Visualization & Inferences from Data

Hypothesis testing & Inference

IMT 547 - Social Media Data Mining and Analysis

25-Feb-2021 (Week 8, Day 16)

Last Class Review Topics

- Mark and Channels - building blocks of visualization (*theory* behind visualization)
- 1D visualization (review from EDA class)
- Multi-D visualization
- Lab
- Feedback on pitch given from last class — work in groups to revamp/revise

Today's Topics

- Descriptive Statistics
- Inferential Statistics
- Hypothesis testing
  - T-test
  - Wilcoxon
- Lab
- Survey
- In class Project work

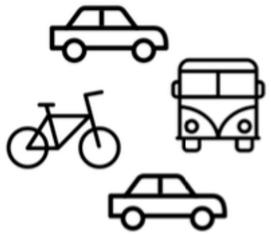
Last class left-over lab

Types of Statistics

Two main branches of statistics

Description Statistics

- Describe and Summarize Data
- Basic descriptive statistics to tell the reader about the participants in the sample that you have collected



- 50% of friends drive to work
- 25% take the bus
- 25% bike

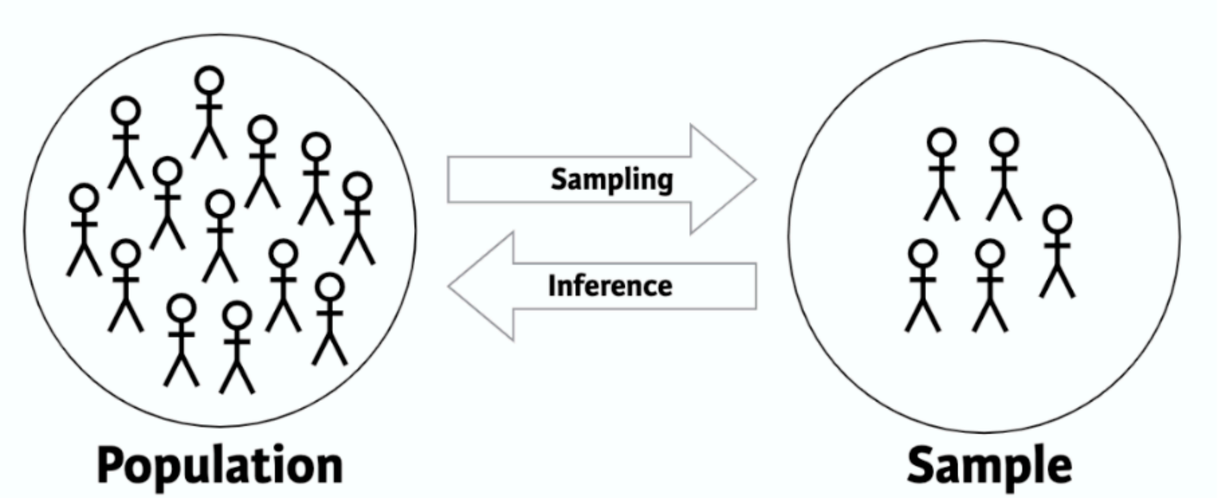
Inferential Statistics

- Use a sample of data to make inference about a larger population

What percent of people drive to work?

Inferential Statistics

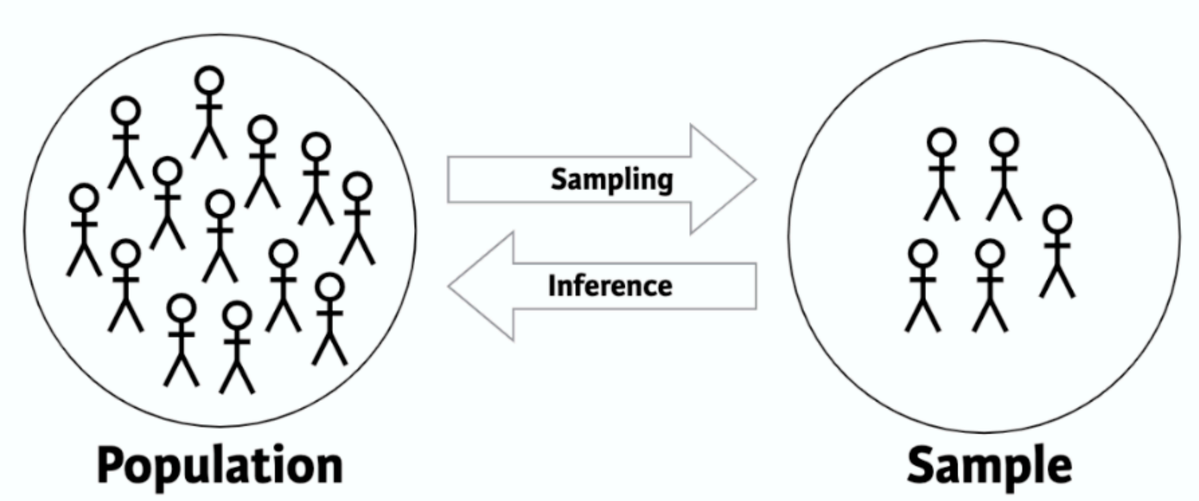
Process of making claims about a population based on information from a sample.



Example: You want to know how does a drug treat diabetes? What you really want to know is how the does the drug treats all people with diabetes in the population, not just a few.

Inferential Statistics Another example

Process of making claims about a population based on information from a sample.

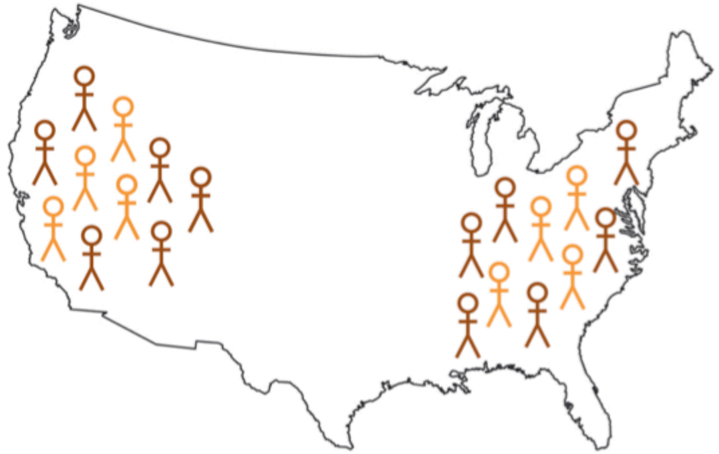


Example election polls: You conducted survey of voters to know their voting preference so as to estimate how will the overall population vote on election day.

Inference Statistics

Another example

You are trying to convince your marketing director that the people on the East coast prefer cola over orange soda compared to people in the West Coast. You start by collecting sample data to prove your point about the population.



