Introduction to Statistics

PSY 611: FALL 2023

Overview of the Course

- Prepare you for conducting analyses on real-world data
  - Preparing data for analysis, describing the data, analyzing the data in R using correct analyses, checking assumptions, interpreting/reporting results

- Two overarching purposes of statistics:
  - Descriptive Statistics: describing data
  - Inferential Statistics: drawing inferences about populations

PSY 611 / 612 / 613

- PSY 611
  - Linear regression analyses using *categorical* predictors
- PSY 612
  - Linear regression analyses using *continuous* predictors
- PSY 613
  - Advanced analyses like multilevel modeling, components analysis, & structural equation modeling

PSY 611 Plan

Week 1:	Preparing & Describing Data	**Week 6:**	Linear Regression with a Single Categorical Predictor with 2 Related Levels (aka, Paired Samples t-Test)
Week 2:	The Logic of Hypothesis Testing	**Week 7:**	Linear Regression with a Single Categorical Predictor with 2+ Related Levels (aka, One-Way Repeated-Measures ANOVA)
Week 3:	The General Linear Model Approach	**Week 8:**	Linear Regression with Multiple Categorical Predictors (aka, Factorial ANOVA)
Week 4:	Linear Regression with a Single Categorical Predictor with 2 Independent Levels (aka, Independent Samples t-Test)	**Week 9:**	Handling Missing Data, Outliers, & Unmet Assumptions
Week 5:	Linear Regression with a Single Categorical Predictor with 2+ Independent Levels (aka, One-Way Between-Subjects ANOVA)	**Week 10:**	Pre-registration & Open Science

The Steps of Data Analysis

1. Pre-registration (https://osf.io/)
   - Specify your methods & analysis plans *prior to analyzing (or even looking at) your data*
2. Data Cleaning / Data Wrangling
   - Get the empirical data into a form that is ready for analysis
3. Descriptive Statistics
   - Describe the data using numerical & visual summaries
4. Inferential Statistics
   - Choose the appropriate analysis to conduct and fit the chosen model to the data
   - Check whether the test assumptions were met
5. Interpret & Report the Results
   - Parameter estimates, effect sizes, *p*-values, confidence intervals, simple slopes analyses, planned & post-hoc comparisons, etc.
6. Replication

The General Linear Model

- The **General Linear Model** is a general framework for analyzing the effects of **continuous** and **categorical predictor variables** on a **continuous outcome variable** using a multiple linear regression model (or simple linear regression in the case of only one predictor).

  - from Cohen (1968): Multiple Regression as a General Data-Analytic System

> If you should say to a mathematical statistician that you have discovered that linear multiple regression analysis and the analysis of variance (and covariance) are identical systems, he would mutter something like, "Of course—general linear model," and you might have trouble maintaining his attention. If you should say this to a typical psychologist, you would be met with incredulity, or worse. Yet it is true, and in its truth lie possibilities for more relevant and therefore more powerful exploitation of research data.

The General Linear Model

- In other words, regression and ANOVA (and thus, t-tests) are both special cases of the general linear model
  - Regardless of whether one's predictor variables are continuous or categorical, a regression analysis can be used to analyze the predictors' relationships with an outcome variable

- Why are regression and ANOVA taught as independent analysis techniques, then?
  - A brief history of statistics

A Brief History of Statistics

- *1880s-1890s*: Regression developed in biology and psychology to analyze correlation between observed characteristics of people (e.g., height of parents and their adult children)
- *1908*: W. S. Gosset develops the *t*-test for his job brewing beer at Guinness in Dublin, Ireland. Brewing the perfect beer required adding the correct amount of yeast (too little and the fermentation would be incomplete, too much and the beer would taste bitter). Gosset was tasked with counting the number of yeast colonies in entire jars of beer. He developed the *t*-test to infer how much yeast was in an entire jar (population) based on only a small sample taken from the jar (sample).
- *1920s-1930s*: Ronald A. Fisher developed the analysis of variance (ANOVA) technique for his job analyzing agricultural experimental data (e.g., how different crop conditions affect agricultural output)

A Brief History of Statistics

- *1940s*: Logistic regression developed as a method for analyzing categorical outcome variables
- *1970s*: The term general*ized* linear models was coined to refer to statistical analysis techniques for analyzing both continuous and categorical outcome variables
- Prior to the 1980s, limits in computing technology made complicated statistical analyses computationally prohibitive because researchers had to perform them by hand.
- *1980s*: Advances in computing power allowed researchers to develop techniques for analyzing *non-linear* relationships
- Today we have many statistical software options to choose from: R, Python, SPSS, SAS, Jamovi

GLM and ANOVA

- The GLM approach focuses on fitting regression models
- The ANOVA approach focuses on analyzing mean differences

- Both analytic techniques will result in the same conclusion
  - Strengths of GLM: more parsimonious across various types of analyses
  - Strengths of ANOVA: can be a helpful way of presenting the results of an analysis

- In this course, I will teach mainly from the GLM perspective, but I will also incorporate ANOVA concepts and techniques when there are strengths in adding it in.

Variables

Measurement Scales

Variables are phenomena that can take on more than one value.

- The type of scale on which a variable is measured has important implications for the types of analyses that can be conducted with that variable.

