What Statistical Analysis Should I Run?: Statistical Analysis Workflow and Statistical Tests

CMU-Q Statistical Consulting Center Workshop Series

Taeyong Park
Carnegie Mellon University in Qatar

CMU-Q Statistical Consulting Center

Taeyong Park

Co-Director. Assistant Teaching Professor of Statistics.

Daniel C. Phelps

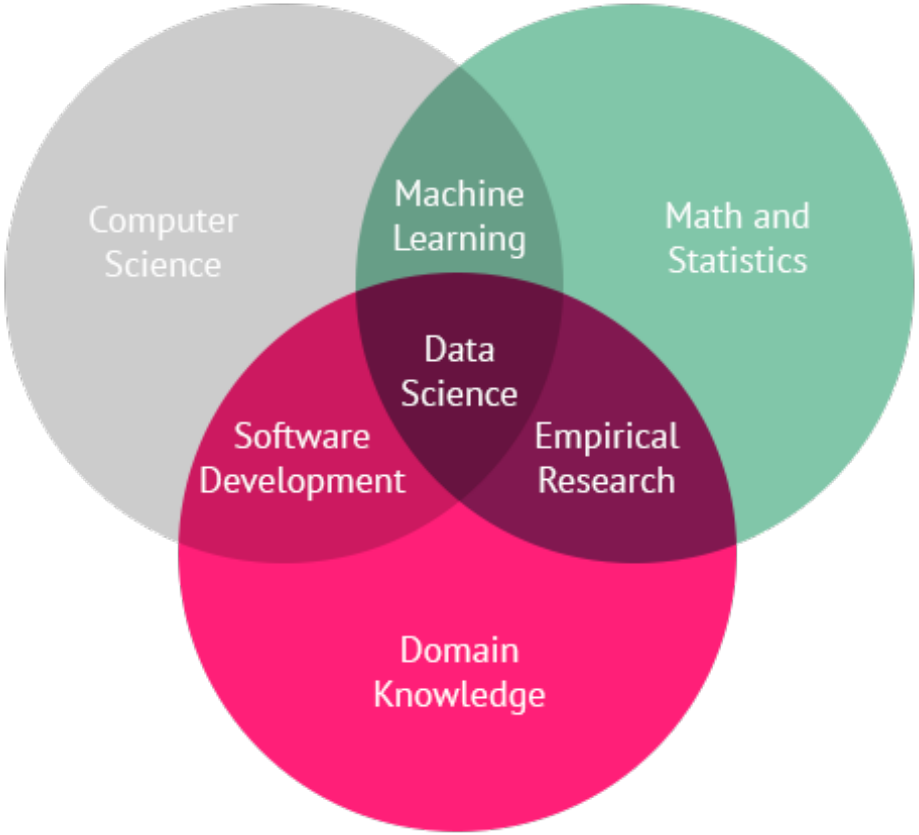
Co-Director. Associate Teaching Professor of Information Systems.

- Statistical advice for research design, modeling, measurements, and data analysis and visualization, etc.
  - You can schedule an appointment here: https://www.qatar.cmu.edu/statistical-consulting-center/
  - Google "statistical consulting center cmu."
- Open to faculty, staff, and students from Education City universities and Qatar Foundation.

Overview of today's workshop

1. Big picture: Descriptive statistics and Inferential statistics.
2. Descriptive statistics: Numerical and graphical summaries.
3. Inferential statistics: Hypothesis tests.
4. Hands-on practice: Use R to run several statistical tests.

What is statistics?



- Statistics is a component of data science and consists of a body of **methods for obtaining and analyzing data**.

Methods for analyzing data

- Methods for analyzing data: derive patterns, insights, or conclusions to allow organizations and companies to make better decisions as well as verify and disprove existing theories or models.

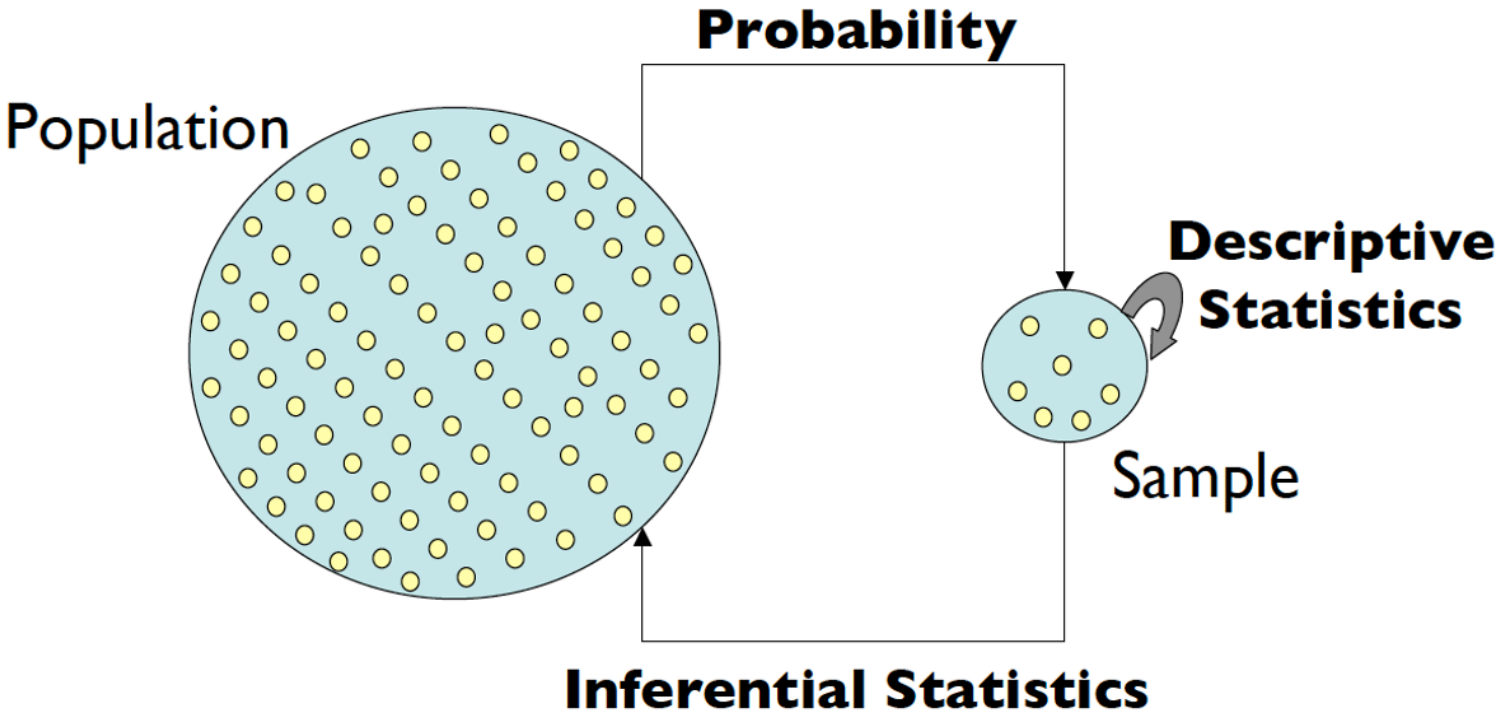
Analyzing data

Two methods for analyzing data.

- Descriptive statistics (a.k.a. exploratory data analysis): discover patterns and extract insights with the help of summary statistics and graphical representations.

- Inferential statistics: make an estimation/prediction about population data using sample data.

Descriptive statistics and Inferential statistics



From Joshua Akey, https://www.gs.washington.edu/academics/courses/akey/56008/lecture.htm.

Descriptive statistics

- **Descriptive statistics** or **Exploratory data analysis (EDA)**: Extracts meaningful patterns by summarizing the main characteristics of the data. Generally, the first step for analyzing data before inferential statistics.
  - Non-graphical methods: Numbers, such as averages and percentages, and frequency tables.
  - Graphical methods: Bar charts, box plots, histograms, etc.

- How to summarize? Depends on the data types.

Data types

	Quantitative		Qualitative
	Continuous	*Discrete*	*Discrete*
Ordinal	Height	Family size, Trust in government (1,2,3)	Trust in government (trust, neutral, not trust)
Nominal			College major

- Quantitative: Numerical data.
- Qualitative: Non-numerical data, such as characters.

- Continuous: Given in an infinite continuum of values. An infinite number of data points between two data points.
- Discrete (a.k.a. categorical): A finite number of data points between two data points.

- Ordinal: Ordering. No measurable differences between data points.
- Nominal: No natural ordering.

