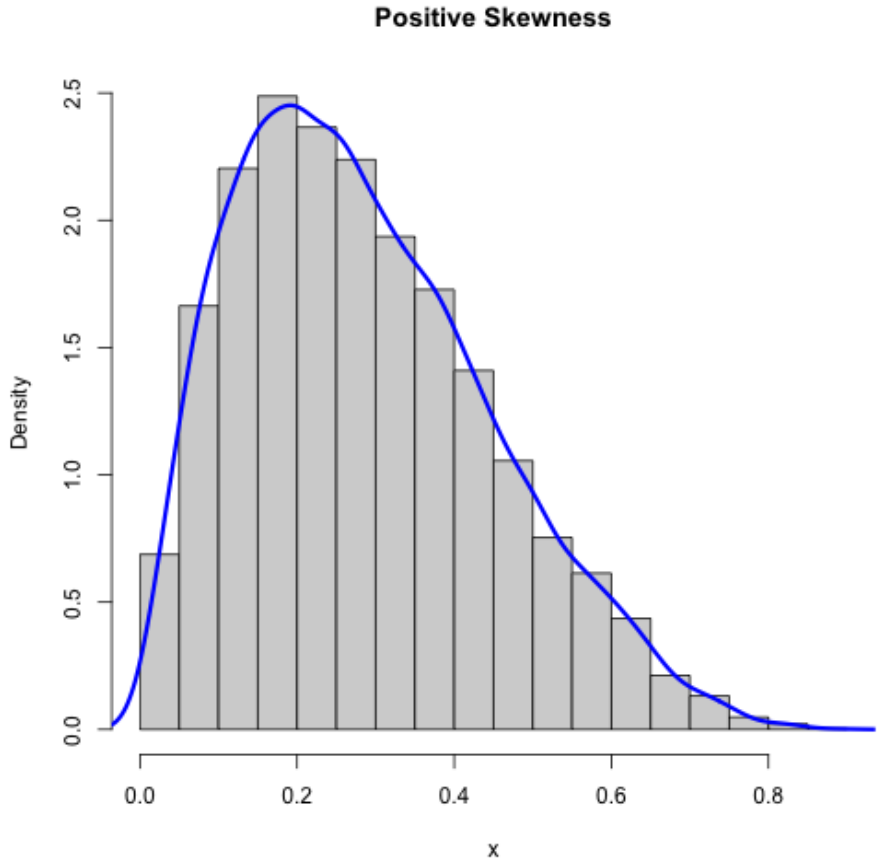


Distribution

Positive Skewness

```
set.seed(322)
x = rbeta(10000,2,5)
hist(x, main="Positive Skewness", freq=FALSE)
lines(density(x), col='blue', lwd=3)
```

