

UX Research

Lesson 10: Exploratory Data Analysis

Descriptive vs. Inferential Analysis

- **Descriptive analysis (Exploratory Data Analysis)**
 - Methods to describe data
 - Find patterns and relationships among variables
 - Cannot test any hypotheses

- **Inferential analysis**
 - Uses statistics / mathematics
 - Make claims about an unknown variable (or multiple variables) based on data
 - Can test hypotheses



Exploratory Data Analysis (EDA)

- A way to learn more about your data
 - Summarize main features
 - Show distributions of each variable (or relationships among variables)
 - Spot anomalies and discover patterns
 - Identify errors (duplicate values, null values)
- Look at data before making any assumptions
 - Helps you confirm that the analytic methods you want to use are applicable
 - Once you finish EDA, you may move on to more rigorous statistics / modeling



Exploratory Data Analysis (EDA)

Summary statistics

- Measures of “central tendency”
 - **Mean, median, mode**
 - A “typical” measurement in a group of measurements
 - A value that is representative of the group
- Measures of “dispersion”
 - **Range, standard deviation, variance**
 - Quantifies how different the measurements are from each other
 - Measures how spread out the measurements are from the c

Exploratory Data Analysis (EDA)

Summary statistics

- Measures of “central tendency”
 - **Mean, median, mode**
- Measures of “dispersion”
 - **Range, standard deviation, variance**

Age of Participants

count	mean	std	min	25%	50%	75%	max
5113.0	44.2	16.2	18.0	31.0	43.0	57.0	91.0

Exploratory Data Analysis (EDA)

Tables

Frequency Table

- Shows the frequency (counts, percents) of occurrence of each unique value in a group of observations
- Not that useful if there's lots of unique values
- Nominal or ordinal data

Cross Tabulation

- Shows the frequency of one type of observation at different values of a second type of observation
- Nominal or ordinal data
- Can be a comparison of means and SDs across different groups of a nominal category



Exploratory Data Analysis (EDA)

Tables

Frequency Table

Country	Participants
France	983
Germany	872
India	891
Italy	892
UK	621
US	854

Cross Tabulation

	Italy	UK	USA
Alone	10.54 %	9.98 %	17.10 %
Child	23.88 %	28.02 %	33.96 %
Elderly	1.68 %	2.25 %	2.34 %
Friend	0.56 %	5.15 %	2.22 %
Parent	16.03 %	26.57 %	15.34 %
Partner	63.68 %	49.28 %	56.56 %

Exploratory Data Analysis (EDA)

Plots

Bar Chart

- Frequency table in visual form
- Only useful when the measurement is discrete (e.g., ordinal, nominal)

Histogram

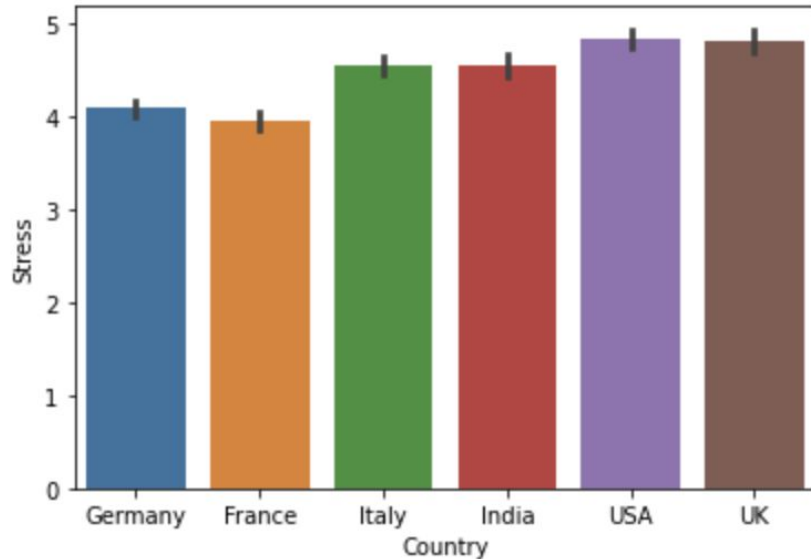
- Represents a continuous distribution
- Takes a continuous measurement, breaks it into pieces, and displays the number of counts within certain ranges of the data