Nominal (aka, Categorical): Measuring a variable using categories with no intrinsic ordering

- Ex: Which political party do you belong to?: Independents, Republicans, Democrats

Ordinal: Measuring a variable using categories with an intrinsic ordering, but the distance between adjacent values is not necessarily equal

- Ex: Rank the following political parties in order of how strongly you identify with them: Independents, Republicans, Democrats

Measurement Scales

Interval: Measuring a variable using a numerical scale with equal distance between adjacent values

- Example: How strongly do you identify as being a Republic on a scale from 1 (*not at all*) to 5 (*extremely identify*)?

1	2	3	4	5
Not at all	Slightly Identify	Moderately Identify	Very Much Identify	Extremely Identify

Digression: Are Likert scales really measured on interval scales?

- Many statistical sources conclude that Likert scales are actually measured on **ordinal scales**
- For psychologists, this is very inconvenient because many analyses, like calculating an average, cannot be performed on ordinal variables
- Instead, many argue that participants are psychologically applying approximately the same distance between adjacent values on a Likert scale
  - (It's helpful to add numbers to go along with each scale option).

Measurement Scales

Ratio: Measuring a variable using a numerical scale with equal distance between values *and* a zero indicates the absence of the phenomenon

- Example: How many friends do you have that identify with being part of the Democratic party?

Example of descriptive statistics that can be conducted with each measurement scale:

Measurement Scale	
Nominal	Frequency, mode
Ordinal	Frequency, mode, median
Interval	Frequency, mode, median, mean
Ratio	Frequency, mode, median, mean

Describing Data

Thought Exercise

- Say I asked 24 former University of Oregon graduate students how satisfied they are with their lives one year after graduating on a scale from 1 *(not at all satisfied)* to 10 *(extremely satisfied)*. The collected scores are shown in the table to the right.
  - Come up with **five** different ways of concisely describing this data.

Life Satisfaction		
9	6	1
3	9	7
10	7	8
4	4	6
6	2	7
4	7	6
8	6	8
2	7	3

Frequency Table

- A frequency table displays the number of times that each value on a variable occurred.

```
> count(life_satisfaction) # plyr
   x freq
1  1    1
2  2    2
3  3    2
4  4    3
5  6    5
6  7    5
7  8    3
8  9    2
9 10    1
```

```
> tab1(life_satisfaction) # epiDisplay
life_satisfaction :
         Frequency Percent Cum. percent
1                1     4.2          4.2
2                2     8.3         12.5
3                2     8.3         20.8
4                3    12.5         33.3
6                5    20.8         54.2
7                5    20.8         75.0
8                3    12.5         87.5
9                2     8.3         95.8
10               1     4.2        100.0
  Total         24   100.0        100.0
```

Measures of Central Tendency

Measures of central tendency attempt to capture the **typical** response given on a variable.

- **Mean**
  - Most commonly used measure of central tendency
  - The average of a set of values on a variable
  - Takes all values equally into consideration

The mean is a *balancing point*. The total distance between the scores above the mean and the mean itself equals the total distance between the scores below the mean and the mean itself.

Sample Mean:

$$M = \frac{\Sigma X}{n}$$

Population Mean:

$$\mu = \frac{\Sigma X}{N}$$

```
> mean(life_satisfaction)
[1] 5.833333
```

Measures of Central Tendency

- **Median**
  - The midpoint of a set of values when they are listed from lowest to highest.
  - The location of the median can be found using: $(n + 1) / 2$

```
> median(life_satisfaction)
[1] 6
```

- **Mode**
  - The most frequently occurring score.
  - There can be more than one mode.

```
> mfv(life_satisfaction) # statip
[1] 6 7
```

Measures of Central Tendency

	Strengths	Weaknesses
Mean	Takes all values into consideration; widely used	The *most* affected by extreme scores (i.e., outliers).
Median	Not as affected by extreme scores (i.e., outliers) as the mean.	Not all values contribute *equally* (i.e., you could change some of the values without the median being affected).
Mode	The only measure of central tendency that can be used to describe categorical data.	Only descriptive of a single value in the data set.

Measures of Variability

The typical score only captures a small aspect of describing an entire data set. Variability describes how spread out the values are. Did participants score very **similarly** or very **differently** from one another?

- Minimum & Maximum

```
> min(life_satisfaction)
[1] 1
> max(life_satisfaction)
[1] 10
```

- Range
  - The distance between the maximum and minimum values:

$$R = x_{\max} - x_{\min}$$

```
> diff(range(life_satisfaction))
[1] 9
```

Measures of Variability

Interquartile Range: A measure of the range for the middle 50% of the data.

