Data Analytics Applications

* Netflix: Personalized movie recommendations
* Facebook: News feed ranking algorithm
* OkCupid: To find suitable relationship matches between individuals
* And many more..
* [**Facebook's blog post**](https://research.fb.com/exposure-to-diverse-information-on-facebook-2/) and [**paper**](https://research.fb.com/publications/exposure-to-ideologically-diverse-information-on-facebook/) on exposure to ideologically diverse information
* [**OkCupid's blog post**](http://blog.okcupid.com/index.php/the-best-questions-for-first-dates/) on the best questions to ask on a first date
* [**Article**](http://www.dezyre.com/article/how-big-data-analysis-helped-increase-walmart-s-sales-turnover/109) on how Walmart used big data analysis to increase sales
* [**Wikipedia page**](https://en.wikipedia.org/wiki/Bill_James) on how Bill James applied data analysis to baseball
* [**Numerate's post**](http://www.numerate.com/numerates-ranking-technology-pharmaceutical-rd-gains-u-s-patent/) on using data analysis to design pharmaceutical drugs

Statistical Modelling and Analysis

* Question, Wrangle, Explore(EDA), Draw Conclusions, Communicate
* Question
  + What would
* Wrangle(To make the dataset work)
  + Gather
  + Assess
    - Df.info()
      * Missing data
      * Duplicates
      * Incorrect Data Types
  + Clean
    - Replacing missing values
    - Duplcates
      * When all the columns match
      * When one patient\_id has multiple rows – how do we choose?

Merge the rows? Select the most recently updated row?

* EDA
  + Explore and then augment our data to maximize the potential of our analyses, visualizations, and models.
  + Exploring involves finding patterns in your data, visualizing relationships in your data, and building intuition about what you’re working with.
  + After exploring, you can do things like remove outliers and create better features from your data, also known as feature engineering.
* Draw Conclusion
  + Inferential Statistics
  + Machine Learning
* Communication
  + You explore and then augment your data to maximize the potential of your analyses, visualizations, and models. Exploring involves finding patterns in your data, visualizing relationships in your data, and building intuition about what you’re working with. After exploring, you can do things like remove outliers and create better features from your data, also known as feature engineering.

Feature and Description tabular format

| **Feature** | **Description** |
| --- | --- |
| datetime | hourly date + timestamp |
| season | 1 = spring, 2 = summer, 3 = fall, 4 = winter |
| holiday | whether the day is considered a holiday |
| workingday | whether the day is neither a weekend nor holiday |
| weather \* | 1, 2, 3, 4 (see descriptions below) |
| temp | temperature in Celsius |
| atemp | "feels like" temperature in Celsius |
| humidity | relative humidity |
| windspeed | wind speed |
| casual | number of non-registered user rentals initiated |
| registered | number of registered user rentals initiated |
| count | number of total rentals |

**\* Keys for Weather Feature**

1 = clear, few clouds, partly cloudy, partly cloudy  
2 = mist + cloudy, mist + broken clouds, mist + few clouds, mist  
3 = light snow, light rain + thunderstorm + scattered clouds, light rain + scattered clouds  
4 = heavy rain + ice pallets + thunderstorm + mist, snow + fog

<http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)>

Ques

How do we know if a tumor is cancerous?

Does the size of the tumor determine its malignancy?

| **Feature** | **Description** |
| --- | --- |
| Radius | Mean of distances from center to points on the perimeter |
| Texture | Standard deviation of gray-scale values |
| Perimeter |  |
| Area |  |
| Smoothness | Local variation in radius lengths |
| Compactness | Perimeter2 / Area - 1.0 |
| Concavity | Severity of concave portions of the contour |
| Concave Points | Number of concave portions of the contour |
| Symmetry |  |
| Fractal Dimension | "Coastline approximation" - 1 |

<https://archive.ics.uci.edu/ml/datasets/Census+Income>

What is the relationship between temperature and electrical output?

Does the humidity variable appear to be normally distributed?

Which variable appears to have the most outliers?

<http://archive.ics.uci.edu/ml/datasets/Wine+Quality>

Which of the following questions would be relevant to this dataset?