- Trailing off toward the right end
- Score distribution of difficult exams

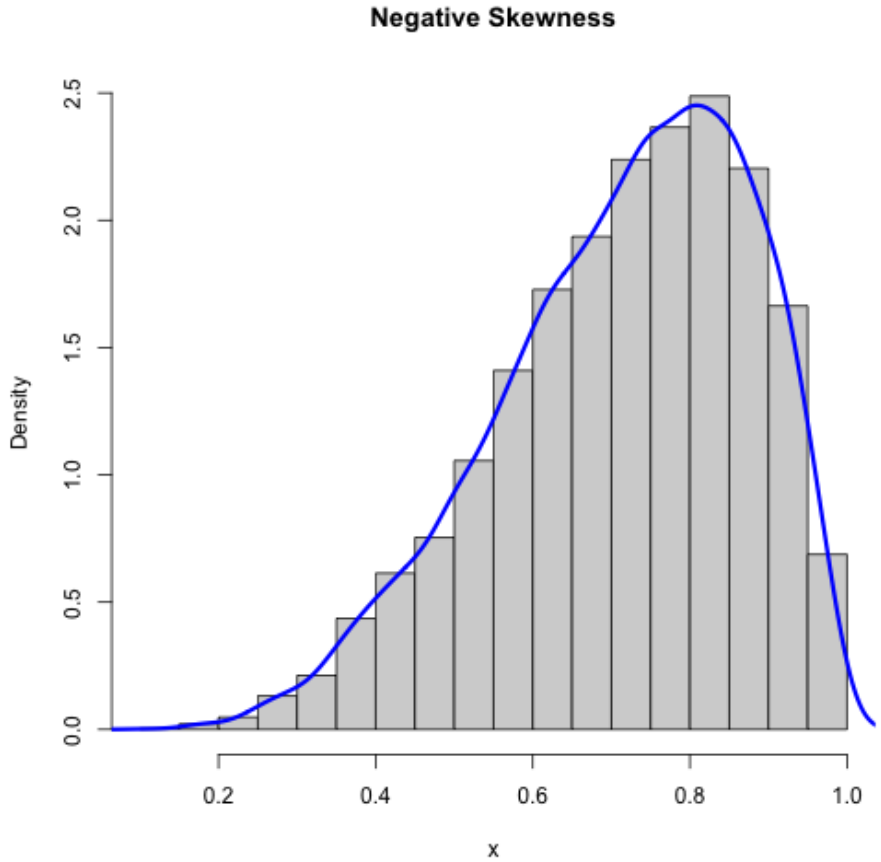


Distribution

Negative Skewness

```
set.seed(322)
x = rbeta(10000,5,2)
hist(x, main="Negative Skewness", freq=FALSE)
lines(density(x), col='blue', lwd=3)
```

- Trailing off toward the left end
- Score distribution of easy exams

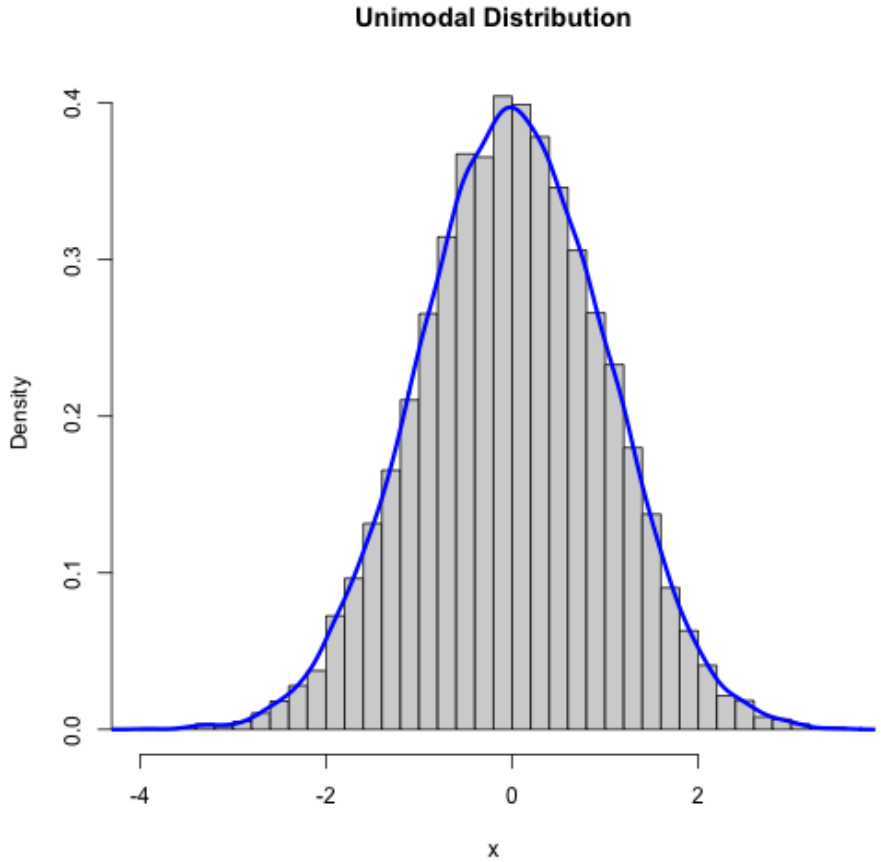


Distribution

Unimodal distribution

```
set.seed(2093557522)
x = rnorm(10000,0,1)
hist(x, main="Unimodal Distribution", freq=FALSE,
     breaks=50)
lines(density(x), col='blue', lwd=3)
```