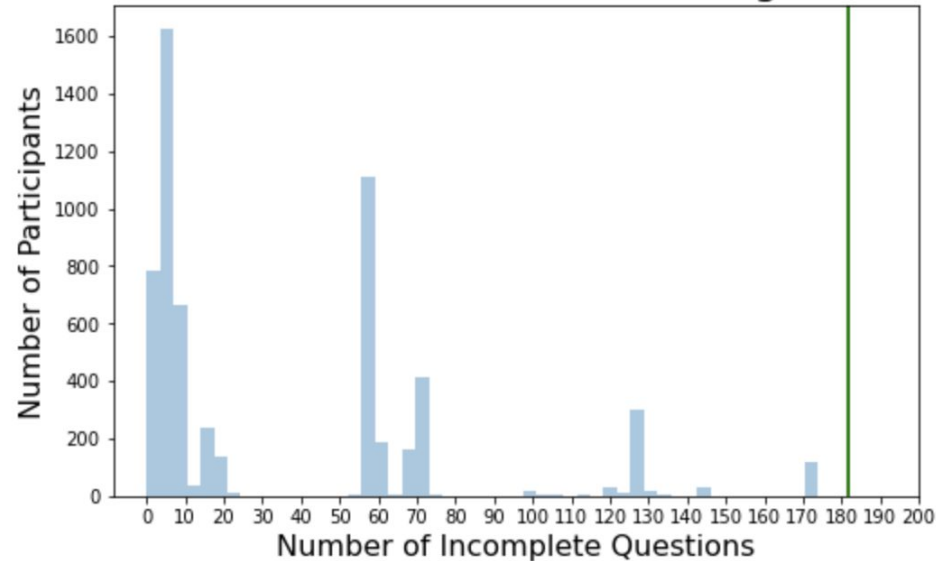
Exploratory Data Analysis (EDA)

Plots

Bar Chart



Histogram

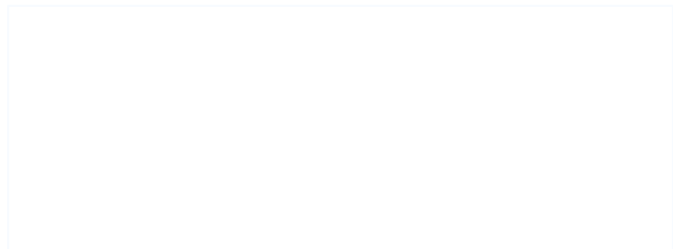


Exploratory Data Analysis (EDA)

Plots

Density plots

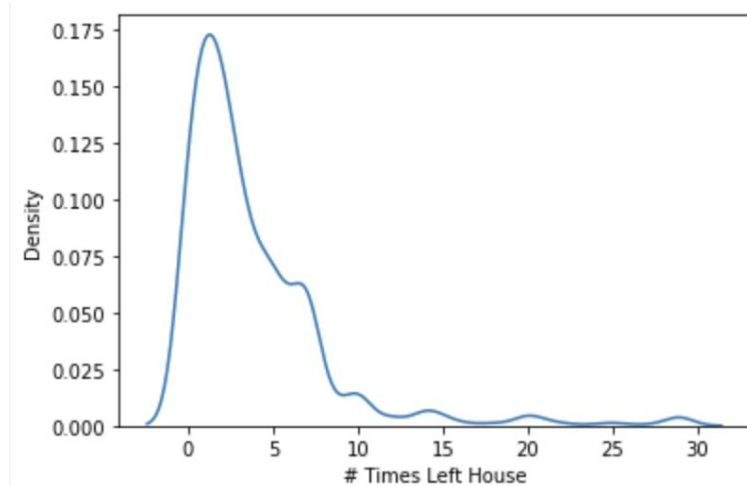
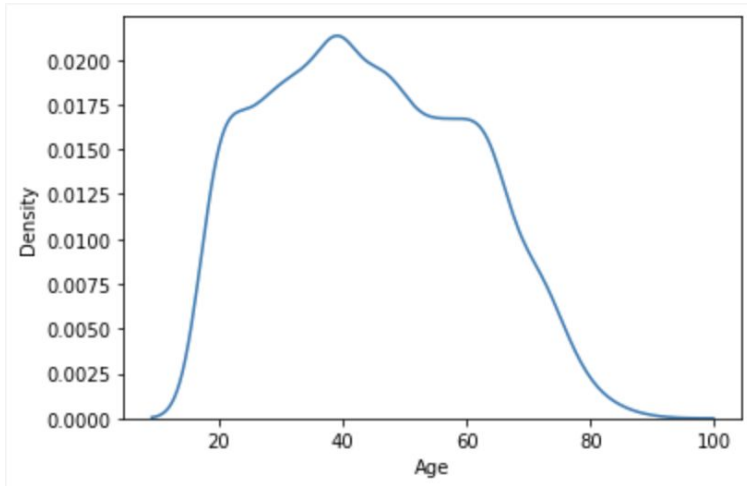
- Aka kernel density estimation (KDE) plots
- Smoothed version of a histogram
- Represents a continuous distribution