Steps for calculating the IQR:

1. Arrange scores from smallest to largest
2. Find the median. This is the second quartile (Q2).
3. To fine Q1, find median for the scores below Q2.
4. To find Q3, find the median for the scores above Q2.
5. IQR = Q3 – Q1

```
> quantile(life_satisfaction, probs = seq(0,1,0.25), type = 2)
  0%  25%  50%  75% 100%
 1.0  4.0  6.0  7.5 10.0
> IQR(life_satisfaction, type = 2)
[1] 3.5
```

Measures of Variability

- Standard deviation and variance are such commonly used, and important to understand, measures of variability that it is worth calculating them by hand to see exactly what they are computing.
  - Let's take a subset of five participants from our Life Satisfaction variable to do this calculation (shown to the right)

Life Satisfaction
9
3
8
4
6

- The standard deviation is commonly thought of as the average amount by which the values in a data set tend to deviate from the mean of the set of values
  - **Go ahead and calculate the standard deviation from scratch from this set of 5 values based on this definition**

Calculating Variance and Standard Deviation

Life Satisfaction (X)	(X – Mean)	$(X - Mean)^2$
9	3	9
3	-3	9
8	2	4
4	-2	4
6	0	0
$Mean = 6$		$SS = \Sigma(X - Mean)^2 = 26$

- **Variance:** 26/5 = 5.20
- **Standard deviation:** sqrt(Variance) = sqrt(5.20) = 2.28

Measures of Variability

- **Variance**: average squared deviation of scores from the mean
- **Standard Deviation**: the square root of the variance

To add even more complexity, the calculation of the variance and standard deviation differs depending on whether you have **sample data** or **population data**.

- Up to calculating the *SS*, the steps are the same. But then →

	Population	Sample
Variance	$\sigma^2 = \frac{SS}{N}$	$s^2 = \frac{SS}{n-1}$
Standard Deviation	$\sigma = \sqrt{\frac{SS}{N}}$	$s = \sqrt{\frac{SS}{n-1}}$

When we have sample data, why do we divide by $n - 1$ instead of N?

Calculating Variance and Standard Deviation

Life Satisfaction (X)	(X – Mean)	(X – Mean)^2
9	3	9
3	-3	9
8	2	4
4	-2	4
6	0	0
Mean = 6		SS= 26

Let's recalculate the variance and standard deviation for our life satisfaction values treating them as if they were **sample data**.

- Variance: 26/(5-1) = 6.50
- Standard deviation: sqrt(6.50) = 2.55

Which version do the *R* functions var() and sd() calculate?

```
> var(life_satisfaction)
[1] 6.5
> sd(life_satisfaction)
[1] 2.54951
```

Measures of Variability

- Mean Absolute Deviation

Life Satisfaction (*X*)	(X – Mean)	|(X – Mean)|
9	3	3
3	-3	3
8	2	2
4	-2	2
6	0	0
Mean = 6		Sum = 10

- MAD = 10/5 = 2.00

```
> madstat(life_satisfaction) # ie2misc
[1] 2
```

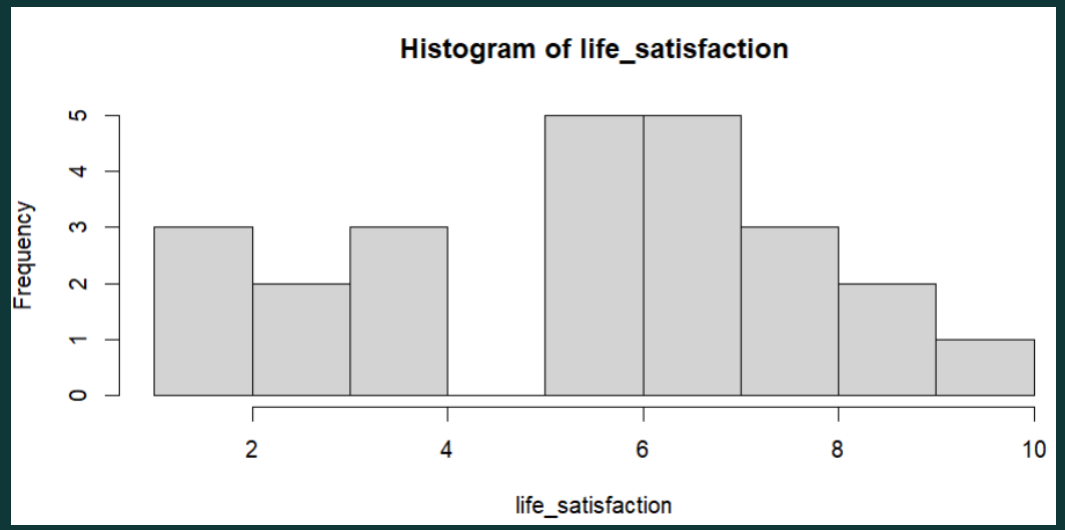
Review of Notation

	Sample Data	Population Data
Sample or Population Size	n	N
Mean	M	μ
Variance	s^2	σ^2
Standard Deviation	s, SD	σ