You start by assuming that there is no difference between the populations. They are the same (H0 or null hypothesis)

Inference Statistics

Now you start sampling (first sample)

Here the samples are same and hence the population. Both east and west coast have 2/3 people who prefer cola

Inference Statistics

The sample data (second sample)

But here the samples are different now. Twice are high…

If you continue repeated sampling and see the data is extremely different between the two population, then your initial assumption of equal population is invalid. And you have proved your point that soda preference is different between East and West coast.

Ha (alternative hypothesis) holds

Descriptive / Inferential

Exercise

1. Given data on all 100,000 people who viewed an ad, what percent of people clicked on it?

2. After interviewing 100 customers, what percent of *all* of your customers are satisfied with your product?

3. Given data on 20 fish caught in a lake, what's the average weight of all fish in the lake?

4. Given data on every customer service request made, what's the average time it took to respond?

Major aspects of Descriptive Statistics

You need all three to get a good description of the data!

- Measures of central tendency (*what's the typical value or central value of the data?*)
- Measures of variability (*what's the spread of the data?*)
- Measures of distribution (*how is the data distributed or shape of the data?*)

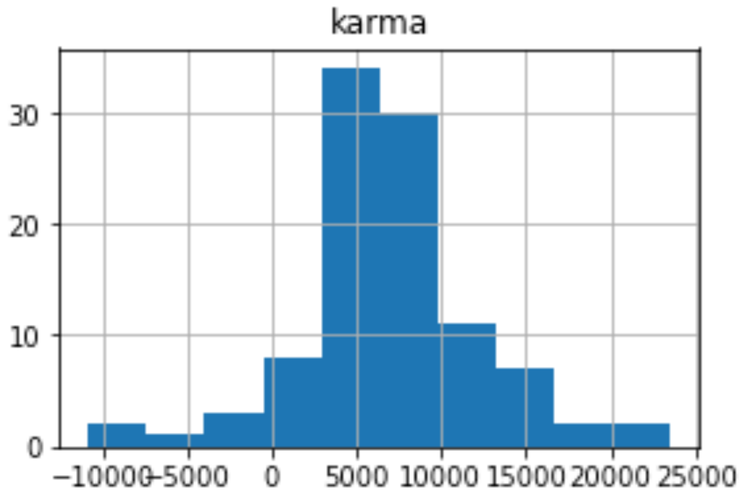
Measures of central tendency

Typically how verbose are users? What's the typical value of karma?

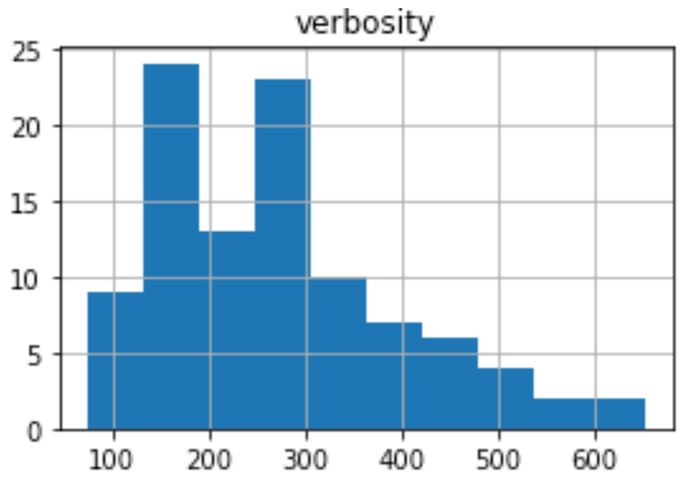
What's a **typical value** of this distribution?

Where is the center of the data?