Data types

	Quantitative		Qualitative
	Continuous	*Discrete*	*Discrete*
Ordinal			
Nominal			

- Exercise
  - cell size (in mm)
  - cell size level (1-5)
  - favorite colors
  - commuting time (from home to work)
  - letter grade A, B, C, D

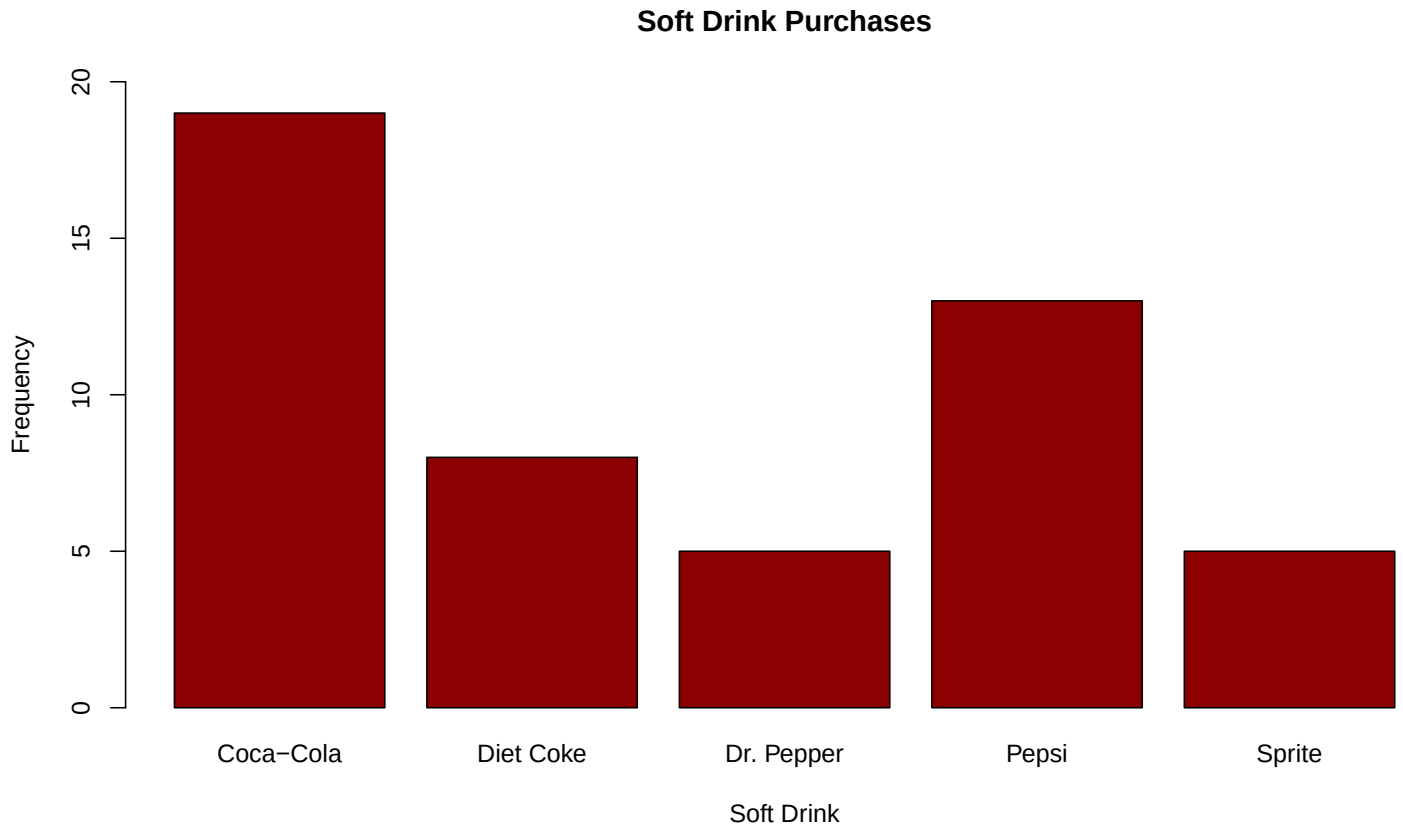
Data types

	Quantitative		Qualitative
	Continuous	*Discrete*	*Discrete*
Ordinal	commuting time, cell size	N. of siblings, cell size level	Letter grade
Nominal			favorite colors

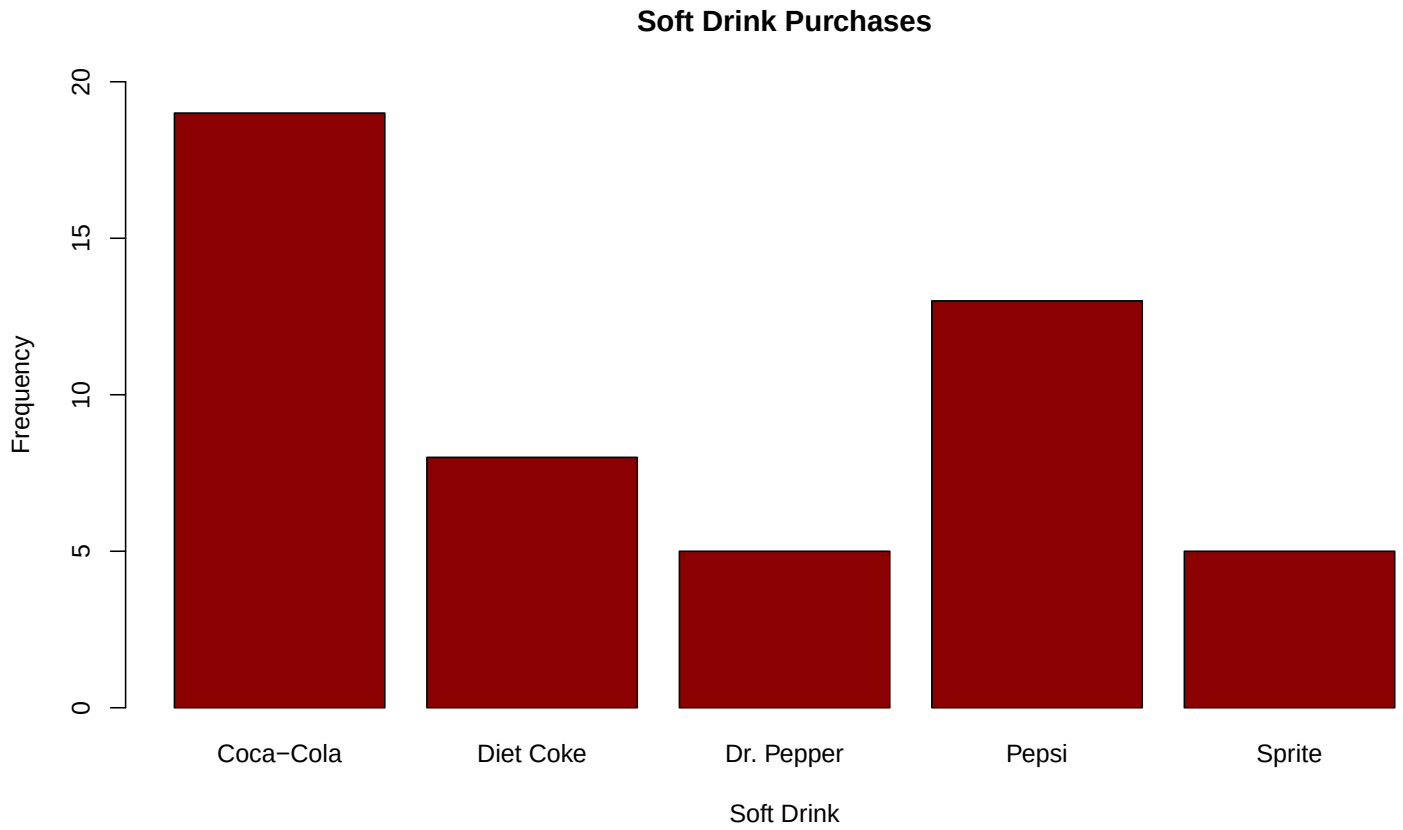
Summarizing qualitative-nominal data: Frequency distribution

Soft Drink	Frequency	Proportion	Percentage
Coca-Cola	19	.38	38
Diet Coke	8	.16	16
Dr. Pepper	5	.10	10
Pepsi	13	.26	26
Sprite	5	.10	10
Total	50	1.00	100

Summarizing qualitative-nominal data: Bar plot



Summarizing qualitative-nominal data: Bar plot



- A bar plot also works for discrete-ordinal data, but the data on the X-axis must be presented in order.