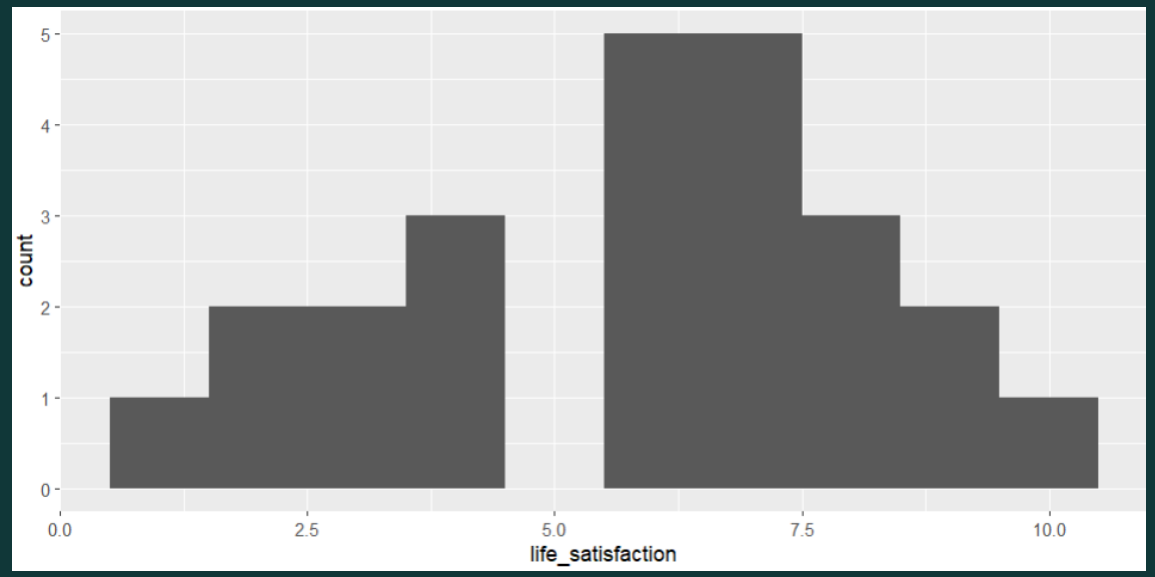
Graphs

- Histogram: A graphical representation of the frequency of each value on a continuous variable

```
> hist(life_satisfaction, breaks = 10)
```



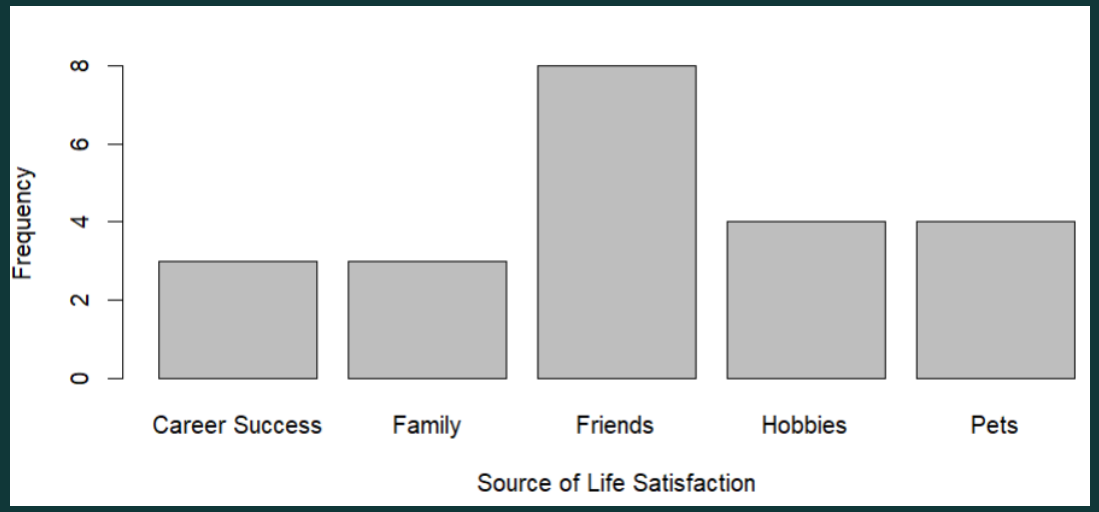

```
> ggplot(df_life_sat, aes(x = life_satisfaction)) +
+   geom_histogram(binwidth = 1)
```