- Mean
- Median
- Mode

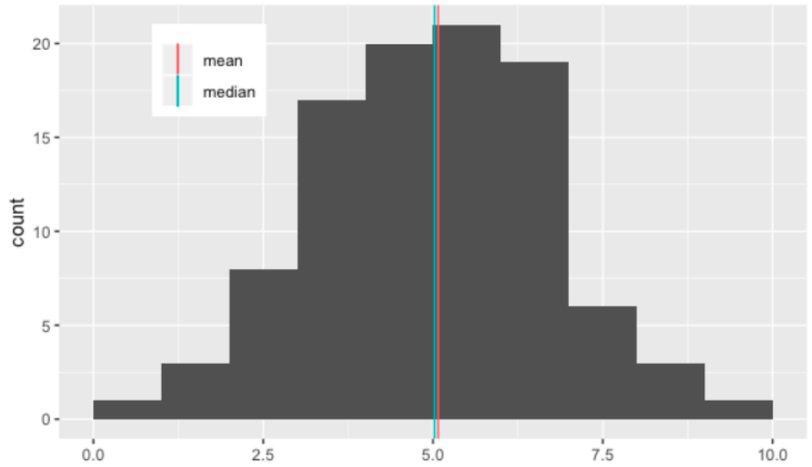


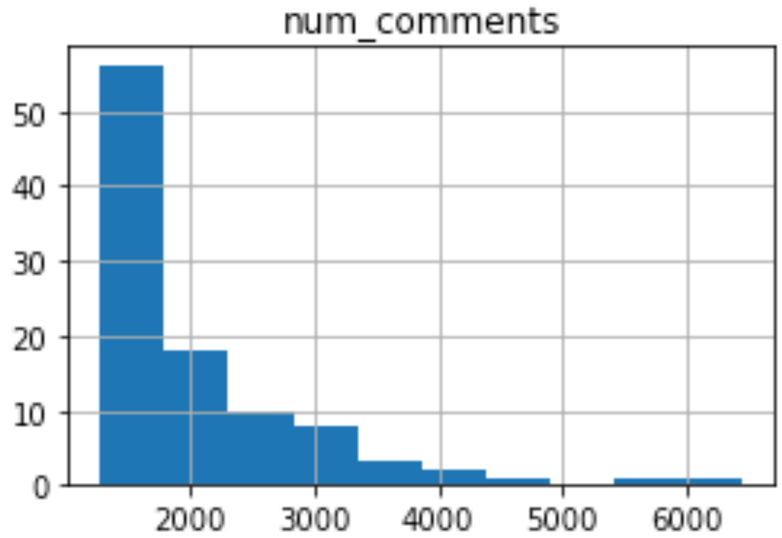
Viz lab



Mean gets effected by outliers

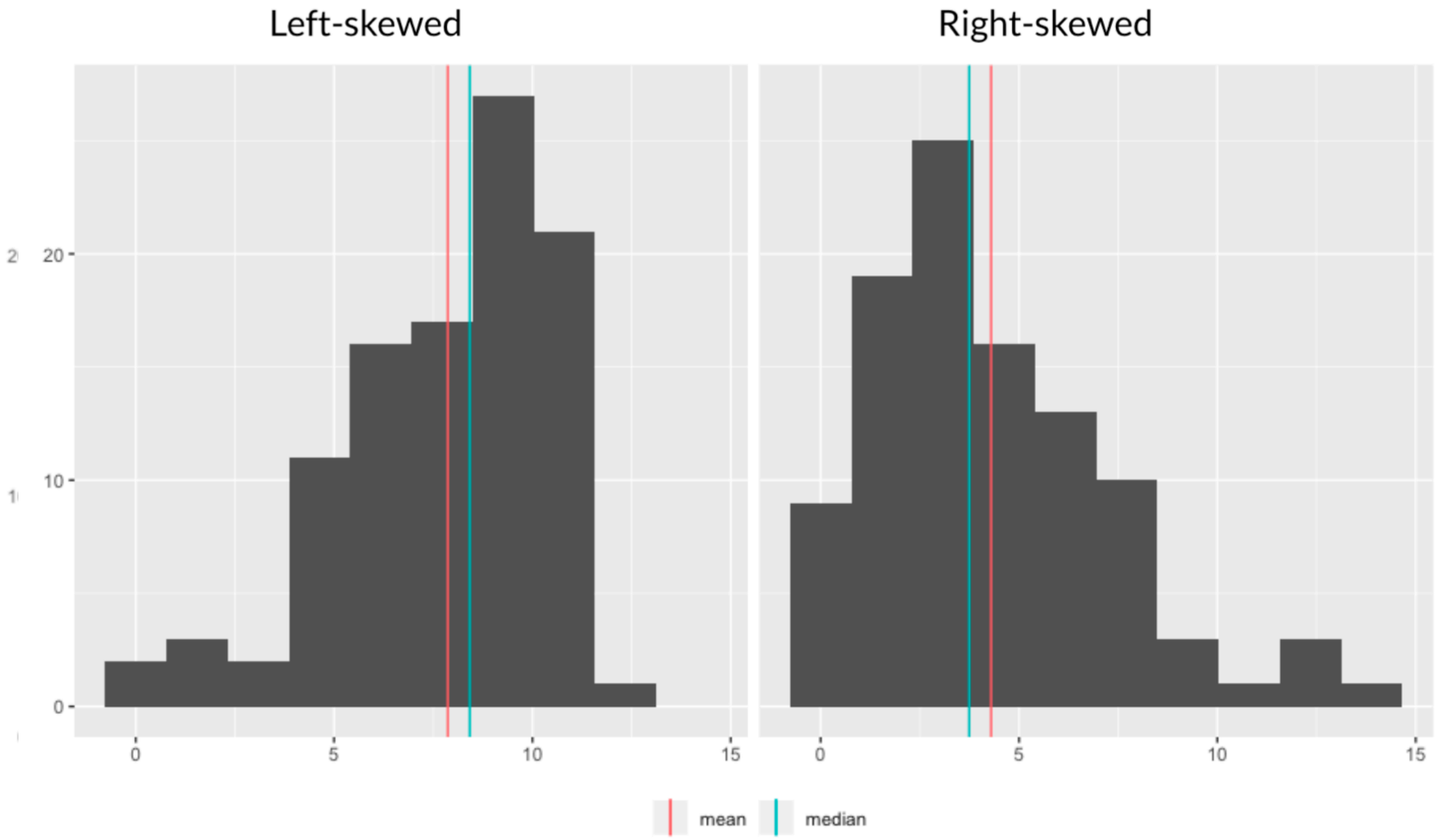
Since mean is sensitive to extreme values, it works better for symmetrical data like below





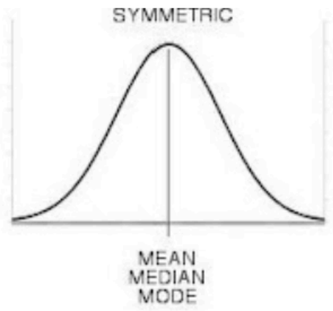
Which measure to use?

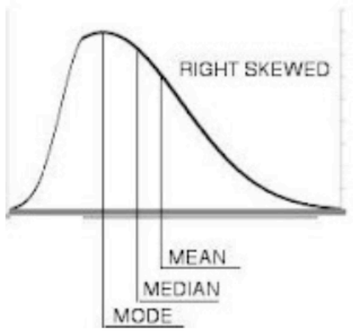
When data is skewed (not symmetric) better to use median

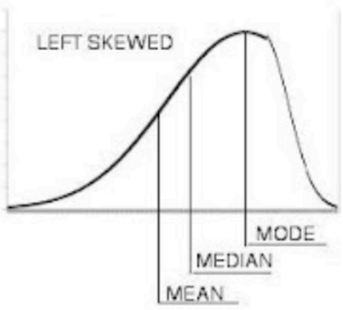


Mean, Median, and Mode

Measures of Central Tendency







http://thenormalgenius.blogspot.com/

Mean: arithmetic average
- Application: interval or ratio data
- Formula: add up all values and divided by the number of observations

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

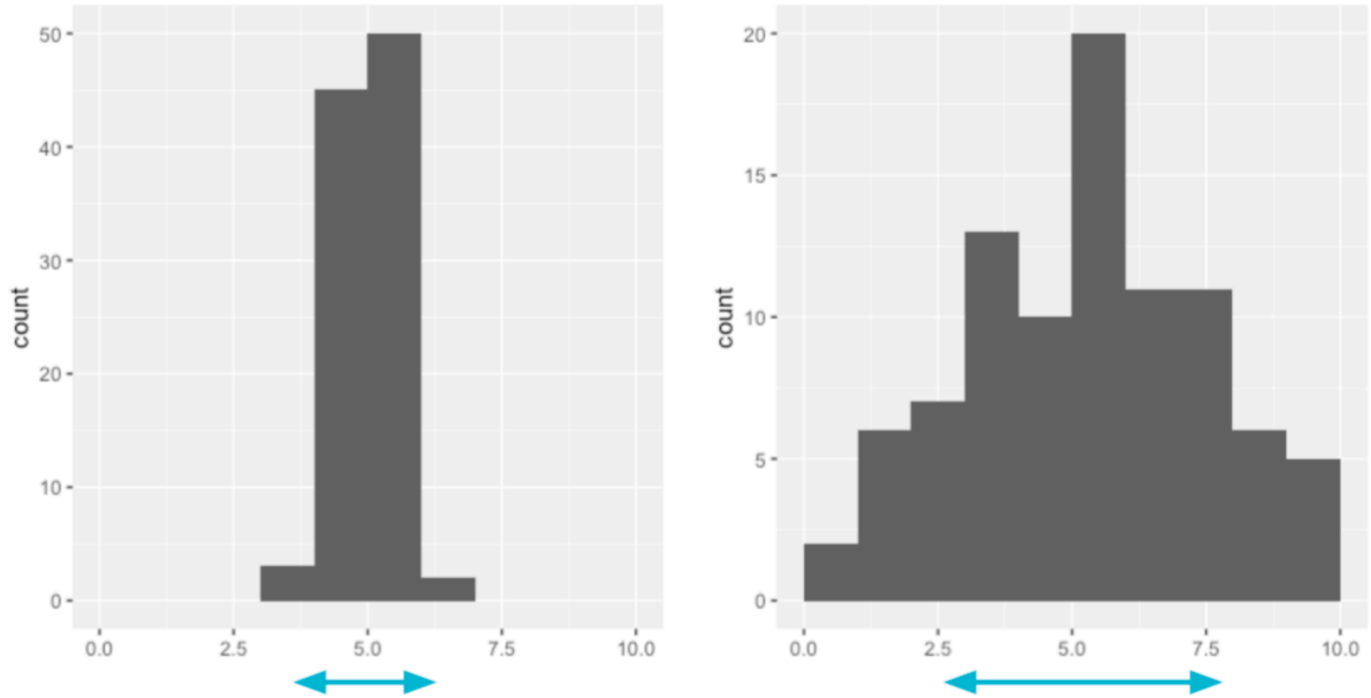
Median: midpoint of the distribution
- Application: ordinal data, when concern is skewed distributions
- Formula: arrange scores in order and find the middle