Exploratory Data Analysis (EDA)

Plots

Density plots



Exploratory Data Analysis (EDA)

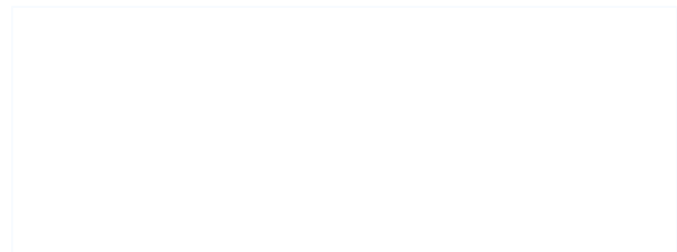
Plots

Box plots

- Represents a continuous distribution
- Displays a summary of the data: minimum value, first quartile, median, third quartile, and maximum value

Violin plots

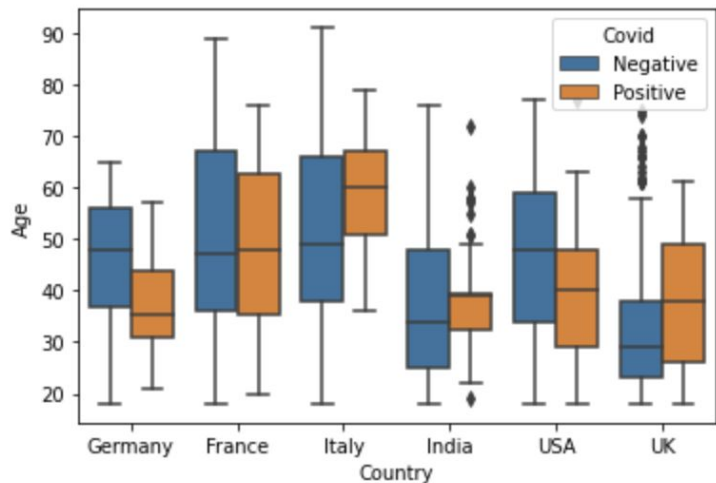
- Represents a continuous distribution
- Hybrid of KDE plot and a box plot
- Unlike a box plot, it shows the full distribution of the data



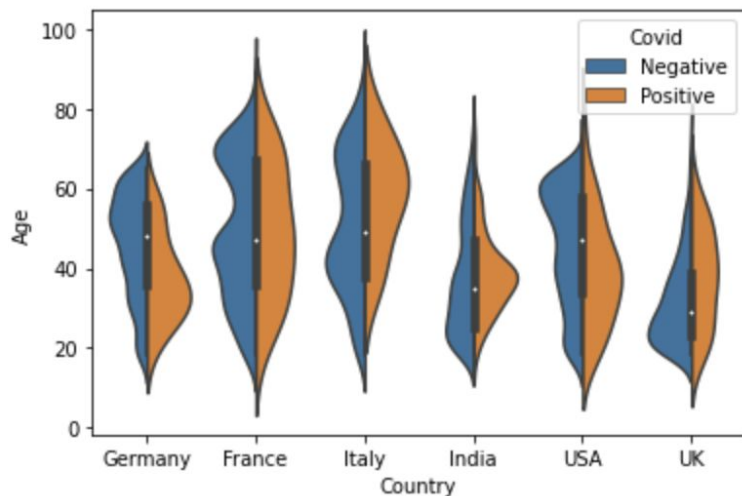
Exploratory Data Analysis (EDA)

Plots

Box plots



Violin plots



Exploratory Data Analysis (EDA)