* What chemical characteristics are most important in predicting the quality of wine?
* Is a certain type of wine (red or white) associated with higher quality?
* Do wines with higher alcoholic content receive better ratings?
* Do sweeter wines (more residual sugar) receive better ratings?
* What level of acidity is associated with the highest quality?
* number of samples in each dataset
* number of columns in each dataset
* features with missing values
* duplicate rows in the white wine dataset
* number of unique values for quality in each dataset
* mean density of the red wine dataset

Based on histograms of columns in this dataset, which feature variables appear skewed to the right?

* Fixed Acidity
* Total Sulfur Dioxide
* pH
* Alcohol

SUBMIT

### QUESTION 2 OF 2

Based on scatterplots of quality against different feature variables, which of the following is most likely to have a positive impact on quality?

* Volatile Acidity
* Residual Sugar
* pH
* Alcohol

Group by

Mean for each quality rating

# Drawing Conclusions Using Groupby

In the notebook below, you're going to investigate two questions about this data using pandas' [**groupby**](https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.groupby.html) function. Here are tips for answering each question:

### Q1: Is a certain type of wine (red or white) associated with higher quality?

For this question, compare the average quality of red wine with the average quality of white wine with groupby. To do this group by color and then find the mean quality of each group.

### Q2: What level of acidity (pH value) receives the highest average rating?

This question is more tricky because unlike color, which has clear categories you can group by (red and white) pH is a quantitative variable without clear categories. However, there is a simple fix to this. You can create a categorical variable from a quantitative variable by creating your own categories. [**pandas' cut**](https://pandas.pydata.org/pandas-docs/stable/generated/pandas.cut.html) function let's you "cut" data in groups. Using this, create a new column called acidity\_levels with these categories:

#### Acidity Levels:

1. High: Lowest 25% of pH values
2. Moderately High: 25% - 50% of pH values
3. Medium: 50% - 75% of pH values
4. Low: 75% - max pH value

Here, the data is being split at the 25th, 50th, and 75th percentile. Remember, you can get these numbers with pandas' describe()! After you create these four categories, you'll be able to use groupby to get the mean quality rating for each acidity level.

Is the mean quality of red wine greater than, less than, or equal to that of white wine?

* Greater
* Less
* Equal

### QUESTION 2 OF 2

What level of acidity receives the highest mean quality rating?

* High
* Moderately High
* Medium
* Low

### Q1: Do wines with higher alcoholic content receive better ratings?

To answer this question, use query to create two groups of wine samples:

1. Low alcohol (samples with an alcohol content less than the median)
2. High alcohol (samples with an alcohol content greater than or equal to the median)

Then, find the mean quality rating of each group.

### Q2: Do sweeter wines (more residual sugar) receive better ratings?

Similarly, use the median to split the samples into two groups by residual sugar and find the mean quality rating of each group.

* [**EPA Fuel Economy Testing**](https://www.epa.gov/compliance-and-fuel-economy-data/data-cars-used-testing-fuel-economy)
* [**DOE Fuel Economy Data**](http://www.fueleconomy.gov/feg/download.shtml/)

| **Attribute** | **Description** |
| --- | --- |
| Model | Vehicle make and model |
| Displ | Engine displacement - the size of an engine in liters |
| Cyl | The number of cylinders in a particular engine |
| Trans | Transmission Type and Number of Gears |
| Drive | Drive axle type (2WD = 2-wheel drive, 4WD = 4-wheel/all-wheel drive) |
| Fuel | Fuel Type |
| Cert Region\* | Certification Region Code |
| Sales Area\*\* | Certification Region Code |
| Stnd | Vehicle emissions standard code |
| Stnd Description\* | Vehicle emissions standard description |
| Underhood ID | This is a 12-digit ID number that can be found on the underhood emission label of every vehicle. It's required by the EPA to designate its "test group" or "engine family." This is explained more [**here**](https://www.epa.gov/vehicle-and-engine-certification/information-about-family-naming-conventions-vehicles-and-engines) |
| Veh Class | EPA Vehicle Class |
| Air Pollution Score | Air pollution score (smog rating) |
| City MPG | Estimated city mpg (miles/gallon) |
| Hwy MPG | Estimated highway mpg (miles/gallon) |
| Cmb MPG | Estimated combined mpg (miles/gallon) |
| Greenhouse Gas Score | Greenhouse gas rating |
| SmartWay | Yes, No, or Elite |
| Comb CO2\* | Combined city/highway CO2 tailpipe emissions in grams per mile |

\* Not included in 2008 dataset  
\*\* Not included in 2018 dataset

Which of the following are relevant questions we could ask about this data?