Quantitative data

Data points differ in magnitude; Numerical values.

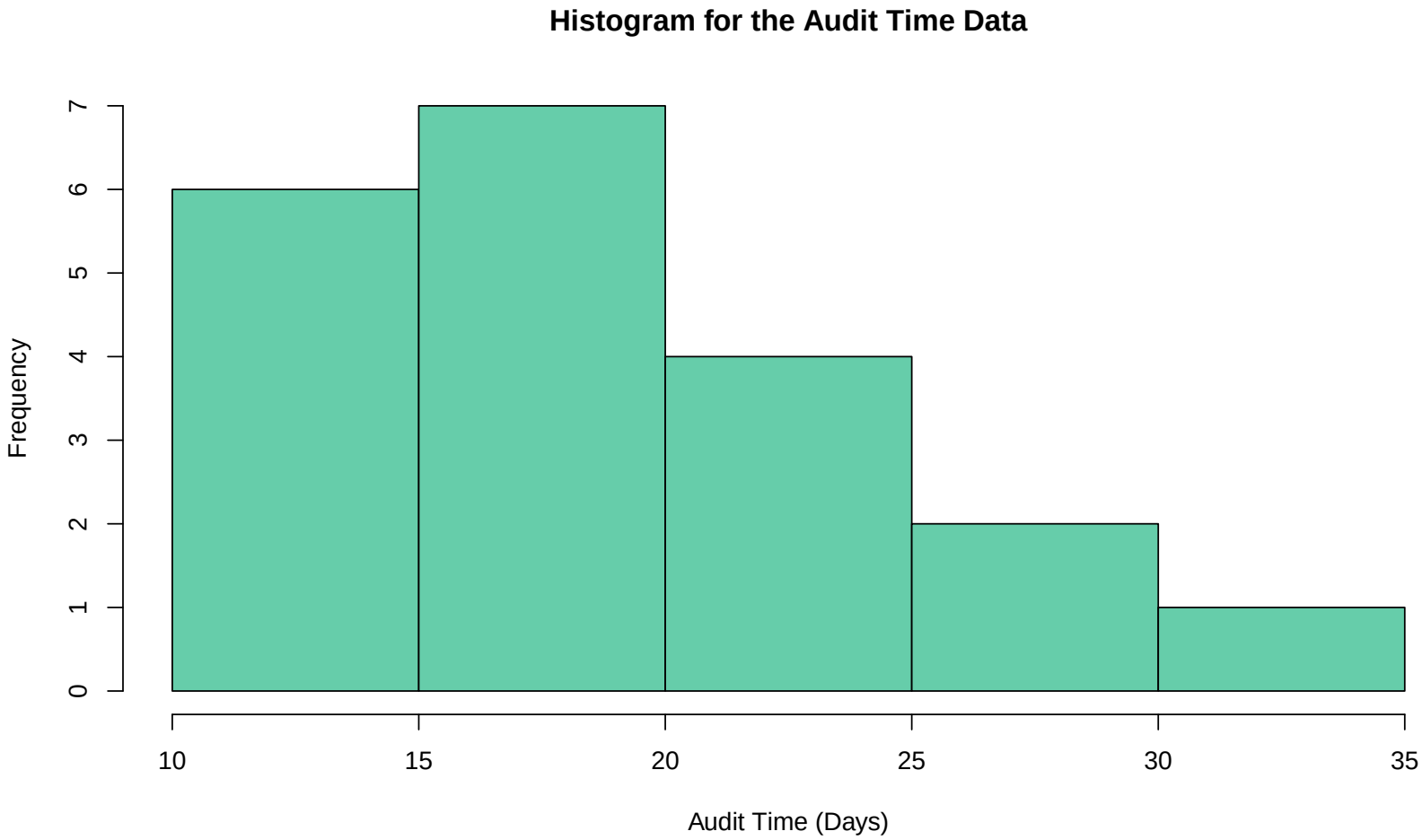
- Example: Year-End Audit Times (in days) for a sample of 20 clients of a public accounting firm.

12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

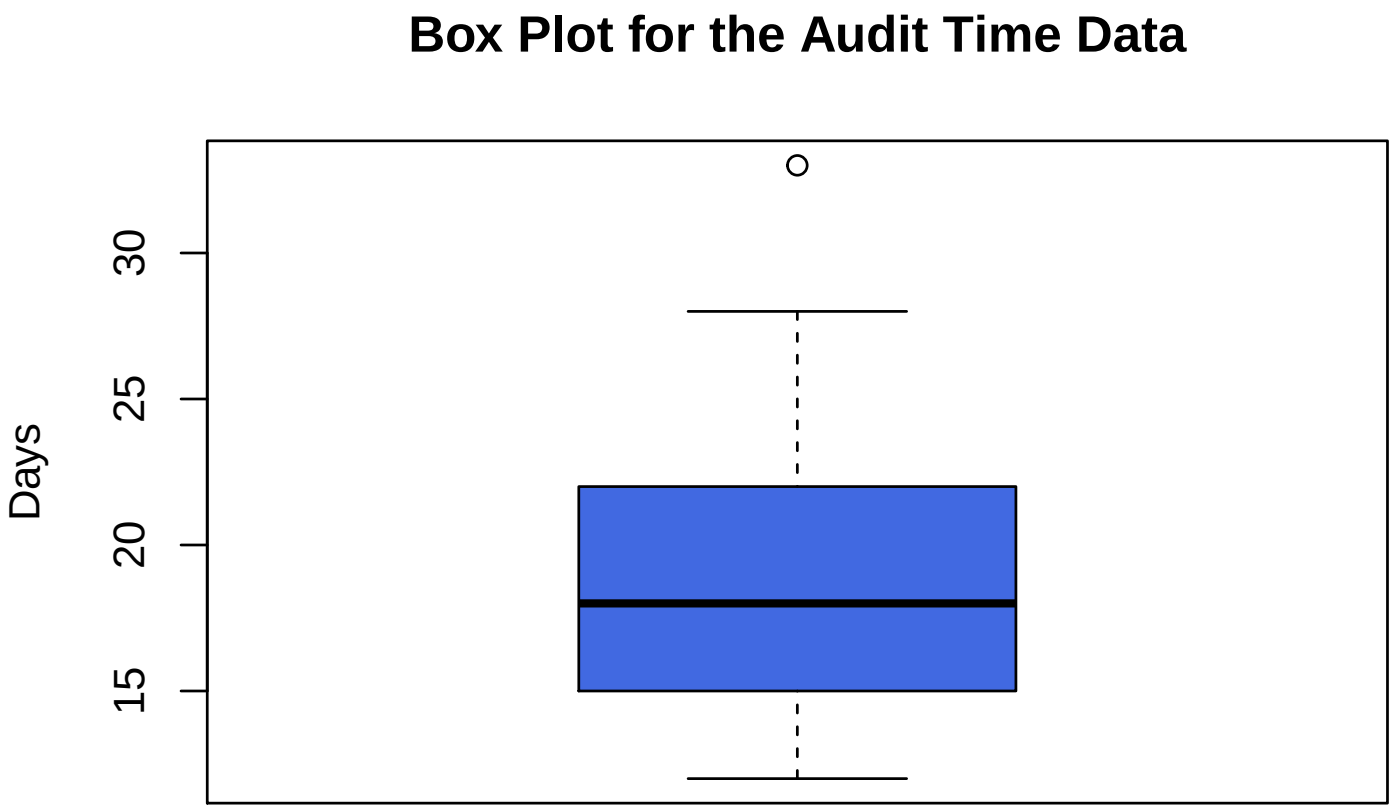
Summarizing quantitative data: Numerical measures

- Location (mean, median, mode, percentiles, quartiles)
- Variability / Dispersion (range, variance, standard deviation)

Summarizing quantitative data: Histogram



Summarizing quantitative data: Box plot



EDA: One variable (1D) vs. Two variables (2D)

EDA (Exploratory data analysis) refers to descriptive statistics.

- Consumers' soft drink purchases
- Accounting firm's audit time in days
- Passengers' flight satisfaction ratings

EDA: One variable (1D) vs. Two variables (2D)

EDA (Exploratory data analysis) refers to descriptive statistics.

- Consumers' soft drink purchases
- Accounting firm's audit time in days
- Passengers' flight satisfaction ratings

Research questions about the relationship between two variables:

- Consumers' soft drink purchases - Gender: Do females and males have different patterns?
- Accounting firm's audit time in days - Position: Do junior accountants take more audit time than senior accountants?
- Passengers' flight satisfaction ratings - ?