Correlation

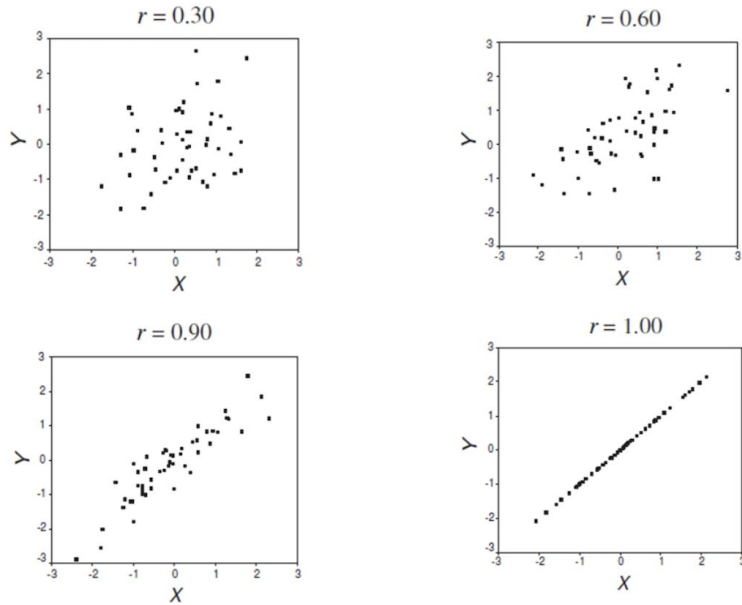
- + value = if you have a high X value, you likely have a high Y value (or low X and low Y)
- - value = if you have a high X value, you likely have a low Y value (or low X and high Y)

- The number's distance from 0 shows the strength of the association
- Closer to +1 or -1 means the association is stronger
- Closer to 0 means the two variables could be linearly independent

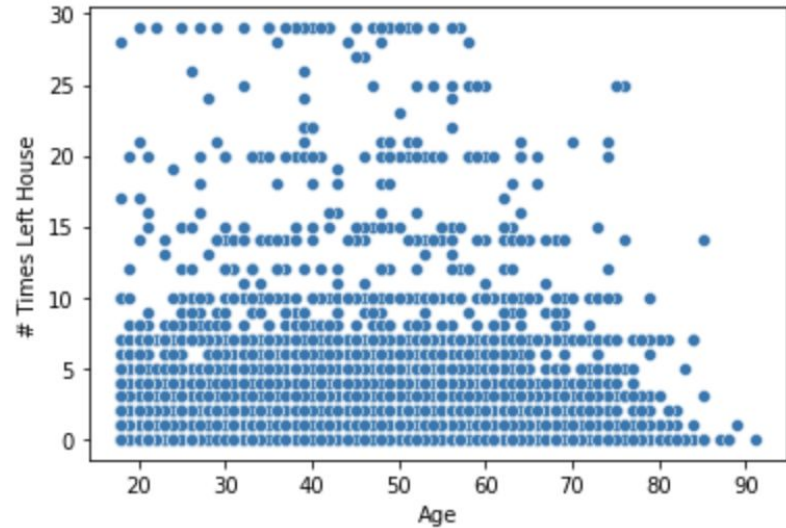
Exploratory Data Analysis (EDA)

Plots

Scatterplot (synthetic)



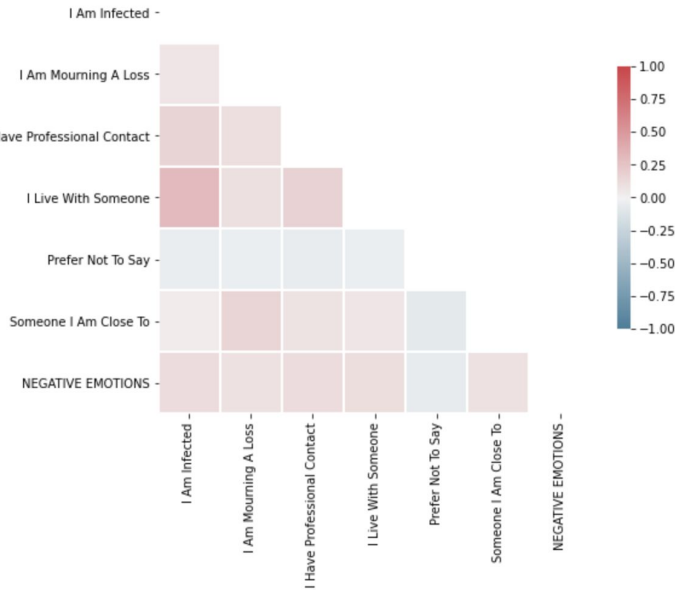
Real data ($r = 0.03$)