* Are SmartWay vehicles more expensive?
* Are more models using alternative sources of fuel? By how much?
* How much have vehicle classes improved in fuel economy?
* What are the characteristics of SmartWay vehicles?
* What features are associated with better fuel economy?
* For all of the models that were produced in 2008 that are still being produced in 2018, how much has the mpg improved and which vehicle improved the most?
* number of samples in each dataset
* number of columns in each dataset
* duplicate rows in each dataset
* datatypes of columns
* features with missing values
* number of non-null unique values for features in each dataset
* what those unique values are and counts for each

# Filter, Drop Nulls, Dedupe

## Fix cyl datatype

* 2008: extract int from string.
* 2018: convert float to int.

## Fix air\_pollution\_score datatype

* 2008: convert string to float.
* 2018: convert int to float.

## Fix city\_mpg, hwy\_mpg, cmb\_mpg datatypes

* 2008 and 2018: convert string to float.

## Fix greenhouse\_gas\_score datatype

* 2008: convert from float to int.

Skewness

Compare the distributions of greenhouse gas score in 2008 and 2018.

Correlation : +ve -ve or 0

correlation between displacement and combined mpg.

correlation between greenhouse gas score and combined mpg.

**Q1:** Are more unique models using alternative fuels in 2018 compared to 2008? By how much?  
**Q2:** How much have vehicle classes improved in fuel economy (increased in mpg)?  
**Q3:** What are the characteristics of SmartWay vehicles? Have they changed over time? (mpg, greenhouse gas)  
**Q4:** What features are associated with better fuel economy (mpg)?

### 1. Rename 2008 columns to distinguish from 2018 columns after the merge

To do this, use pandas' rename() with a lambda function. See example [**here**](https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.rename.html). In the lambda function, take the first 10 characters of the column label and and concatenate it with \_2008. (Only take the first 10 characters to prevent really long column names.)

The lambda function should look something like this: lambda x: x[:10] + "\_2008"

In your rename, don't forget to specify the parameter columns= when you add the lambda function!

### 2. Perform inner merge

To answer the last question, we are only interested in how the same model of car has been updated and how the new model's mpg compares to the old model's mpg.

Perform an inner merge with the left on model\_2008 and the right on model. See documentation for pandas' merge [**here**](https://pandas.pydata.org/pandas-docs/stable/merging.html#database-style-dataframe-joining-merging).

**Q5: For all of the models that were produced in 2008 that are still being produced now, how much has the mpg improved and which vehicle improved the most?**

Here are the steps for answering this question.

### 1. Create a new dataframe, model\_mpg, that contain the mean combined mpg values in 2008 and 2018 for each unique model

To do this, group by model and find the mean cmb\_mpg\_2008 and mean cmb\_mpgfor each.

### 2. Create a new column, mpg\_change, with the change in mpg

Subtract the mean mpg in 2008 from that in 2018 to get the change in mpg

### 3. Find the vehicle that improved the most

Find the max mpg change, and then use query or indexing to see what model it is!

<https://s3.amazonaws.com/video.udacity-data.com/topher/2018/July/5b57919a_data-set-options/data-set-options.pdf>

**Quantitative**

**Numbers (which we can perform operations upon)**

Categorical

Not numbers

**Categorical Ordinal** data take on a ranked ordering (like a ranked interaction on a scale from Very Poor to Very Good with the dogs).