Mode: the most common observed value
- Application: nominal data (think about bimodal distributions!)
- Formula: arrange scores or distribution find most common
- Formula: arrange scores or distribution and find the most common

Measures of Variability

- *what's the spread of the data?*

How spread apart of how close together data points are.



Measures of Variability

- *what's the spread of the data?*

How spread apart of how close together data points are.

- Range: distance between smallest and largest values
- Quartile range: quarters of the rank ordering (e.g. 25th - 75th is the inter-quartile range) . Quartiles split up the data into 4 equal parts.
- Variance: average squared deviation from the mean $\text{Variance} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$
- Standard deviation: degree of dispersion in the data

$$SD = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$$

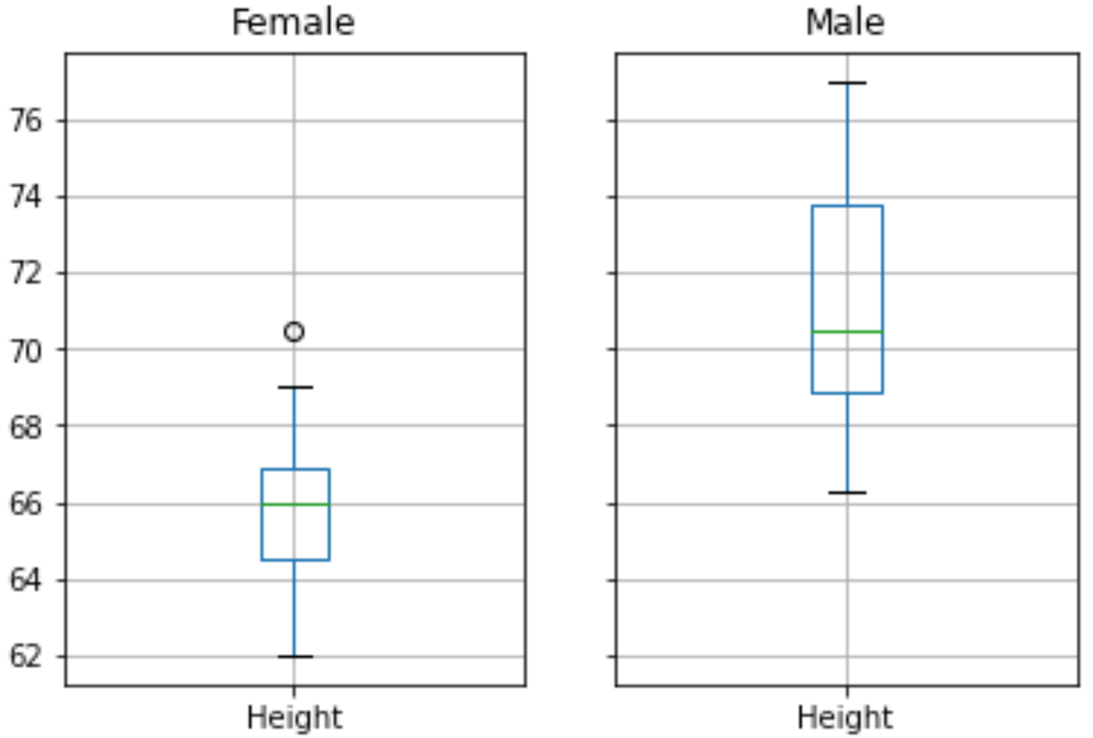
Quartiles

Quartiles split the data into 4 equal parts

Rank order your data

25% of your data is between 1.90 to 7.85.

The next 25% is between…..



Second quartile or 50th percentile = ? Median

1st quartile: Bottom of box

2nd quartile: Top of box

Middle line: Median

Outliers

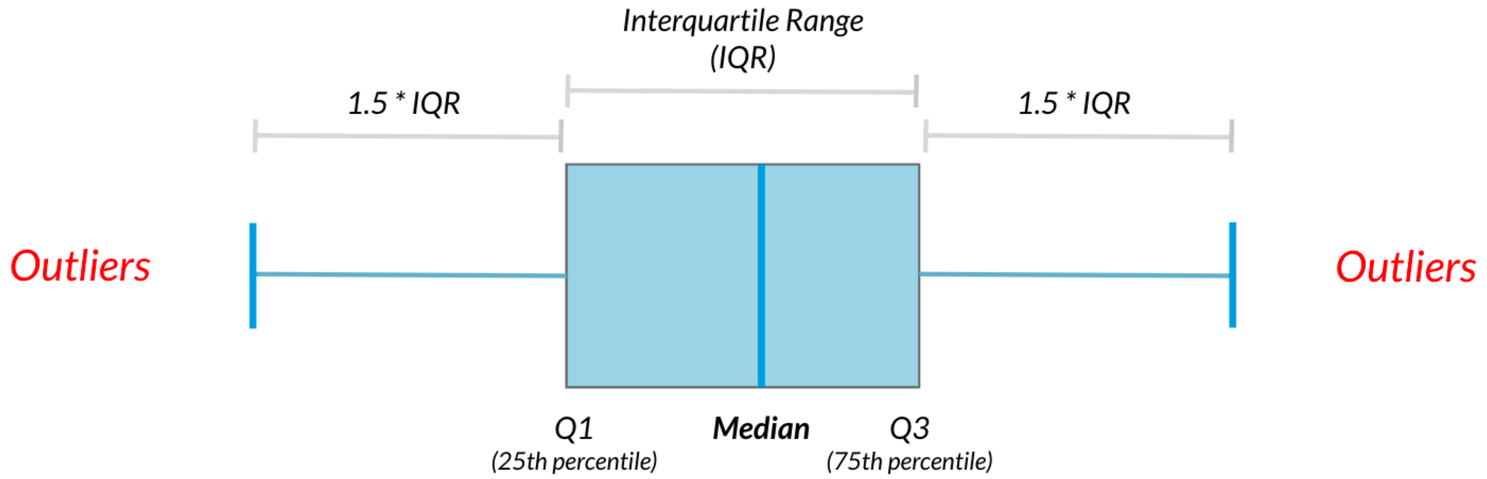
Data point that is substantially different from the others