Relationship between two variables

- Consumers' soft drink purchases (Qualitative) - Gender (Qualitative)
- Accounting firm's audit time in days (Quantitative) - Position (Qualitative)
- Passengers' flight satisfaction ratings (Quantitative) - Age (Quantitative)

Relationship between two variables

- Consumers' soft drink purchases (Qualitative) - Gender (Qualitative)
- Accounting firm's audit time in days (Quantitative) - Position (Qualitative)
- Passengers' flight satisfaction ratings (Quantitative) - Age (Quantitative)
- Example: YouTube data and Intro Statistics Survey data.

YouTube data

VARIABLES

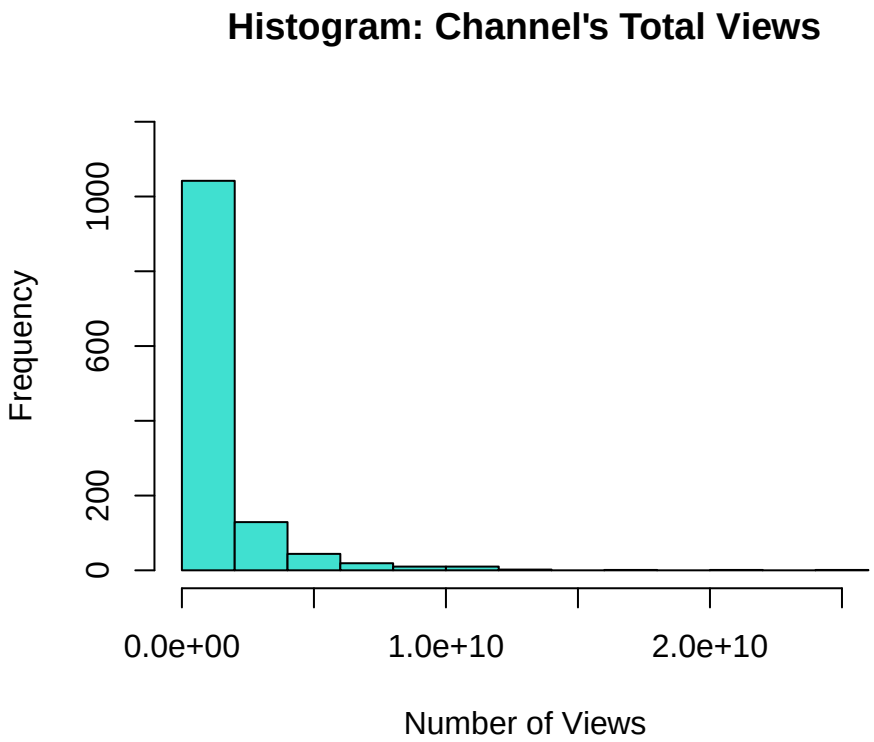
- name: Name of the YouTube channel
- category: User-defined channel topic
- country: The country of origin of the channel
- accountAge: The age of the account in weeks. Note that for consistency the age calculation was performed on December 31 2018.
- videoUploads: The amount of videos uploaded by the channel.
- subscribers: The number of subscribers to the channel
- views: The total views across all videos
- viewsPerVideo: Total views divided by videos
- continent: Continent of origin of the channel

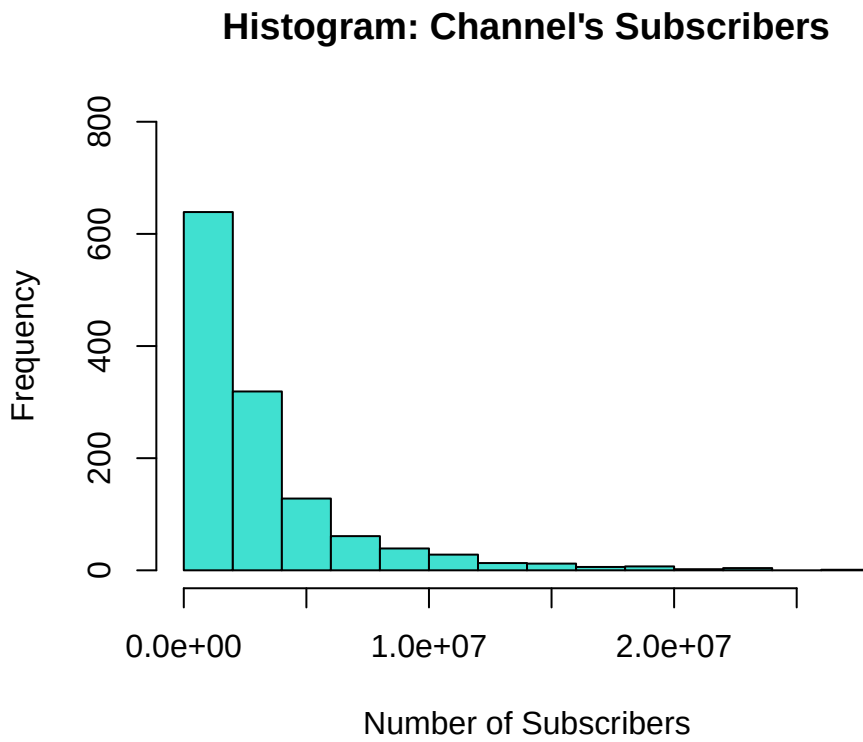
Building a hypothesis

- subscribers: The number of subscribers to the channel
- views: The total views across all videos

EDA for two quantitative variables: Graphical summary

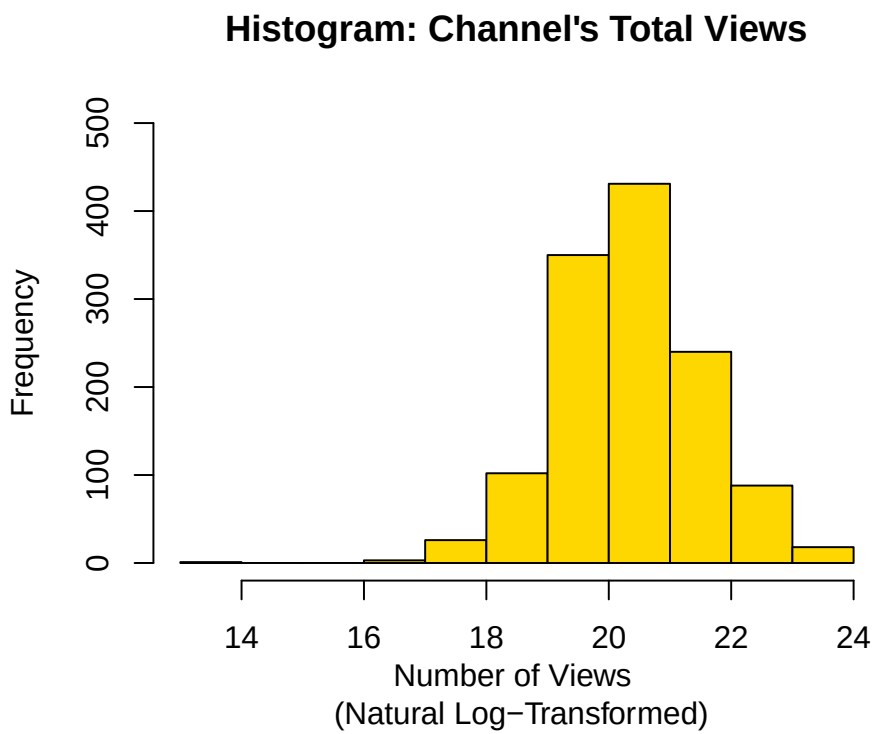
Histogram for each variable:

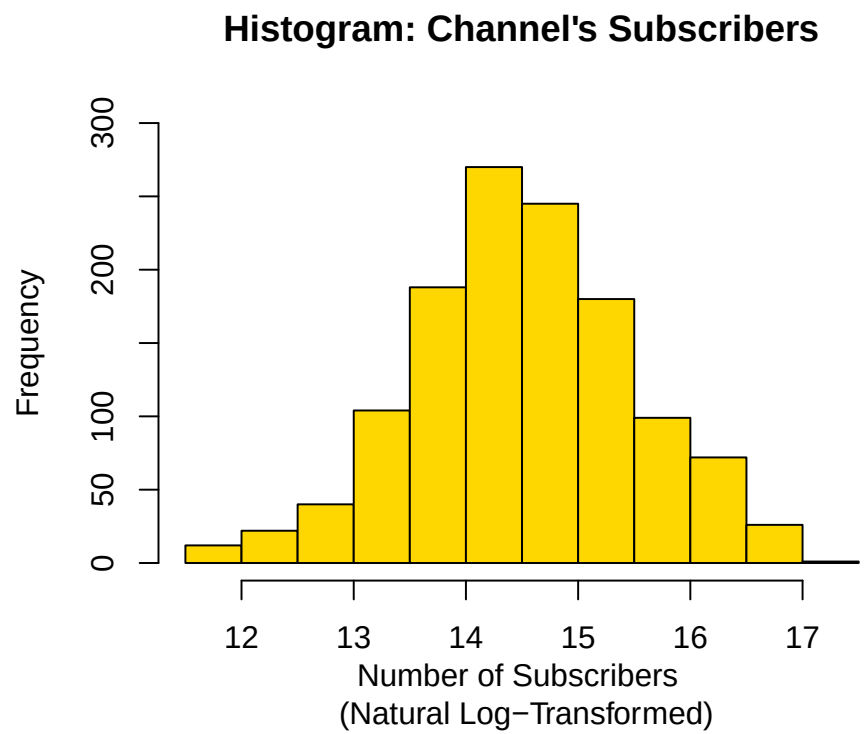




EDA for two quantitative variables: Graphical summary

Histogram for each variable after log transformation:

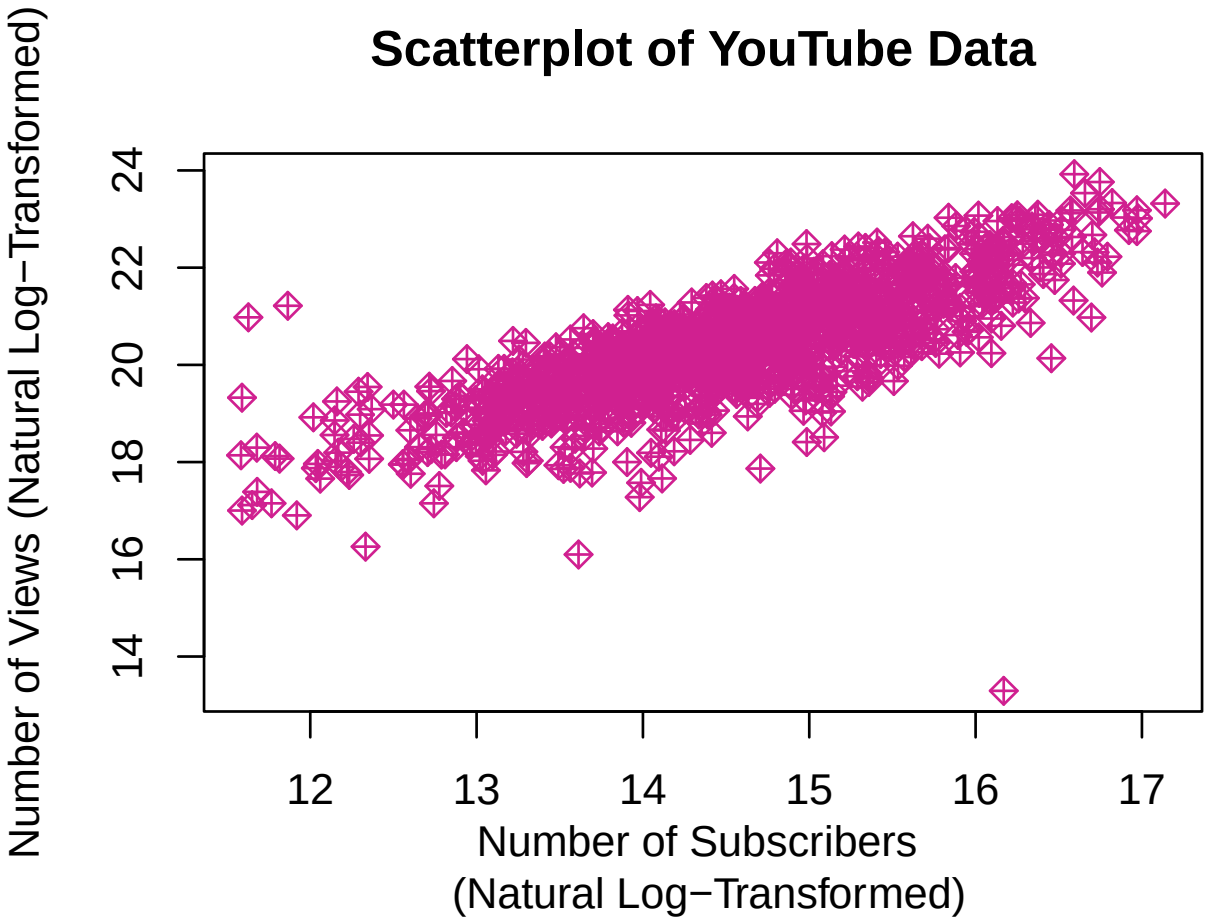




$log_e(24154953) = 17; \quad e^{17} = \exp(17) = 24154953.$

EDA for two quantitative variables: Graphical summary

Scatterplot after log transformation:



EDA for two quantitative variables: Numerical summary

Correlation coefficient

For sample data (x_i, y_i), $i = 1, 2, 3, ..., n$,

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- r_{xy} = correlation coefficient
- s_{xy} = covariance
- s_x = sample standard deviation of x
- s_y = sample standard deviation of y

- $-1 \leq r_{xy} \leq 1$.
- As a rule of thumb,
  - a correlation coefficient between $|1|$ and $|0.7|$: high correlation.
  - between $|0.7|$ and $|0.5|$: moderate.
  - between $|0.5|$ and $|0.3|$: low.
  - between $|0.3|$ and 0: negligible.
- In outer YouTube example, $r_{xy} = 0.767$.

Intro Statistics Survey data

VARIABLES:

- Course: Course that the Respondent was enrolled in
- Math: Math SAT Score
- Verbal: Verbal SAT Score
- HT: Respondent's Height (in inches)
- Shoe: Shoe Size (US)
- Gender: Respondent's Gender
- MomHT: Height of Respondent's Mother (in inches)
- DadHT: Height of Respondent's Father (in inches)
- Color: Favorite Color
- WT: Respondent's Weight (in pounds)
- Major: Declared Major