**Categorical Nominal** data do not have an order or ranking (like the breeds of the dog).

Analyzing Quantitative data

Measure of Center

Mean

Median

Mode

Measure of Spread

Range

Interquartile Range

Standard Deviation

Variance

The shape of data

Left Skewed ( Mean>Median>Mode) |--------------|--|-----|

Right Skewed ( Mode<Median<mean) |-----|--|------------|

Symmetric ( Mean = Median = Mode)

Outliers

Typos – remove/fix

Anamoly Detection

The five number summary consist of 5 values:

Minimum: The smallest number in the dataset.

Q1Q1​: The value such that 25% of the data fall below.

Q2Q2​: The value such that 50% of the data fall below.

Q3Q3​: The value such that 75% of the data fall below.

Maximum: The largest value in the dataset.

| **Shape** | **Mean vs. Median** | **Real World Applications** |
| --- | --- | --- |
| Symmetric (Normal) | Mean equals Median | Height, Weight, Errors, Precipitation |
| Right-skewed | Mean greater than Median | Amount of drug remaining in a blood stream, Time between phone calls at a call center, Time until light bulb dies |
| Left-skewed | Mean less than Median | Grades as a percentage in many universities, Age of death, Asset price changes |

* [**Quora**](https://www.quora.com/What-are-some-real-world-examples-of-normally-distributed-quantities)
* [**University of Texas**](https://www.utdallas.edu/~scniu/OPRE-6301/documents/Important_Probability_Distributions.pdf)
* [**Stack Exchange**](https://stats.stackexchange.com/questions/89179/real-life-examples-of-distributions-with-negative-skewness)

When outliers are present we should consider the following points.

**1.** Noting they exist and the impact on summary statistics.

**2.** If typo - remove or fix

**3.** Understanding why they exist, and the impact on questions we are trying to answer about our data.

**4.** Reporting the 5 number summary values is often a better indication than measures like the mean and standard deviation when we have outliers.

**5.** Be careful in reporting. Know how to ask the right questions.

Below are my guidelines for working with any column (random variable) in your dataset.

**1.** Plot your data to identify if you have outliers.

**2.** Handle outliers accordingly via the methods above.

**3.** If no outliers and your data follow a normal distribution - use the mean and standard deviation to describe your dataset, and report that the data are normally distributed.

If you aren't sure if your data are normally distributed, there are plots called [**normal quantile plots**](http://data.library.virginia.edu/understanding-q-q-plots/) and statistical methods like the [**Kolmogorov-Smirnov test**](https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test) that are aimed to help you understand whether or not your data are normally distributed.

## Descriptive Statistics

Descriptive statistics **is about describing our collected data** using the measures discussed throughout this lesson: measures of center, measures of spread, shape of our distribution, and outliers. We can also use plots of our data to gain a better understanding.

## Inferential Statistics

Inferential Statistics **is about using our collected data to draw conclusions to a larger population**. Performing inferential statistics well requires that we take a sample that accurately represents our population of interest.

A common way to collect data is via a survey. However, surveys may be extremely biased depending on the types of questions that are asked, and the way the questions are asked. This is a topic you should think about when tackling the first project.

1. **Population** - our entire group of interest.
2. **Parameter** - numeric summary about a population
3. **Sample** - subset of the population
4. **Statistic** numeric summary about a sample

Simpson Paradox

A screenshot of a cell phone

Description automatically generated

Yes – Male

A close up of text on a white background

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Parameter – are fixed

Statistic – change according tot he sample selected

Two important mathematical theorems for working with sampling distributions include:

1. **Law of Large Numbers**
2. **Central Limit Theorem**

The **Law of Large Numbers** says that **as our sample size increases, the sample mean gets closer to the population mean**, but how did we determine that the sample mean would estimate a population mean in the first place? How would we identify another relationship between parameter and statistic like this in the future?

Three of the most common ways are with the following estimation techniques:

* [**Maximum Likelihood Estimation**](https://en.wikipedia.org/wiki/Maximum_likelihood_estimation)
* [Method of Moments Estimation][[**https://en.wikipedia.org/wiki/Method\_of\_moments]**](https://en.wikipedia.org/wiki/Method_of_moments%5D)
* [**Bayesian Estimation**](https://en.wikipedia.org/wiki/Bayes_estimator)

The Central Limit Theorem states that with a large enough sample size the sampling distribution of the mean will be normally distributed.