Rule of thumb: A data point is an outlier if:

data < Q1 — 1.5 X IQR

Or

data > Q3 + 1.5 X IQR



Why we need to visualize: Statistics might not give you the full picture

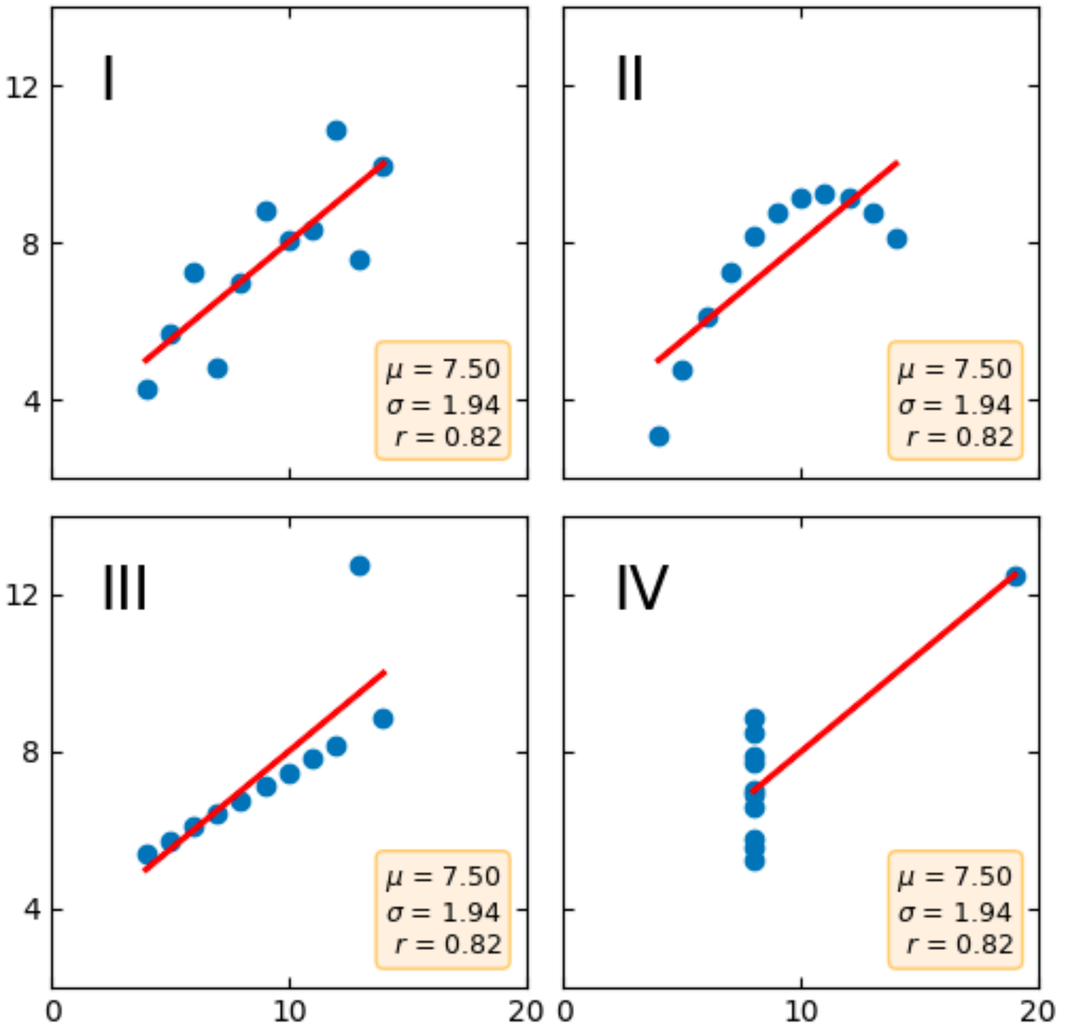
Example: Anscombe's quartet. Dataset consists of 4 pairs of x and y data (x1,y1), (x2,y2)…that have the same mean, standard deviation, and regression line, but which are qualitatively different.

Non-Visual approach

The 6 statistical assessments of these four datasets are identical

Property	Value
Mean of x in each case	9
Sample variance of x in each case	11
Mean of y in each case	7.50
Sample variance of y in each case	4.12
Correlation between x and y in each case	0.816
Linear regression line in each case	y = 3.00 + 0.500x

Visual approach



Statistics might not give you the full picture

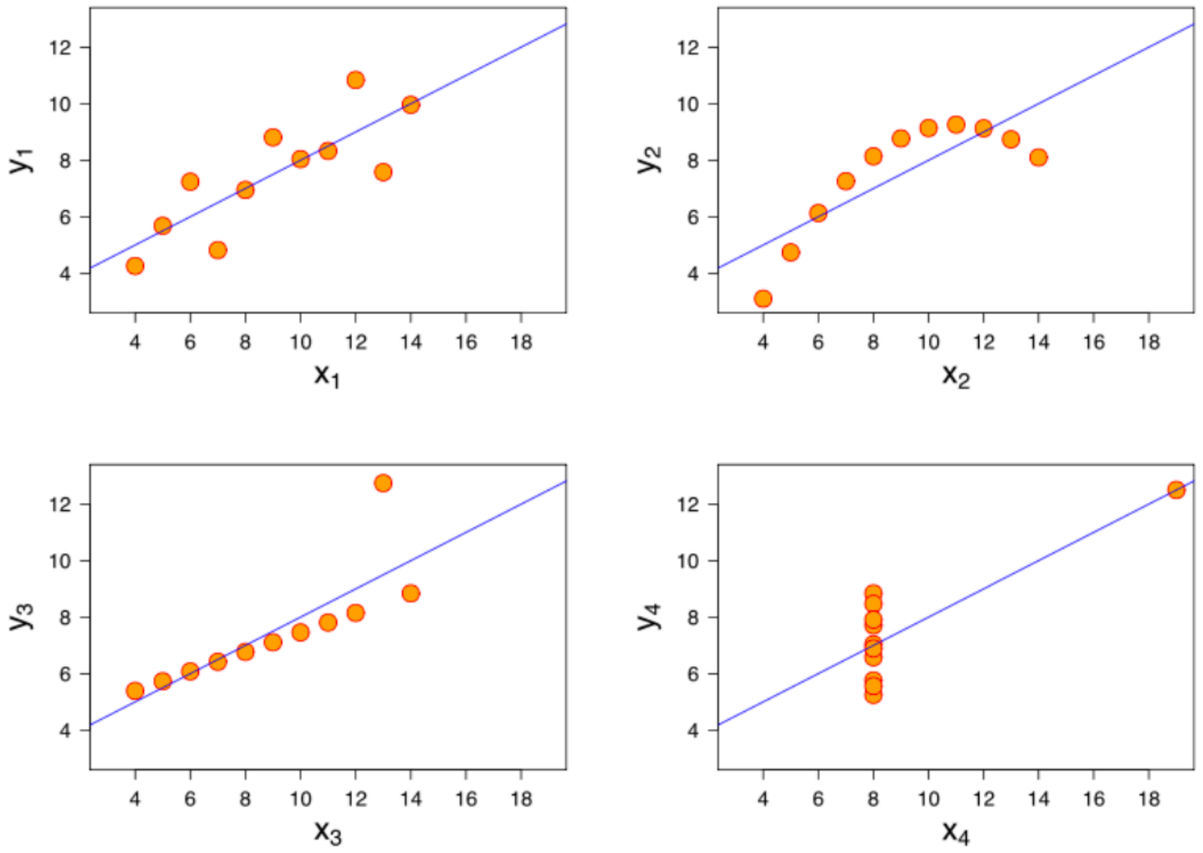
Example: Anscombe's quartet. Dataset included in R and consists of 4 pairs of x and Y data (x1,y1), (x2,y2)…