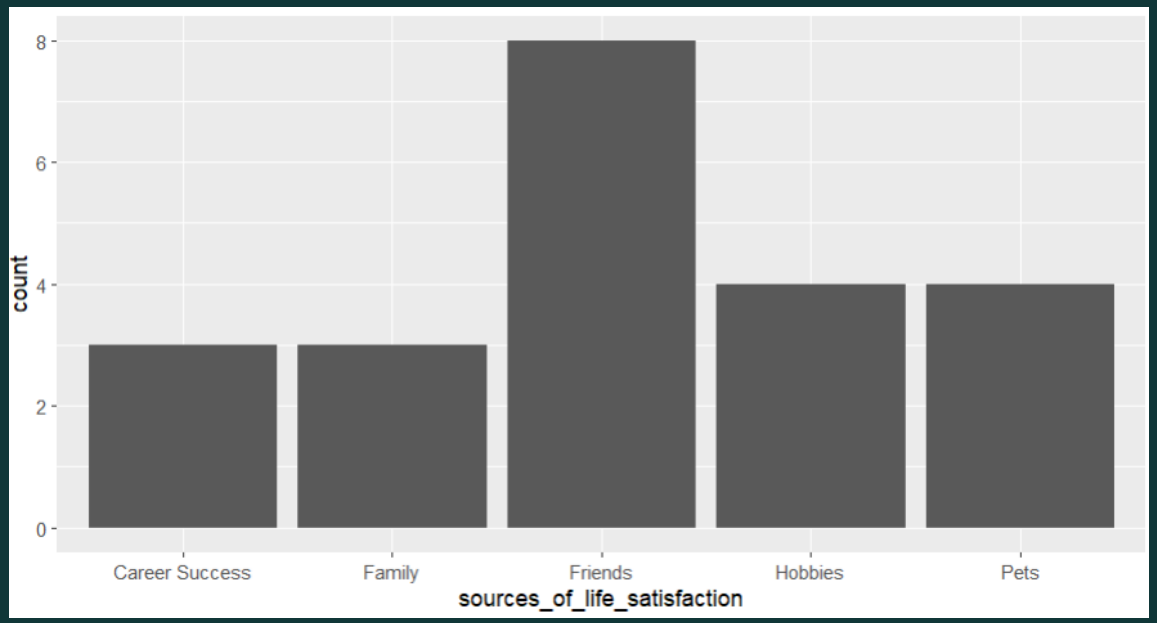
Graphs

- Bar graph: A graphical representation of the frequency of each value on a categorical variable

```
barplot(count(sources_of_life_satisfaction)$freq, names.arg =
c("Career Success", "Family", "Friends", "Hobbies", "Pets"),
xlab = "Source of Life Satisfaction", ylab = "Frequency")
```




```
ggplot(df_sources_of_life_sat) +
  geom_bar(aes(x = sources_of_life_satisfaction))
```



Graphs

```
> quantile(life_satisfaction, probs = seq(0,1,0.25), type = 2)
  0%  25%  50%  75% 100%
 1.0  4.0  6.0  7.5 10.0
> IQR(life_satisfaction, type = 2)
[1] 3.5
```

- Box plot: A graphical representation of the interquartile range

```
boxplot(life_satisfaction, xlab = "Life Satisfaction")
```

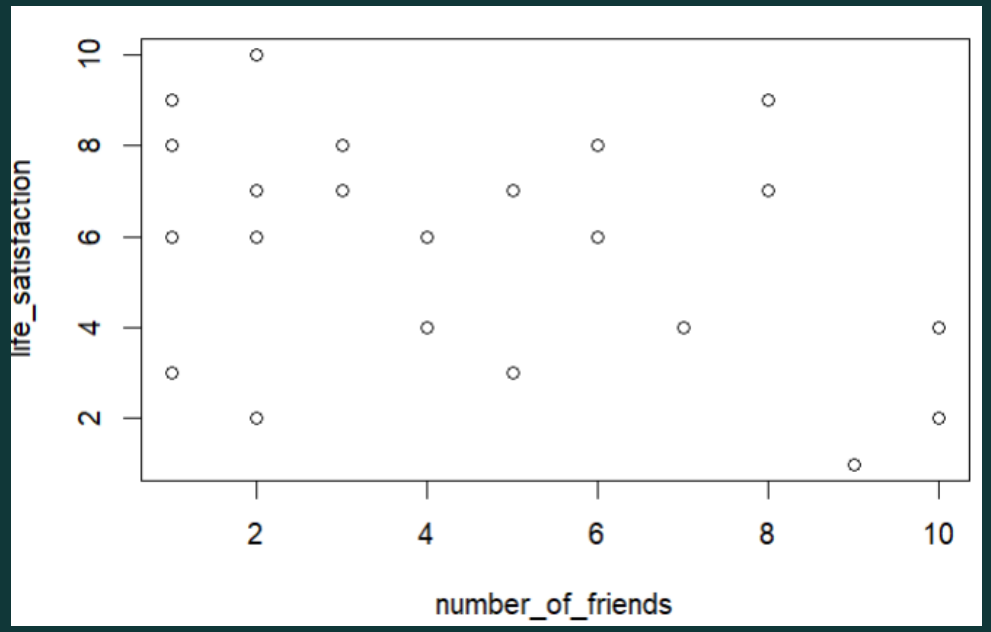
- The bottom edge of the box corresponds to Q1
- The top edge of the box corresponds to Q3
- The box represents the IQR

- Boxplots can be used to identify potential outliers
- The **whiskers** extend to the most extreme value in the data that is no more than 1.5*IQR from either edge of the box (by default)
  - If you would like to modify the 1.5, change the “range” argument in boxplot( )

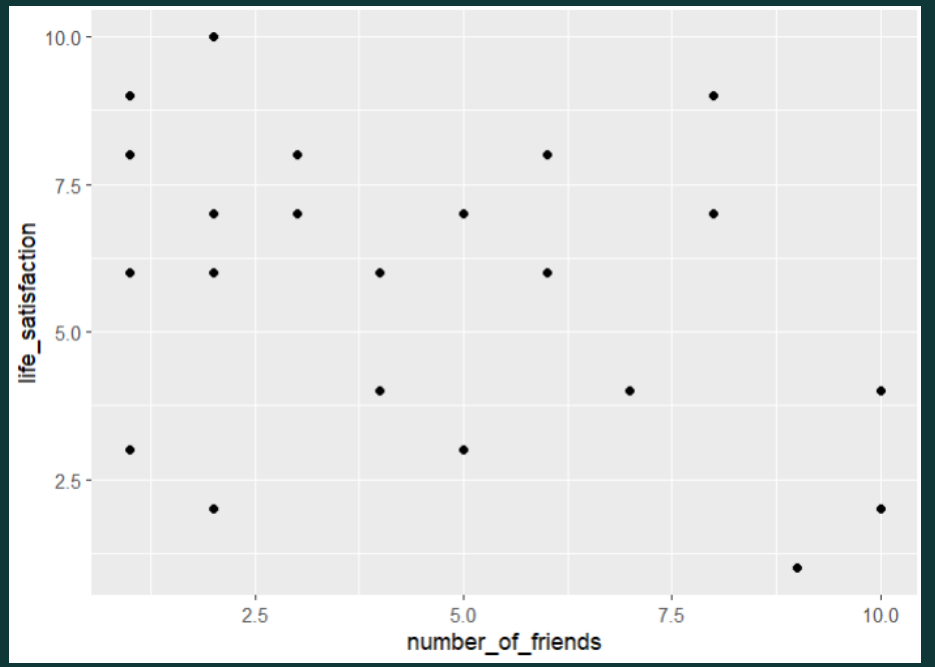
Graphs

- Scatterplots: A graphical representation of the relationship between two continuous variables

```
plot(number_of_friends, life_satisfaction)
```