The Central Limit Theorem actually applies for these well known statistics:

1. Sample means (x¯)
2. Sample proportions (p)
3. Difference in sample means (x¯1​−x¯2​)
4. Difference in sample proportions (p1​−p2​)

Proportion : mean of 0/1 data values

Clt

Doesnt work for variance, r value etc

Instead of CLT, we can use bootstrapping to simulate sampling distributions

BS – sampling with replacement

Three of the most common ways are with the following estimation techniques for finding "good statistics" are as shown previously:

* [**Maximum Likelihood Estimation**](https://en.wikipedia.org/wiki/Maximum_likelihood_estimation)
* [**Method of Moments Estimation**](https://onlinecourses.science.psu.edu/stat414/node/193)
* [**Bayesian Estimation**](https://en.wikipedia.org/wiki/Bayes_estimator)
* You can learn more about Bradley Efron [**here**](https://en.wikipedia.org/wiki/Bradley_Efron).
* Additional notes on why bootstrapping works as a technique for inference can be found [**here**](https://stats.stackexchange.com/questions/26088/explaining-to-laypeople-why-bootstrapping-works).

#### Sampling Distributions

* **Sampling Distributions** are the distribution of a statistic (any statistic).
* There are two very important mathematical theorems that are related to sampling distributions: **The Law of Large Numbers** and **The Central Limit Theorem**.
* **The Law of Large Numbers** states that as a sample size increases, the sample mean will get closer to the population mean. In general, if our statistic is a "good" estimate of a parameter, it will approach our parameter with larger sample sizes.
* **The Central Limit Theorem** states that with large enough sample sizes our sample mean will follow a normal distribution, but it turns out this is true for more than just the sample mean.

#### Bootstrapping

* **Bootstrapping** is a technique where we sample from a group with replacement.
* We can use bootstrapping to simulate the creation of sampling distribution, which you did many times in this lesson.
* By bootstrapping and then calculating repeated values of our statistics, we can gain an understanding of the sampling distribution of our statistics.

Confidence Intervals CI

Pop parameter is fixed

We dont estimate the exact value rather we specify an interval where our statistic/estimate might fall into

We can use bootstrapping and sampling distributions to build confidence intervals for our parameters of interest.

By finding the statistic that best estimates our parameter(s) of interest (say the sample mean to estimate the population mean or the difference in sample means to estimate the difference in population means), we can easily build confidence intervals for the parameter of interest.

With 95% confidence we can say that our statistic falls in this interval (cut 2.5% from left and 2.5% from right oft he normal distribution)

Statistical Significance:

Evidence from Hyposthesis tests and CI that H1 is true

Practical Significance:

COnsiders real world aspects and not just numbers to make decision

Traditional methods vs (bootstrapping and CI which we did)

To learn more about the traditional methods, see the documentation [**here on the Stat Trek site**](http://stattrek.com/hypothesis-test/hypothesis-testing.aspx) on the corresponding hypothesis tests.

Traditional methods also provide the same results as our BS and CI

Assuming you control all other items of your analysis:

1. Increasing your sample size will decrease the width of your confidence interval.
2. Increasing your confidence level (say 95% to 99%) will increase the width of your confidence interval.

You saw that you can compute:

1. The confidence interval **width** as the difference between your upper and lower bounds of your confidence interval.
2. The **margin of error** is half the confidence interval width, and the value that you add and subtract from your sample estimate to achieve your confidence interval final results.

### Confidence Intervals (& Hypothesis Testing) vs. Machine Learning

Confidence intervals take an aggregate approach towards the conclusions made based on data, as these tests are aimed at understanding population parameters (which are aggregate population values).

Alternatively, machine learning techniques take an individual approach towards making conclusions, as they attempt to predict an outcome for each specific data point.

HT

Convert questions into Hypothesis

Use data to verify the hypothesis

Used to estimate pop paramter using sample statistic

In the hypothesis tests you build in the upcoming lessons, you will be able to choose a type I error threshold, and your hypothesis tests will be created to minimize the type II errors after ensuring the type I error rate is met.