Non-Visual approach

The 6 statistical assessments of these four datasets are identical

Property	Value
Mean of x in each case	9
Sample variance of x in each case	11
Mean of y in each case	7.50
Sample variance of y in each case	4.12
Correlation between x and y in each case	0.816
Linear regression line in each case	$y = 3.00 + 0.500x$

Visual approach



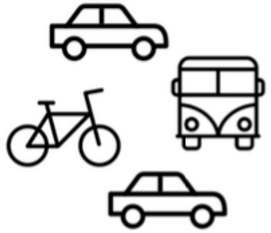
Anscombe's Quartet, 1973

Types of Statistics

Two main branches of statistics

Description Statistics

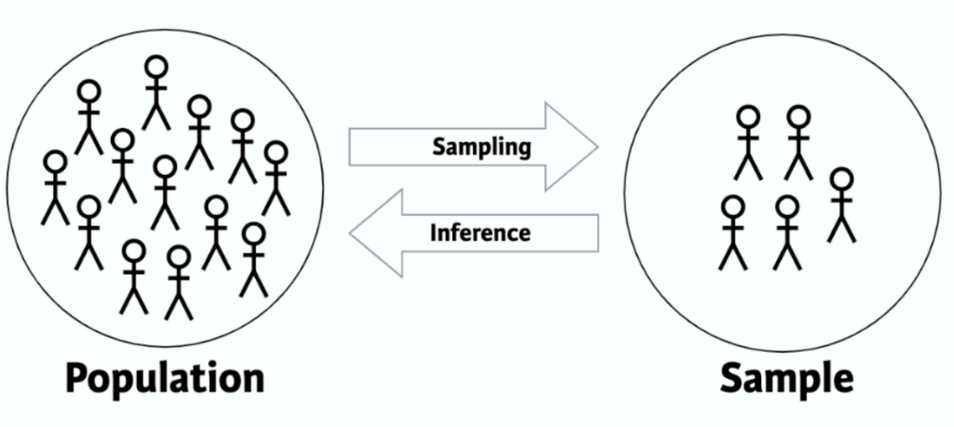
- Describe and Summarize Data
- Basic descriptive statistics to tell the reader about the participants in the sample that you have collected



- 50% of friends drive to work
- 25% take the bus
- 25% bike

Inferential Statistics

- Use a sample of data to make inference about a larger population



Early Public Responses to the Zika-Virus on YouTube: Prevalence of and Differences Between Conspiracy Theory and Informational Videos

Adina Nerghes Peter Kerkhof Iina Hellsten

Table 1: Descriptive statistics of video metrics

	Informational (n=23)		Conspiracy(n=12)	
	M	SD	M	SD
Views	205.097	211.859	159.224	163.452
Top Level Comments	404	367	274	293
Replies	517	473	361	359
Likes	3340	4417	1627	1981
Dislikes	122	109	102	151
Shares	109	833	689	630

Statistical inference in the papers that you read earlier!

This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News

Benjamin D. Horne and Sibel Adalı
Rensselaer Polytechnic Institute
110 8th Street, Troy, New York, USA
{horneb, adalis}@rpi.edu

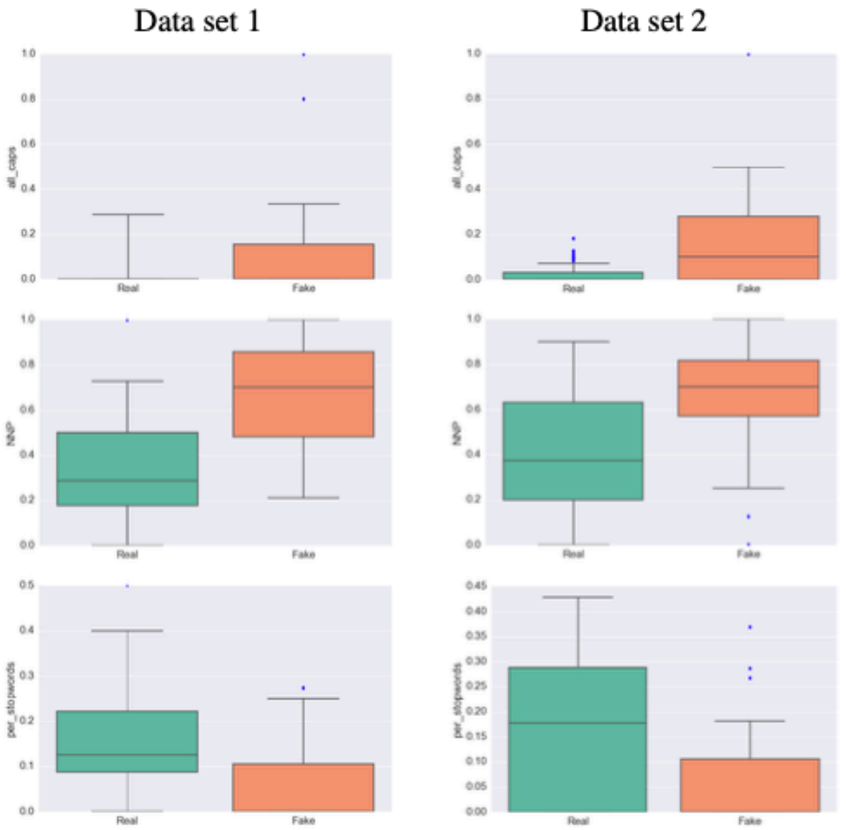


Table 7: 95% Confidence plots of all_caps, NNP, and per_stop. These features were found to be significant across both data sets 1 and 2 for fake and real titles. **Top:** all_caps **Middle:** NNP **Bottom:** per_stop