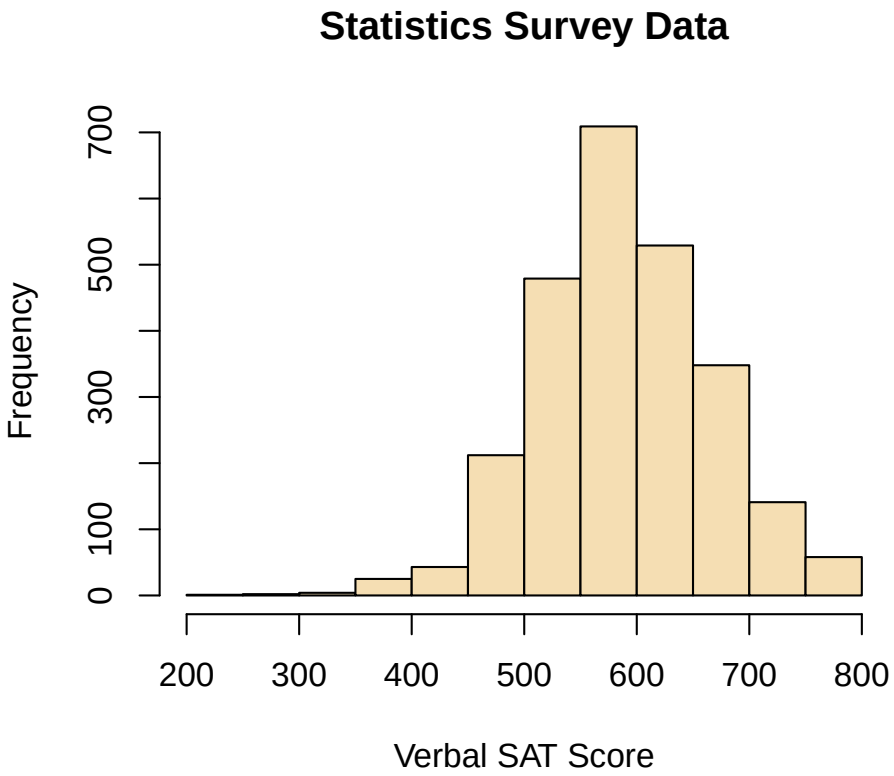
Building a hypothesis

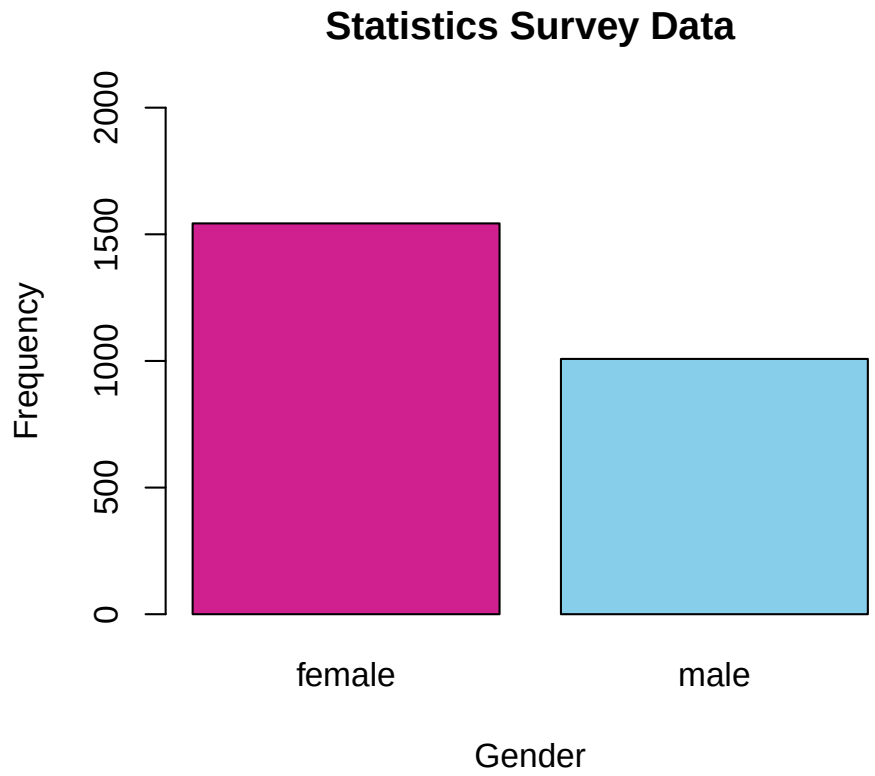
- Verbal: Verbal SAT Score
- Gender: Respondent's Gender

Is gender associated with verbal SAT scores?

EDA for two variables (qualitative-quantitative): Graphical summary

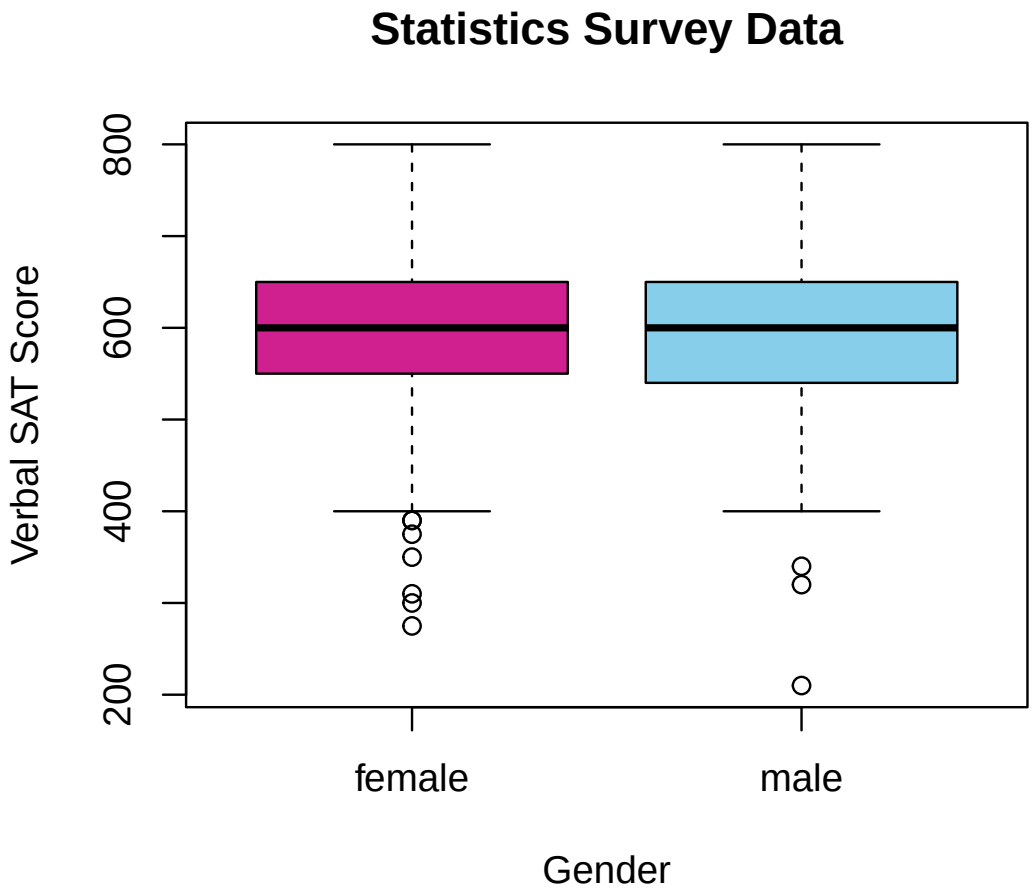
Each variable:





EDA for two variables (qualitative-quantitative): Graphical summary

A side-by-side box plot:



EDA for two variables (qualitative-quantitative): Numerical summary

	Female	Male
Sample Mean	598.39	594.88
Sample Median	600	600
Sample Std. Dev.	74.21	78.65

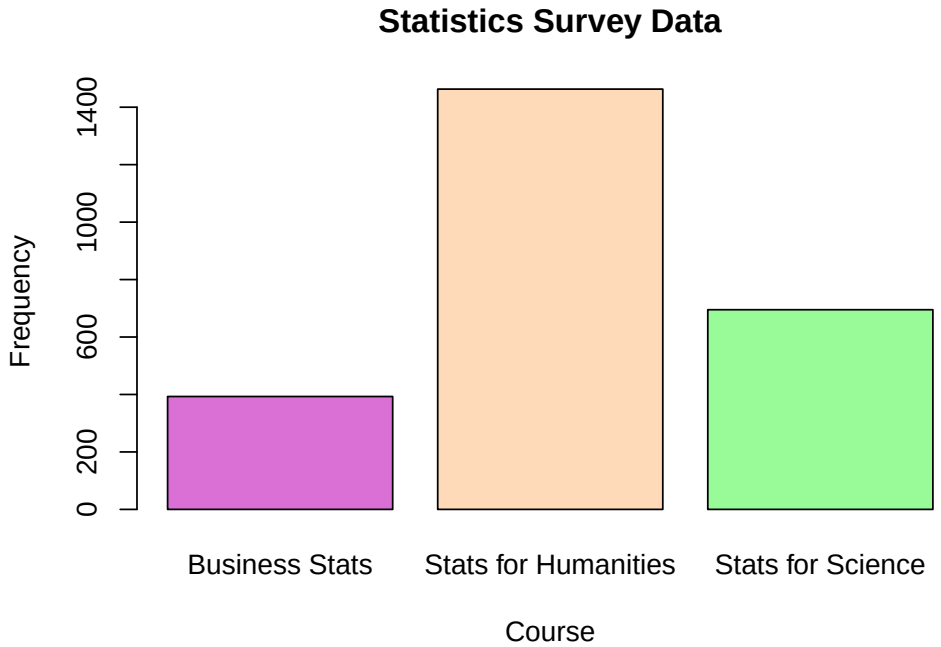
TABLE: Numerical summary for the verbal SAT score by gender