You are always performing hypothesis tests on **population parameters**, never on statistics. Statistics are values that you already have from the data, so it does not make sense to perform hypothesis tests on these values.

Common hypothesis tests include:

1. Testing a population mean [**(One sample t-test)**](http://sites.utexas.edu/sos/guided/inferential/numeric/claim/one-sample-t/).
2. Testing the difference in means [**(Two sample t-test)**](https://www.isixsigma.com/tools-templates/hypothesis-testing/making-sense-two-sample-t-test/)
3. Testing the difference before and after some treatment on the same individual [**(Paired t-test)**](http://www.statstutor.ac.uk/resources/uploaded/paired-t-test.pdf)
4. Testing a population proportion [**(One sample z-test)**](http://stattrek.com/statistics/dictionary.aspx?definition=one-sample%20z-test)
5. Testing the difference between population proportions [**(Two sample z-test)**](https://onlinecourses.science.psu.edu/stat414/node/268)

You can use one of these sites to provide a t-table or z-table to support one of the above approaches:

* [**t-table**](https://s3.amazonaws.com/udacity-hosted-downloads/t-table.jpg)
* [**t-table or z-table**](http://www.z-table.com/t-value-table.html)

**There are literally hundreds of different hypothesis tests!** However, instead of memorizing how to perform all of these tests, you can find the statistic(s) that best estimates the parameter(s) you want to estimate, you can bootstrap to simulate the sampling distribution. Then you can use your sampling distribution to assist in choosing the appropriate hypothesis.

P-value

1. Simulate the values of your statistic that are possible from the null.
2. Calculate the value of the statistic you actually obtained in your data.
3. Compare your statistic to the values from the null.
4. Calculate the proportion of null values that are considered **extreme** based on your alternative.

The p-value is the probability of getting our statistic or a more extreme value if the null is true.

Therefore, small p-values suggest our null is not true. Rather, our statistic is likely to have come from a different distribution than the null.

When the p-value is large, we have evidence that our statistic was likely to come from the null hypothesis. Therefore, we do not have evidence to reject the null.

By comparing our p-value to our type I error threshold (α*α*), we can make our decision about which hypothesis we will choose.

pval≤α⇒*pval*≤*α*⇒ Reject H0*H*0​

pval>α⇒*pval*>*α*⇒ Fail to Reject H0*H*0​

When performing more than one hypothesis test, your type I error compounds. In order to correct for this, a common technique is called the **Bonferroni**correction. This correction is **very conservative**, but says that your new type I error rate should be the error rate you actually want divided by the number of tests you are performing.

Therefore, if you would like to hold a type I error rate of 1% for each of 20 hypothesis tests, the **Bonferroni** corrected rate would be 0.01/20 = 0.0005. This would be the new rate you should use as your comparison to the p-value for each of the 20 tests to make your decision.

### Other Techniques

Additional techniques to protect against compounding type I errors include:

1. [**Tukey correction**](http://www.itl.nist.gov/div898/handbook/prc/section4/prc471.htm)
2. [**Q-values**](http://www.nonlinear.com/support/progenesis/comet/faq/v2.0/pq-values.aspx)

A two-sided hypothesis test (that is a test involving a ≠≠ in the alternative) is the same in terms of the conclusions made as a confidence interval as long as:

1−CI=α1−*CI*=*α*

For example, a 95% confidence interval will draw the same conclusions as a hypothesis test with a type I error rate of 0.05 in terms of which hypothesis to choose, because:

1−0.95=0.051−0.95=0.05

assuming that the alternative hypothesis is a two sided test.

Video on [**effect size here**](https://www.youtube.com/watch?v=z98xODInLCQ).