Exploratory Data Analysis (EDA)

Plots

Heatmap



(Corresponding correlation matrix)

	I Am Infected	I Am Mourning A Loss	I Have Professional Contact	I Live With Someone	Prefer Not To Say	Someone I Am Close To	NEGATIVE EMOTIONS
I Am Infected	1.00	0.07	0.17	0.31	-0.04	0.03	0.12
I Am Mourning A Loss	0.07	1.00	0.10	0.10	-0.04	0.16	0.09
I Have Professional Contact	0.17	0.10	1.00	0.18	-0.05	0.08	0.12
I Live With Someone	0.31	0.10	0.18	1.00	-0.04	0.06	0.11
Prefer Not To Say	-0.04	-0.04	-0.05	-0.04	1.00	-0.08	-0.06
Someone I Am Close To	0.03	0.16	0.08	0.06	-0.08	1.00	0.09
NEGATIVE EMOTIONS	0.12	0.09	0.12	0.11	-0.06	0.09	1.00

UX Research

Lesson 11: Statistical Analysis

Basic Hypothesis Testing

Hypothesis testing involves investigating how an independent variable impacts a dependent variable

- Does store location affect sales rates?
- Does price affect brand perception?
- Does customer age affect loyalty behaviors?



Basic Hypothesis Testing

We need to formulate our theory into two versions of reality that are mutually exclusive.

H_0 = “null hypothesis”

H_a = “alternative hypothesis”



Basic Hypothesis Testing

Hypothesis

You are scouting a location for a new coffee shop. You want to test the idea that coffee shops within 1 mile of campus have higher Saturday morning sales than shops more than 1 mile from campus.

H_0 : sales for shops ≤ 1 mile are ***the same as*** sales for shops > 1 mile

H_a : sales for shops ≤ 1 mile are ***greater than*** sales for shops > 1 mile



Basic Hypothesis Testing

To test our hypothesis, we want to *disconfirm the null*.

H_0 = “null hypothesis”

H_a = “alternative hypothesis”

We always start by assuming the null hypothesis is true.

- If we **do not have enough evidence** to disconfirm the null, we say the null adequately describes reality.
- If we **do have enough evidence** to disconfirm the null, we must accept the alternative.



Basic Hypothesis Testing

Hypothesis

Store A = 1 mile from campus

Store B = 2 miles from campus

H_0 : Shop A sales = Shop B sales

H_a : Shop A sales > Shop B sales

We assume that reality matches the null – there's no difference in sales based on shop distance from campus. We need enough evidence to convince us that the null hypothesis is not true.



Basic Hypothesis Testing

We collect data on Store A and B's Saturday morning coffee sales for one year.

Store A – 1 mile from campus

Week	Sales
1	514.9
2	515.0
3	514.5
4	513.8
5	513.1
6	514.2
7	517.0
8	517.3
9	514.8
10	513.5

Mean: 514.83

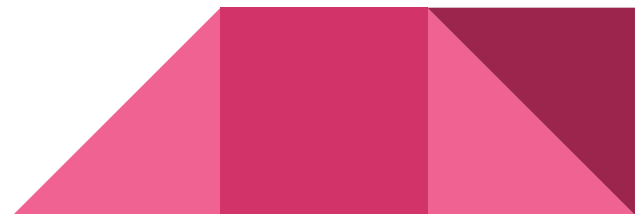
SD: 1.88

Store B – 2 miles from campus

Week	Sales
1	508.7
2	512.9
3	509.7
4	512.2
5	509.5
6	509.1
7	504.7
8	507.2
9	511.2
10	511.5