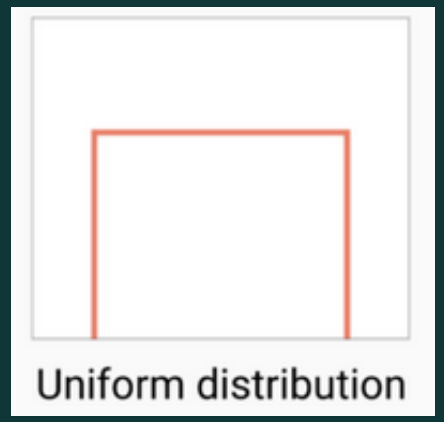


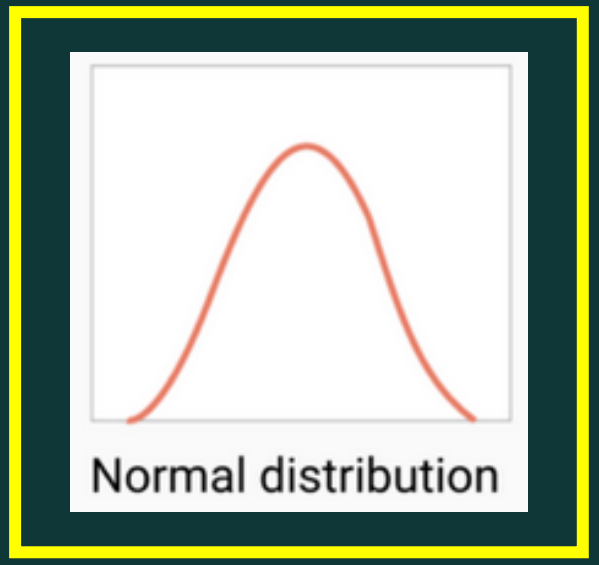

```
ggplot(df_life) +
  geom_point(aes(x = number_of_friends, y = life_satisfaction))
```

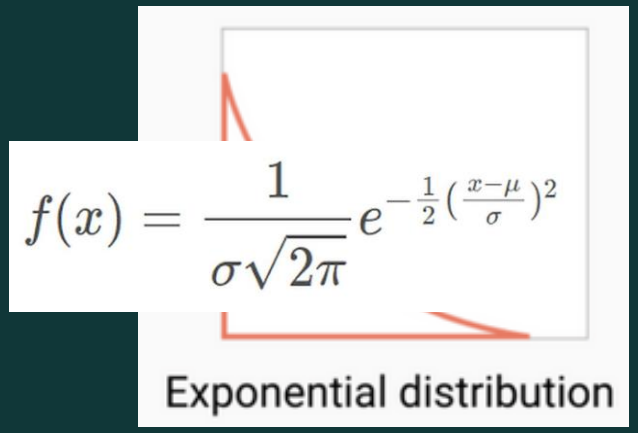


Shape

- Type of distribution



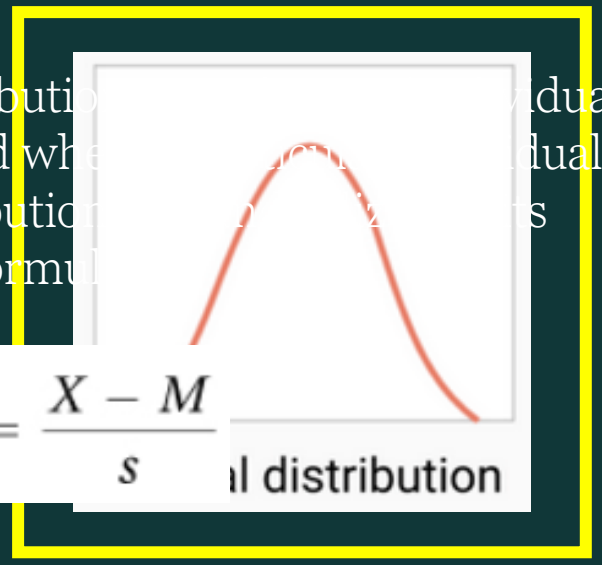




Shape

- Type of distribution

- If the normal distributi[illegible]viduals' scores, you can find wh[illegible]dual lands on the distributio[illegible]s using the *z*-score formu[illegible]