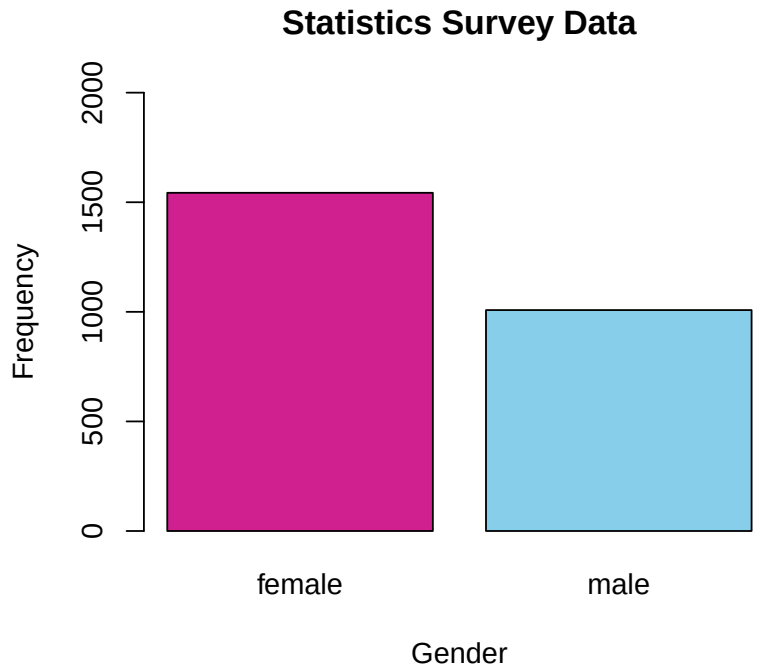
Building a hypothesis

- Course: Course that the Respondent was enrolled in [Qualitiative]
- Gender: Respondent's Gender [Qualitiative]

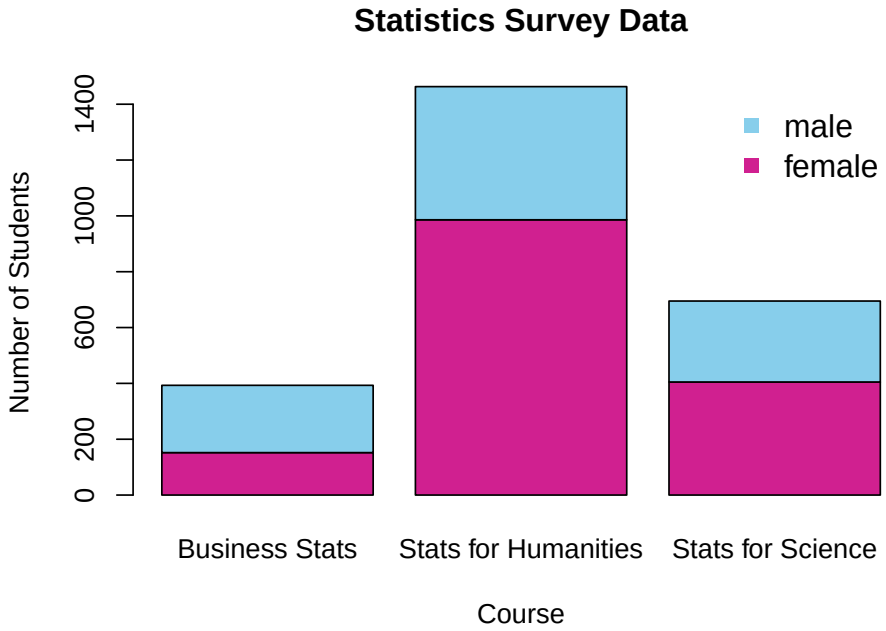
EDA for two qualitative variables: Graphical summary

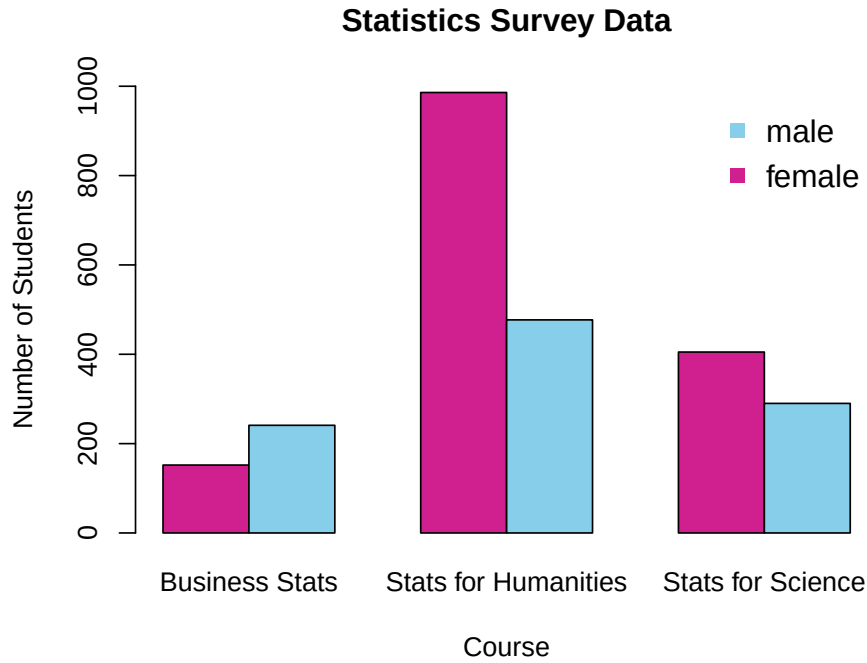
Each variable:





EDA for two qualitative variables: Graphical summary





EDA for two qualitative variables: Numerical summary

	Female	Male	Total
Business Stats	152 (38.7%)	241 (61.3%)	393
Stats for Humanities	986 (67.4%)	477 (32.6%)	1463
Stats for Science	405 (58.3%)	290 (41.7%)	695

TABLE: Contingency table (a.k.a. cross-tabulation) for the enrolled course by gender

Our hypotheses

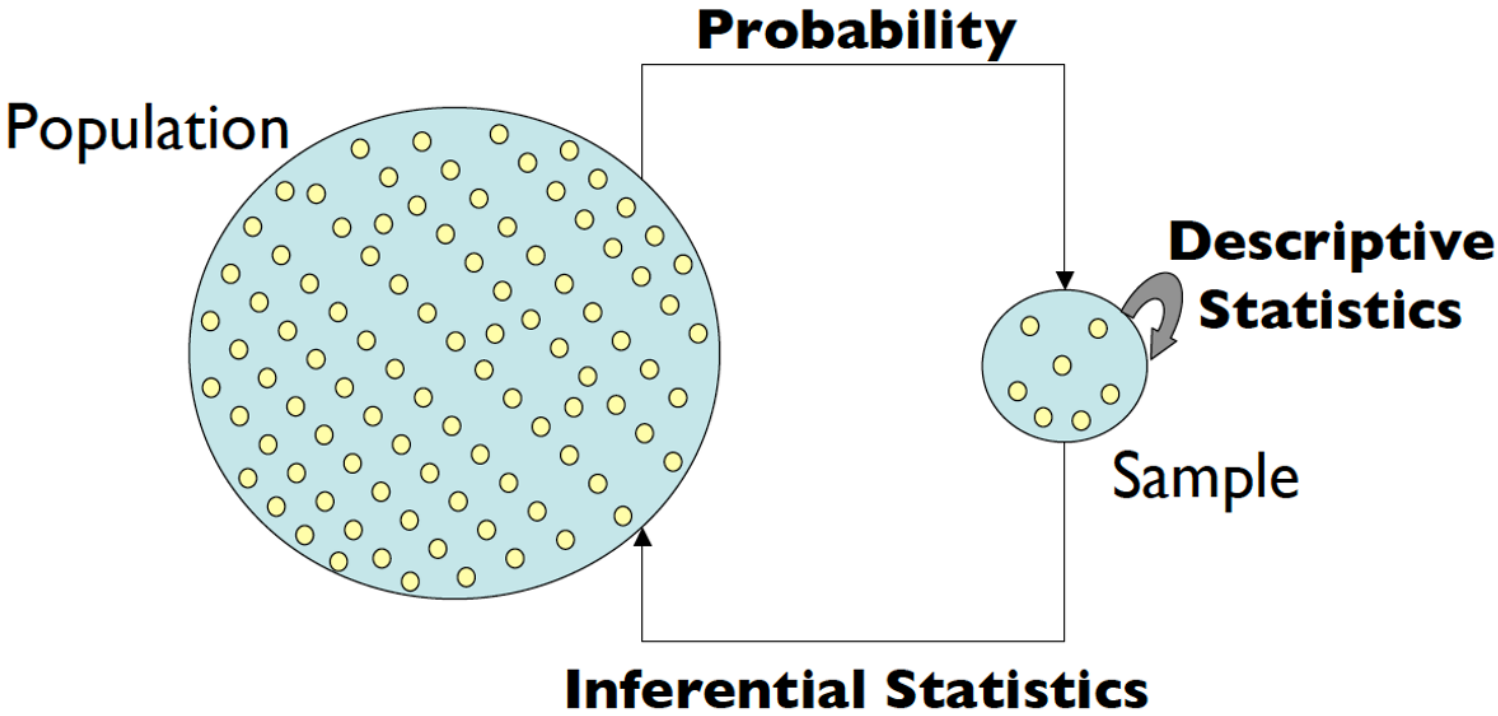
- subscribers: The number of subscribers to the channel
- views: The total views across all videos

- Verbal: Verbal SAT Score
- Gender: Respondent's Gender

- Course: Course that the Respondent was enrolled in
- Gender: Respondent's Gender

For these hypotheses, we conducted EDA. Are we done now? Can we draw our final conclusions from the EDAs?