Mean: 510.02

SD: 2.71



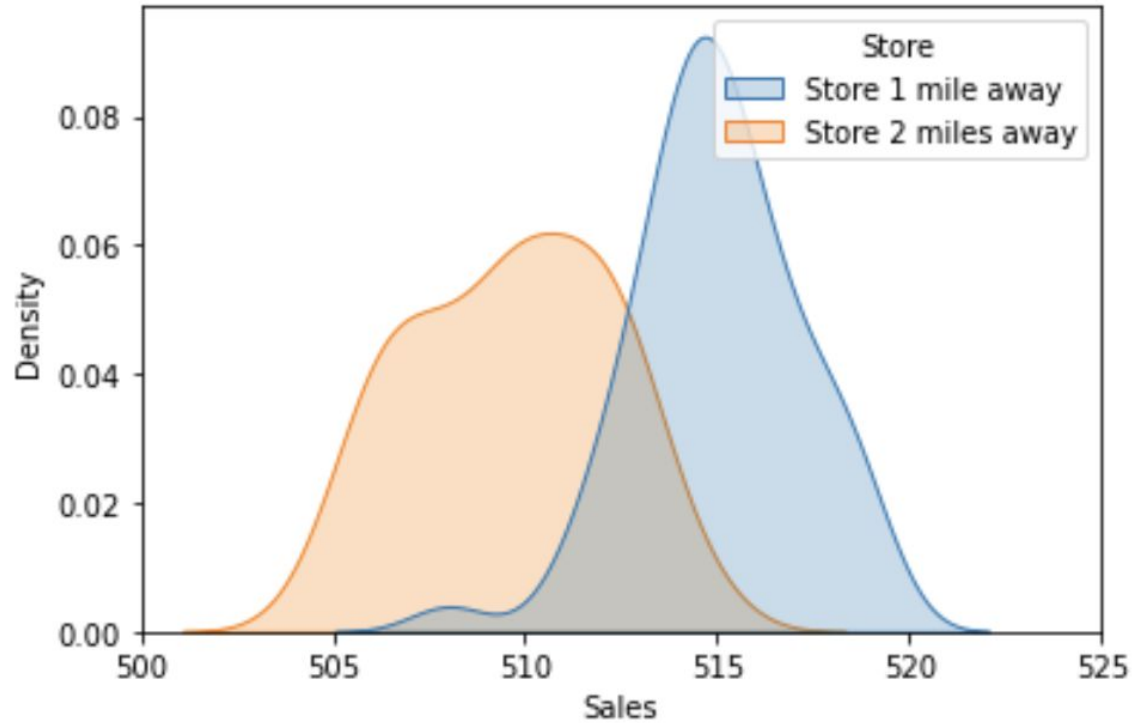
Basic Hypothesis Testing

Now, we need to look at the consistency between the sales result from our data and the sales data under the null hypothesis.

*In other words, do we have enough evidence to disprove the null hypothesis – that the sales from **Store A** and **Store B** are the same?*



Basic Hypothesis Testing




Basic Hypothesis Testing

P values

- Assumes the null hypothesis is true
- The probability that the obtained result, or a result even more discrepant from the null hypothesis than the obtained result

Typically, we reject the null hypothesis if the p-value is < 0.05 .

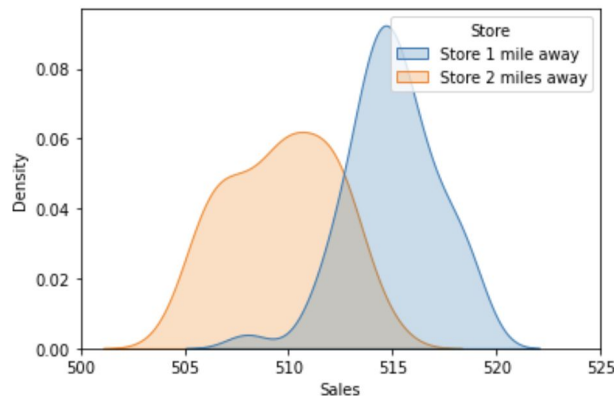
This means that, if the null hypothesis is true, there is a 5% chance of obtaining your result (or a value more extreme than the one you observed).



Basic Hypothesis Testing

T-test

- A statistical test to compare the means of two groups
- A p-value will be associated with the t-test



For our data

A (one-tailed) t-test comparing the Saturday coffee sales of **Store A** and **Store B**

- $t = 11.403$
- $p = 3.26 \text{ e-}20$ – we can reject the null hypothesis!

On average, a store 1 mile from campus sold more coffee (mean = \$514, SD = 1.88) than a store 2 miles from campus (mean = \$510, SD = 2.71), $t(51) = 11.403$, $p < 0.001$ (one-tailed).