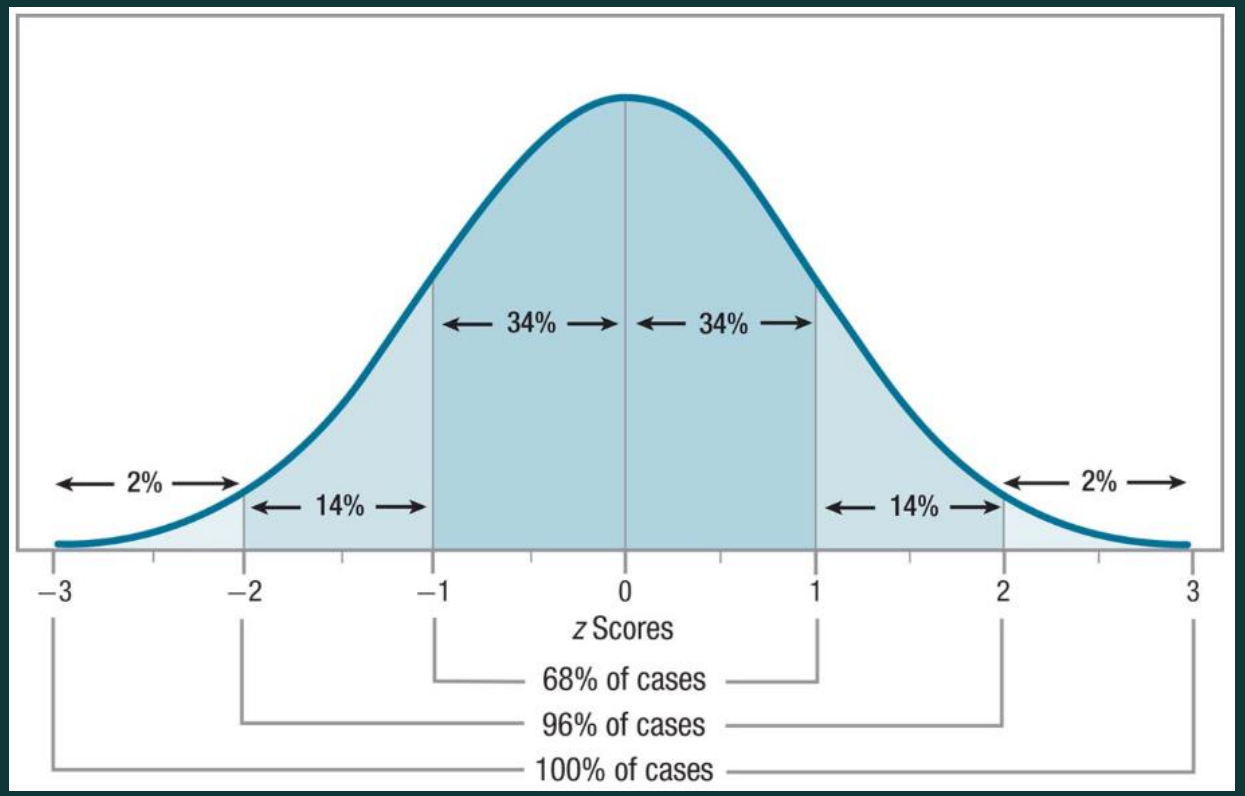


Sample Data:

$$z = \frac{X - M}{s}$$

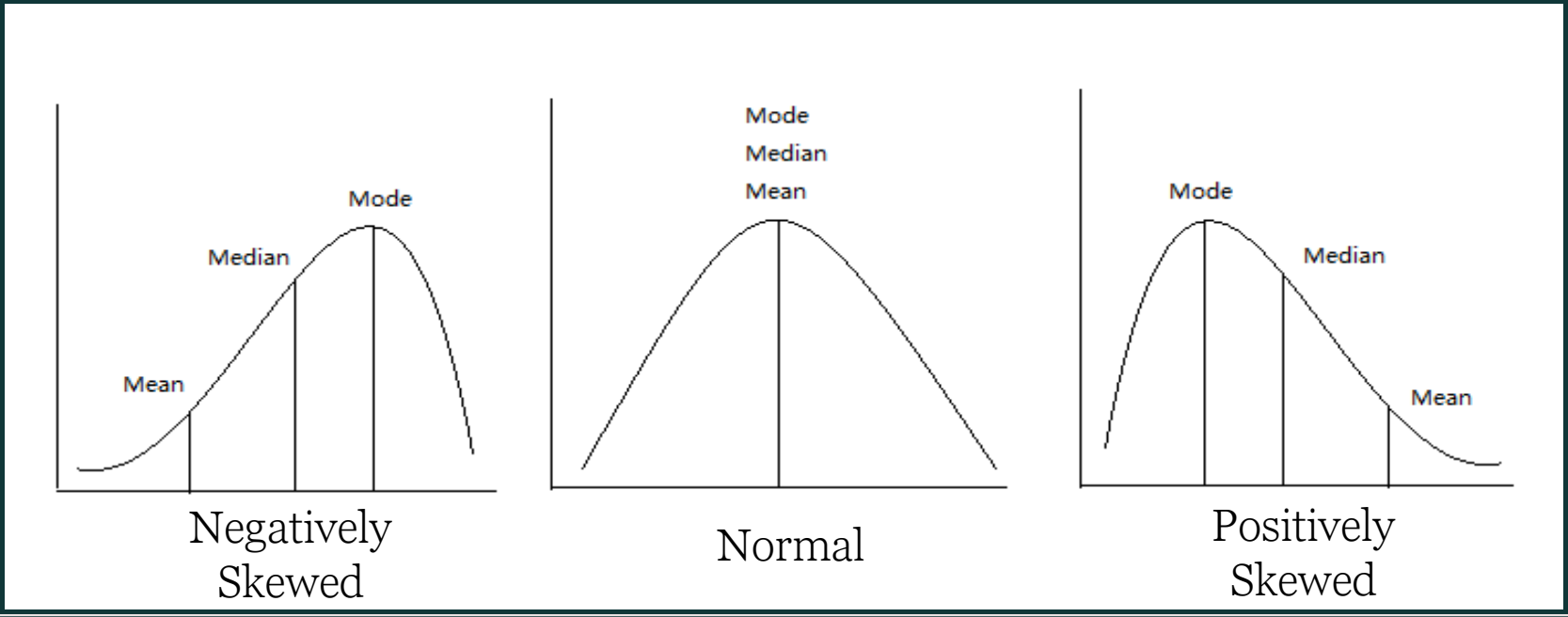
Population Data:

$$z = \frac{X - \mu}{\sigma}$$



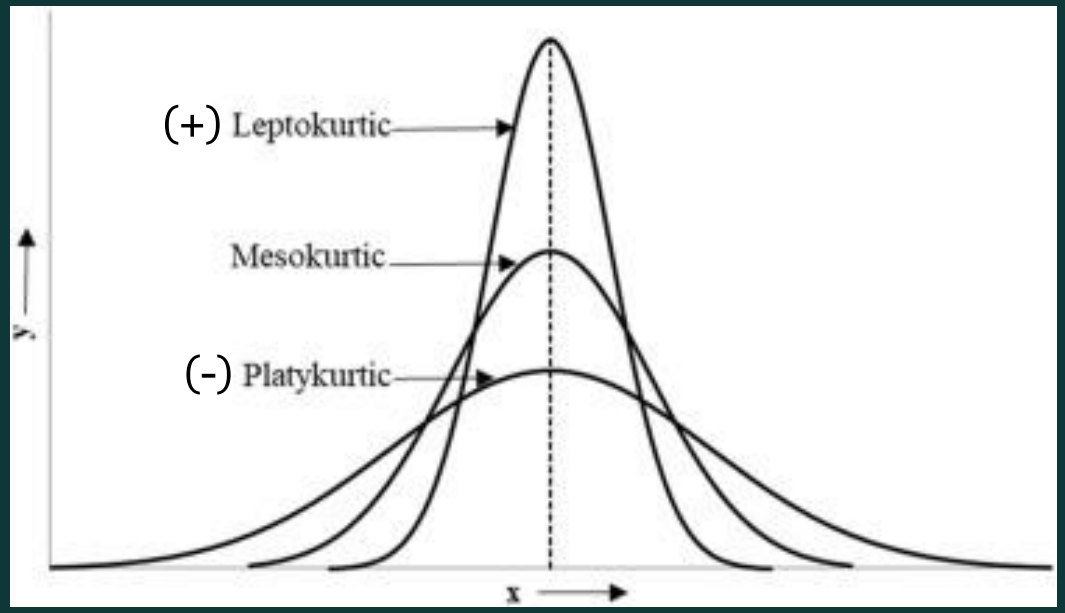
Shape

- Skewness: the normal distribution has a skewness of zero (is symmetrical)
  - The sign indicates the direction of skewness and the value indicates the severity of the skewness

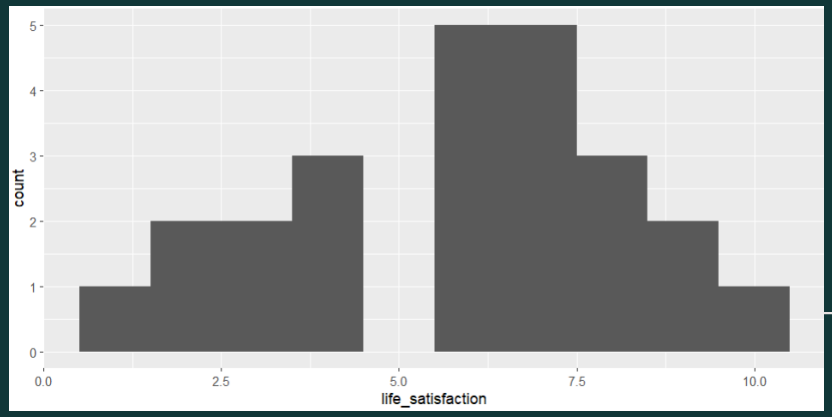


Shape

- Kurtosis: the normal distribution has a kurtosis of 0
  - Platykurtic (Negative Kurtosis): Flatter than a normal distribution
  - Leptokurtic (Positive Kurtosis): More peaked than a normal distribution




```
> describe(life_satisfaction)
   vars  n mean   sd median trimmed  mad min max range  skew kurtosis  se
X1    1 24 5.83 2.44      6     5.9 2.97   1  10     9 -0.33    -0.99 0.5
```



Cleaning Data

Cleaning Data

- Cleaning data is an exceptionally important step to complete prior to conducting your analyses

Data Cleaning Considerations

- Check variable measure types
- Clean variable names
- Remove duplicate rows / IDs / IP addresses
- Identify and fix (if possible) or remove data entry errors
- Get data in tidy format
- More to come on: handling missing data & outliers later in the course