- One single peak in the distribution

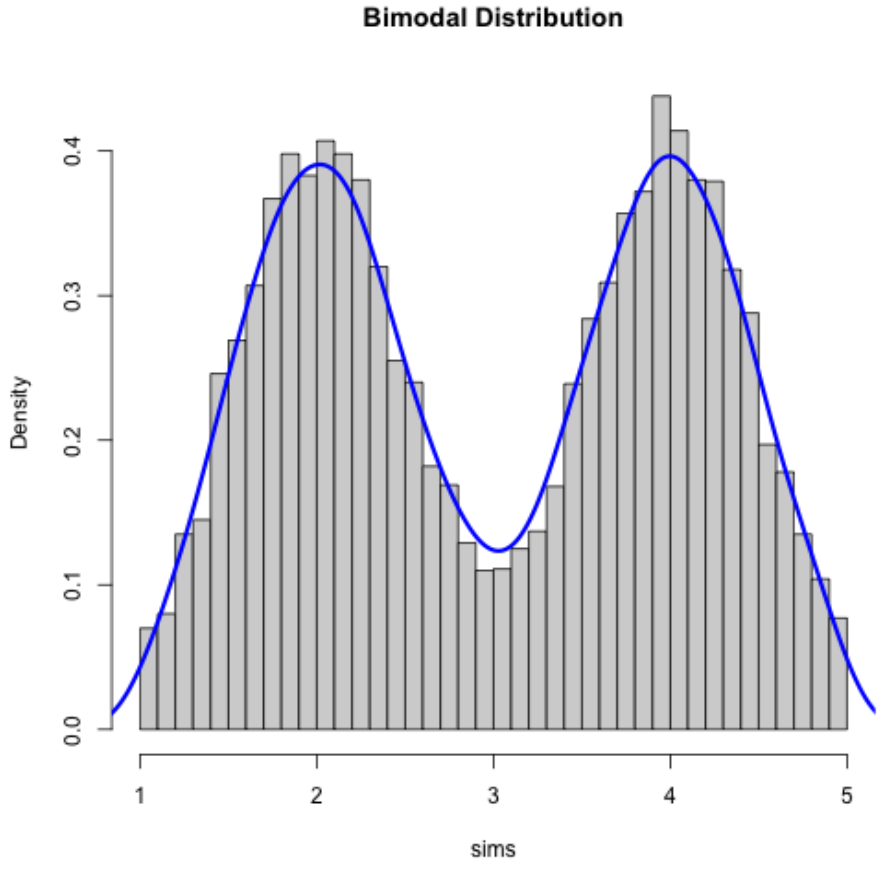


Distribution

Bimodal distribution

```
set.seed(20220201)
library(truncnorm)
nn <- 10000
sims <- c(rtruncnorm(nn/2, a=1, b=5, mean=2, sd=.5)
      rtruncnorm(nn/2, a=1, b=5, mean=4, sd=.5))
hist(sims, main="Bimodal Distribution", freq=FALSE,
     breaks=50)
lines(density(sims), col='blue', lwd=3)
```

- Two peaks in the distribution



Food for thought

Case 1

Melanie, a researcher at FDA, is interested in how people in California give ratings to the taste of ranch pizza. She collected 20 people in LA and made a frequency table on the pizza ratings (1 through 5).

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

Can you answer below?

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

Food for thought

Case 2

Itamar has been working as a data scientist at Marvel Studios. Since Spider-Man: No Way Home 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? Symmetrically distributed, negatively skewed, or positively skewed? Is it unimodal or bimodal?

Before you go home...

Lab materials are available at

https://github.com/lhnwhiHeo/PSY010

Any questions or comments?

Office hours or my email

Thanks! Have a wonderful day!