T-test

- Uses continuous variables
- One-sample t-test
 - Compares the mean of a sample to a known mean (e.g., population mean)
 - E.g., test scores of a class vs. the national average
- Independent two-sample t-test
 - Compares means of two independent groups
 - E.g., test scores from classroom A vs classroom B
- Paired t-test
 - Compares means of two related groups (e.g., same person over time, husband and wife)
 - E.g., test scores from a specific teacher in 2023 vs 2024
- Note: degrees of freedom change the shape of the distribution and differ among these types of t-test

T-test

- Independent two-sample t-test
 - \bar{X} = sample means
 - s^2 = sample variances
 - n = sample sizes

- Paired two-sample t-test
 - \bar{D} = mean of differences between pairs (e.g., after-before; wife-husband)
 - s_D = standard deviation of the differences
 - n = number of pairs (sample size)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$t = \frac{\bar{D}}{s_D / \sqrt{n}}$$

Chi-square Test

- Tests relationships between two categorical variables
 - Variable 1 = Biological Sex → male or female
 - Variable 2 = Favorite Pet → cat, dog, rabbit

	Cat	Dog	Rabbit
Male	35	55	10
Female	42	50	8

- Want to test the independence between these two variables
 - Null hypothesis = no relationship between sex and favorite pet

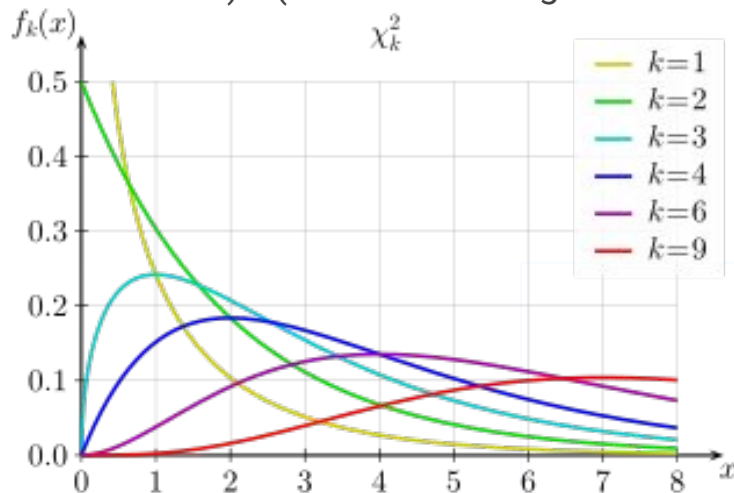
Chi-square Test

- Tests relationships between two categorical variables
- Chi-squared test statistic
 - Compares observed and expected values
 - Null = 0 (variables are independent)
 - Large value (with $p < 0.05$ means the relationship b/w variables are dependent)

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad \text{where} \quad E_i = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

Chi-square Test

- Tests relationships between two categorical variables
- Chi-squared test statistic
- Degrees of freedom define the shape of the distribution
 - (number of categories for first variable - 1) * (number of categories for second variable - 1)



Basic Hypothesis Testing

Type of dependent variable	Type of independent variable						
	Ordinal/categorical				Normal/interval (ordinal)	More than 1	None
	Two groups		More groups				
	Paired	Unpaired	Paired	Unpaired			
2 categories	McNemar Test, Sign-Test	Fisher Test, Chi-squared-Test	Cochran's Q-Test	Fisher Test, Chi-squared-test	(Conditional) Logistic Regression	Logistic Regression	Chi-squared-Test
Nominal	Bowker Test	Fisher Test, Chi-squared-Test		Fisher Test, Chi-squared-test	Multinomial logistic regression	Multinomial logistic regression	Binomial Test
Ordinal	Wilcoxon Test, Sign-Test	Wilcoxon-Mann-Whitney Test	Friedman-Test	Kruskal-Wallis Test	Spearman-rank-test	Ordered logit	Median Test
Interval	Wilcoxon Test, Sign-Test	Wilcoxon-Mann-Whitney Test	Friedman-Test	Kruskal-Wallis Test	Spearman-rank test	Multivariate linear model	Median Test
Normal	t-Test (for paired)	t-Test (for unpaired)	Linear Model (ANOVA)	Linear Model (ANOVA)	Pearson-Correlation-test	Multivariate Linear Model	t-Test
Censored Interval	Log-Rank Test		Survival Analysis, Cox proportional hazards regression				
None	Clustering, factor analysis, PCA, canonical correlation						

Basic Hypothesis Testing

Parametric assumptions:
 (1) Independent samples
 (2) Data normally distributed
 (3) Equal variances