Inferential statistics and descriptive statistics (EDA)



From Joshua Akey, https://www.gs.washington.edu/academics/courses/akey/56008/lecture.htm.

Brief introduction to inferential statistics

- Purpose: Derive conclusions about a **population** based on a **sample** data. In other words, estimate a parameter of the population using a sample statistic of the sample data.

Brief introduction to inferential statistics

- Purpose: Derive conclusions about a **population** based on a **sample** data. In other words, estimate a parameter of the population using a sample statistic of the sample data.

- **Population**: The total set of subjects of interest in a study.
- **Parameter**: Numerical summary of the population. For example, population mean or population median.

- **Sample**: The subset of the population on which the study collects data.
- **Sample statistic**: Numerical summary of the sample data. For example, sample mean or sample median.

Brief introduction to inferential statistics

In our YouTube example,

- **Sample data**: The total number of views and the total number of subscribers for 1,259 YouTube channels from a dataset gathered by Social Blade that contains information up to the end of 2018.

What is the population data for our study?

Brief introduction to inferential statistics

- Inference can't avoid an error due to differences between population data and sample data.

- Inferential statistics use probability theory and statistical methods to take into account differences between population data and sample data, and then draw conclusions.
  - Probability
  - Sampling distribution
  - Confidence interval
  - Statistical significance tests

Statistical significance tests

In our YouTube example,

- Correlation coefficient between subscribers and views $= 0.767$.
- Suggesting a high correlation.

Statistical significance tests

In our YouTube example,

- Correlation coefficient between subscribers and views = 0.767.
- Suggesting a high correlation.

- The coefficient is from our sample data for 1,259 YouTube channels as of 2018.
- What if we collect more recent data as our sample, say from 2018 - 2023 and calculate the correlation coefficient?
- Will we get the same value 0.767?

Statistical significance tests

In our YouTube example,

- Correlation coefficient between subscribers and views = 0.767.
- Suggesting a high correlation.

- The coefficient is from our sample data for 1,259 YouTube channels as of 2018.
- What if we collect more recent data as our sample, say from 2018 - 2023 and calculate the correlation coefficient?
- Will we get the same value 0.767?
- How about our new sample was drawn for 2015 - 2021?

Statistical significance tests

- Sample statistic (e.g., correlation coefficient from the sample data) $\neq$ Population parameter (e.g., true relationship between YouTube subscribers and views).
- A statistical significance test is designed to evaluate whether our estimation (i.e., sample statistic) is **large enough to reject the null hypothesis**, where the null hypothesis claims that the population parameter is zero (generally).

Statistical significance tests

- Sample statistic (e.g., correlation coefficient from the sample data) $\neq$ Population parameter (e.g., true relationship between YouTube subscribers and views).
- A statistical significance test is designed to evaluate whether our estimation (i.e., sample statistic) is **large enough to reject the null hypothesis**, where the null hypothesis claims that the population parameter is zero (generally).

- Sufficiently large sample statistic $\Rightarrow$ A large test statistic value $\Rightarrow$ A small probability that the null hypothesis is true (i.e., A small p-value). $\Rightarrow$ Reject the null hypothesis.
- Conventional significance level for "sufficiently large" is 0.05.
- To summarize, the decision rule is: If p-value < 0.05, you reject the null hypothesis and conclude that your sample statistic (or estimated value) is statistically significant at the 0.05 significance level.

Various statistical significance tests

- t test
- ANOVA
- χ^2 test
- Linear regression analysis
- Logistic regression analysis
- Multinomial/Ordinal logistic regression analysis

How to choose the right method for your research?

Various statistical significance tests

- t test
- ANOVA
- χ^2 test
- Linear regression analysis
- Logistic regression analysis
- Multinomial/Ordinal logistic regression analysis

How to choose the right method for your research?

It is determined by the data type.

Various statistical significance tests for 2D cases

- t test: Qualitative (2 groups) - Quantitative
- ANOVA: Qualitative (3 or more groups) - Quantitative
- χ^2 test: Qualitative - Qualitative
- Bivariate linear regression analysis: Quantitative - Quantitative

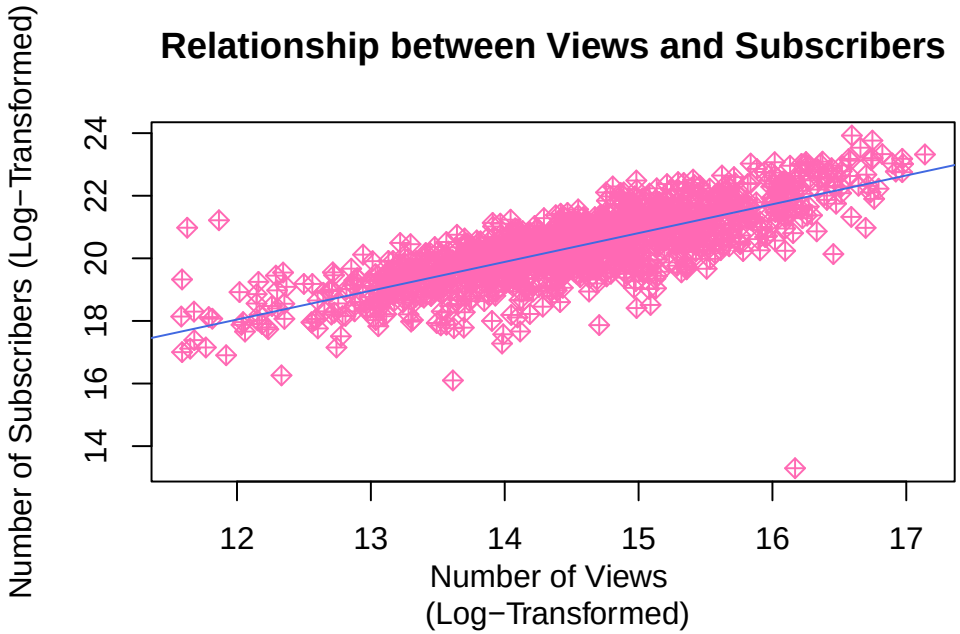
Our hypotheses revisited

- Research hypothesis: The number of subscribers to the channel is influenced by the total views across all videos.
- Null hypothesis: The number of subscribers to the channel is **NOT** influenced by the total views across all videos.

Our hypotheses revisited

- Research hypothesis: The number of subscribers to the channel is influenced by the total views across all videos.
- Null hypothesis: The number of subscribers to the channel is **NOT** influenced by the total views across all videos.

Quantitative - Quantitative: Bivariate linear regression



- $\widehat{\text{log\_views}} = 6.995 + 0.921\ \text{log\_subscribers}$ $\Rightarrow$ A 1% increase in subscribers is predicted to increase views by about 0.92%.
- p-value for the estimated impact $(0.921) \approx 0$. $\Rightarrow$ 0.921 is statistically greater than zero at $\alpha = 0.05$.

Our hypotheses revisited

- Verbal: Verbal SAT Score
- Gender: Respondent's Gender

Our hypotheses revisited

- Verbal: Verbal SAT Score
- Gender: Respondent's Gender

Two independent means (quantitative) from two groups (qualitative): *t*-test.

- Test statistic value $= 1.127$, degrees of freedom $= 2063.1$, *p*-value$=0.26 \Rightarrow$ Female students' verbal SAT scores and Male students' verbal SAT scores are not statistically different at $\alpha = 0.05$.

Our hypotheses revisited

- Course: Course that the Respondent was enrolled in
- Gender: Respondent's Gender

Our hypotheses revisited

- Course: Course that the Respondent was enrolled in
- Gender: Respondent's Gender

Test for independence between Course (qualitative) and Gender (qualitative): χ^2-test.

- Test statistic value $= 108.86$, degrees of freedom $= 2$, p-value$= \approx 0 \quad \Rightarrow$ Course and Gender are statistically associated with each other $\alpha = 0.05$. In other words, female students and male students tend to take significantly different statistics courses at $\alpha = 0.05$.