Hypothesis Test Framework

A framework for using a **sample** to test whether the mean of a population is on one side of a **reference point.**

Intuitive setting:

H0 (Null hypothesis): The claim that is not interesting
Ha (Alternative hypothesis): The claim corresponding to the research questions

The goal is to disprove the null hypothesis

Hypothesis Test Framework

A framework for using a **sample** to test whether the mean of a population is on one side of a **reference point.**

Formal setting:

H0 (Null hypothesis): The reference point we arbitrarily chose. Denote as μ_0.
Ha (Alternative hypothesis): The side of the reference point which we think contains the population mean

The goal is to disprove the null hypothesis, i.e.

- Hypothesis testing attempts to reject the Null hypothesis in favor of the alternative hypothesis
- Hypothesis testing does not directly test the alternative hypothesis, but attempts to reject a reference point far away from it
- Everything will hinge on an arbitrary reference point.

Data Analysis - Inferential Statistics

Hypothesis Testing

Hypothesis: A hypothesis is a statement about some characteristic of a variable or a collection of variable.

Hypothesis Tests: statistical test to test your hypothesis about the population based on the sample data.

1. *Test whether the average writing score (write) differs significantly from 50.*
2. *Are female and male Twitter users distinct in their use of the site?*

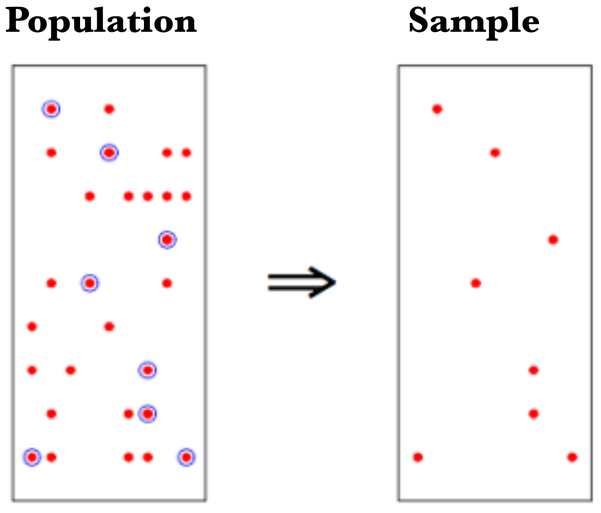
Hypothesis Testing:

H0: Mean of writing score = 50
Ha: Mean of writing score ≠ 50

Sample - 200 observations of high school students (n=200)

Population - all high school students

Larger the sample (large n) the better your estimates



Hypothesis Testing - 5 elements

Assumption - samples are normally distributed, independent.

Test whether the average writing score (write) differs significantly from 50.

Hypothesis Testing:

1. **Hypothesis Write the H0 and Ha**

H0: Mean of writing score = 50
Ha: Mean of writing score ≠ 50

Check your test statistic

Level of significance = p-value
If p-value < 0.05, reject NULL hypothesis

2. Set significance level p-value

3. Calculate Test-statistic

```
t.test(write, mu = 50)
t = 4.1403, df = 199, p-value = 5.121e-05
alternative hypothesis: true mean is not equal to 50
95 percent confidence interval:
 51.45332 54.09668
sample estimates:
mean of x
   52.775
```

4. p-value check.

Decide whether or not to reject the null hypothesis by comparing test statistics

5. Conclusion

p-value < 0.05, reject H0.
Hence, the mean of the writing scores is statistically significantly different from the test value of 50. The mean is significantly higher than 50.

Hypothesis testing

Another example (have to give an election example)

Example: election

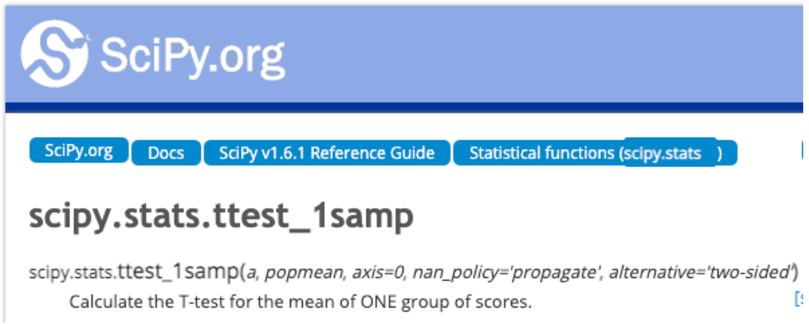
From a sample, the researchers would like to claim that Candidate X will win

H_0: Candidate X will get half the votes

H_A: Candidate X will get more than half the votes

Python's SciPy

SciPy package has many modules for statistical tests.



What test statistic to use?

Previously you saw t-test

- Depends on the **levels of measurement** of the variables
- Depends on the **research question** you are asking

Let's see a few examples

Data Analysis - Inferential Statistics

One sample median test: Test whether a sample median differs significantly from a hypothesized value.

Test whether the median writing score (write) differs significantly from 50.

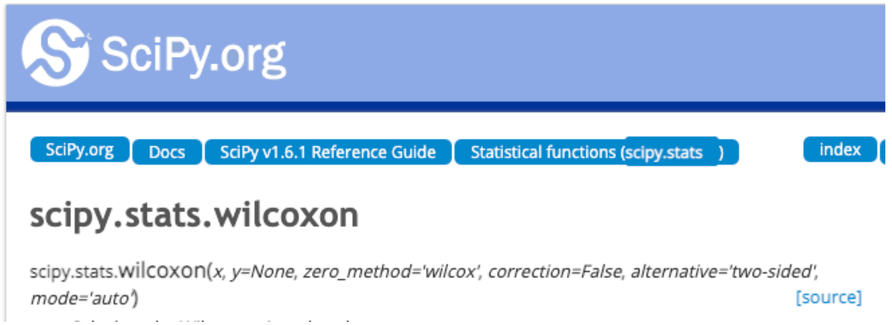
Hypothesis Testing:

H0: Median of writing score = 50
Ha: Median of writing score ≠ 50

Check your test statistic

Level of significance = p-value
If p-value < 0.05, reject NULL hypothesis

One sample median test: Wilcoxon Test



Hypothesis testing of Interval (or numerical) data: t-tests

You will see statisticians also refer to numerical data as interval data.
https://stats.idre.ucla.edu/other/mult-pkg/whatstat/what-is-the-difference-between-categorical-ordinal-and-numerical-variables/

One-sample t-test: test of whether the mean of a population has a value specified in a null hypothesis.

Hypothesis testing of Interval (or numerical) data: t-tests

You will see statisticians also refer to numerical data as interval data.
https://stats.idre.ucla.edu/other/mult-pkg/whatstat/what-is-the-difference-between-categorical-ordinal-and-numerical-variables/

One-sample t-test: test of whether the mean of a population has a value specified in a null hypothesis.

Two-sample t-test (or 2 independent sample t-test): test of the null hypothesis that the means of two populations are equal, i.e., compare differences between 2 independent group means

Eg.: A study was done to compare job stress between 2 employee groups (support staff and administrative staff). Test whether there is any difference in the job stress between the 2 employee groups?

Hypothesis testing of Interval (or numerical) data: t-tests

You will see statisticians also refer to numerical data as interval data.
https://stats.idre.ucla.edu/other/mult-pkg/whatstat/what-is-the-difference-between-categorical-ordinal-and-numerical-variables/

One-sample t-test: test of whether the mean of a population has a value specified in a null hypothesis.

Two-sample t-test (or 2 independent sample t-test): test of the null hypothesis that the means of two populations are equal, i.e., compare differences between 2 independent group means

Eg.: A study was done to compare job stress between 2 employee groups (support staff and administrative staff). Test whether there is any difference in the job stress between the 2 employee groups?

Paired t-test: test of the null hypothesis that the difference between two responses measured on the same statistical unit has a mean value of zero

Eg: A training on R-programming was conducted to improve students's performance in a data science course. Data were collected from selected sample both before and after the training program. Test the hypothesis that the training is effective in improving the students performance

Hypothesis testing of Nominal Data: Chi-Square tests

Categorical data

One-way chi-square (goodness of fit) test: test for significance in the analysis of frequency distributions of a single nominal variable

Two-way chi-square test (test of independence) - test for a relationship between two nominal variables

Data Analysis - Inferential Statistics

Chi-square test: Test whether there is a relationship between two categorical variables.

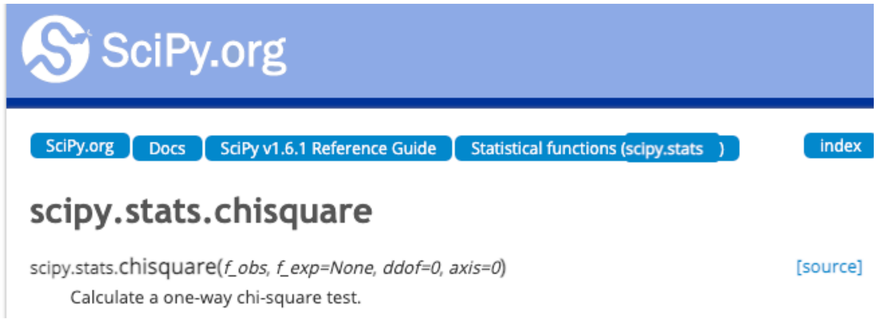
Is there a relationship between the type of school attended (schtyp) and student's gender?

Hypothesis Testing:

H0: schtype and gender are independent
Ha: schtype and gender are not independent

Check your test statistic

Level of significance = p-value
If p-value < 0.05, reject NULL hypothesis



https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1b-statistical-methods/parametric-nonparametric-tests

MANY OTHER CHEATSHEETS also available online

		Outcome variable					
		Nominal	Categorical (>2 Categories)	Ordinal	Quantitative Discrete	Quantitative Non-Normal	Quantitative Normal
Input Variable	Nominal	X^2 or Fisher's	X^2	X^2-trend or Mann - Whitney	Mann-Whitney	Mann-Whitney or log-rank[a]	Student's *t* test
	Categorical (2>categories)	X^2	X^2	Kruskal-Wallis[b]	Kruskal-Wallis[b]	Kruskal-Wallis[b]	Analysis of variance[c]
	Ordinal (Ordered categories)	X^2-trend or Mann - Whitney	e	Spearman rank	Spearman rank	Spearman rank	Spearman rank or linear regression[d]
	Quantitative Discrete	Logistic regression	e	e	Spearman rank	Spearman rank	Spearman rank or linear regression[d]
	Quantitative non-Normal	Logistic regression	e	e	e	Plot data and Pearson or Spearman rank	Plot data and Pearson or Spearman rank and linear regression
	Quantitative Normal	Logistic regression	e	e	e	Linear regression[d]	Pearson and linear regression

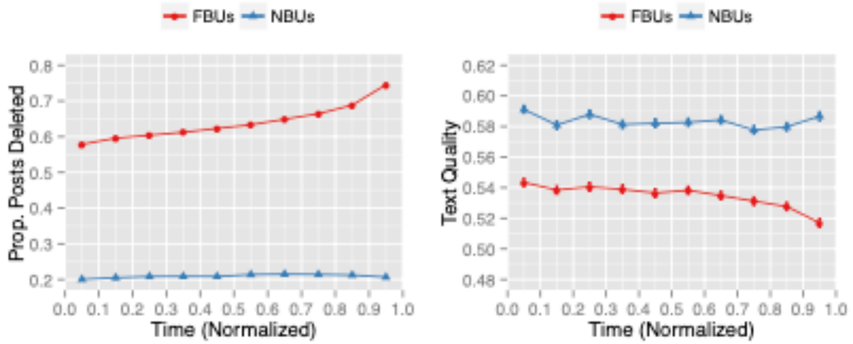
Statistical inference in the papers that you read earlier!

Antisocial Behavior in Online Discussion Communities

Justin Cheng*, **Cristian Danescu-Niculescu-Mizil**†, **Jure Leskovec***

	Mean Post Appropriateness on CNN (1-5)		
	All Posts	**First 10%**	**Last 10%**
FBUs	2.7	3.0	2.3
NBUs	3.3	3.5	3.2

Table 2: FBUs start out writing worse than NBUs and worsen more than NBUs over the course of their life. Higher appropriateness ratings correspond to higher quality posts.



(a) Post deletion rate (b) Text quality

Figure 3: (a) The rate of post deletion increases over time for FBUs, but is effectively constant for NBUs. (b) Similarly, text quality decreases over time for FBUs, but not for NBUs.

tasks (Krippendorff's α=0.35). As Table 2 shows, FBUs enter a community already writing worse than NBUs (3.0 vs. 3.5 for CNN, p<0.05 for all communities). Moreover, for both user types, post ratings also decreased with time (p<0.05), supporting H1 and previous work that showed that users in discussion communities tend to write worse over time (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2014). In fact, post ratings decreased more for FBUs than NBUs (p<0.05, d<0.19 for NBUs, d>0.29 for FBUs). In other words, while both FBUs and NBUs write worse over time, this change in quality is larger for FBUs.

Break

BACK at 9:45am

Lab

Survey

Project work