<https://blog.minitab.com/blog/adventures-in-statistics-2/understanding-hypothesis-tests-confidence-intervals-and-confidence-levels>

<http://www.mit.edu/~6.s085/notes/lecture2.pdf>

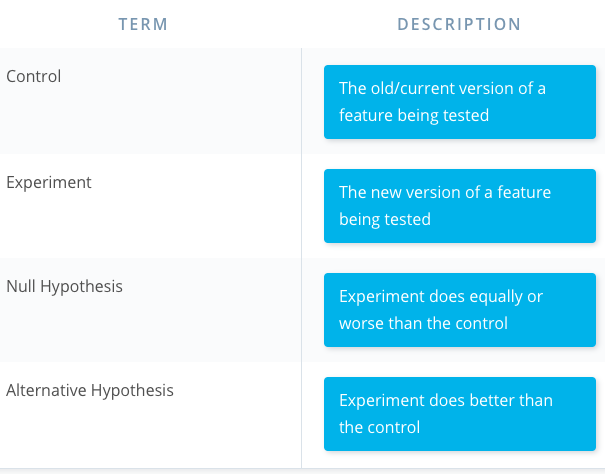
<https://www.analyticsvidhya.com/blog/2015/09/hypothesis-testing-explained/>

<https://online.stat.psu.edu/stat200/lesson/6/6.6>

<https://towardsdatascience.com/p-values-explained-by-data-scientist-f40a746cfc8>

A/B testing also has its drawbacks. It can help you compare two options, but it can't tell you about an option you haven’t considered. It can also produce bias results when tested on existing users, due to factors like change aversion and novelty effect.

* **Change Aversion:** Existing users may give an unfair advantage to the old version, simply because they are unhappy with change, even if it’s ultimately for the better.
* **Novelty Effect:** Existing users may give an unfair advantage to the new version, because they’re excited or drawn to the change, even if it isn’t any better in the long run.



The metric we will use is the click through rate for the Explore Courses button on the home page. **Click through rate (CTR)** is often defined as the the number of clicks divided by the number of views. Since Audacity uses cookies, we can identify unique users and make sure we don't count the same one multiple times. For this experiment, we'll define our click through rate as:

**CTR: # clicks by unique users / # views by unique users**

Now that we have our metric, let's set up our null and alternative hypotheses:

H0:CTRnew≤CTRold*H*0​:*CTRnew*​≤*CTRold*​

H1:CTRnew>CTRold*H*1​:*CTRnew*​>*CTRold*​

Our alternative hypothesis is what we want to prove to be true, in this case, that the new homepage design has a higher click through rate than the old homepage design. And the null hypothesis is what we assume to be true before analyzing data, which is that the new homepage design has a click through rate that is less than or equal to that of the old homepage design. As you’ve seen before, we can rearrange our hypotheses to look like this:

H0:CTRnew−CTRold≤0*H*0​:*CTRnew*​−*CTRold*​≤0  
H1:CTRnew−CTRold>0*H*1​:*CTRnew*​−*CTRold*​>0

1. We computed the **observed difference** between the metric, click through rate, for the control and experiment group.
2. We simulated the **sampling distribution** for the difference in proportions (or difference in click through rates).
3. We used this sampling distribution to simulate the **distribution under the null** hypothesis, by creating a random normal distribution centered at 0 with the same spread and size.
4. We computed the **p-value** by finding the proportion of values in the null distribution that were greater than our observed difference.
5. We used this p-value to determine the **statistical significance** of our observed difference.

<https://en.wikipedia.org/wiki/Multiple_comparisons_problem>

Since the Bonferroni method is too conservative when we expect correlation among metrics, we can better approach this problem with more sophisticated methods, such as the [**closed testing procedure**](http://en.wikipedia.org/wiki/Closed_testing_procedure), [**Boole-Bonferroni bound**](http://en.wikipedia.org/wiki/Bonferroni_bound), and the [**Holm-Bonferroni method**](http://en.wikipedia.org/wiki/Holm%E2%80%93Bonferroni_method). These are less conservative and take this correlation into account.

If you do choose to use a less conservative method, just make sure the assumptions of that method are truly met in your situation, and that you're not just trying to [**cheat on a p-value**](http://freakonometrics.hypotheses.org/19817). Choosing a poorly suited test just to get significant results will only lead to misguided decisions that harm your company's performance in the long run.

# Difficulties in A/B Testing

As you saw in the scenarios above, there are many factors to consider when designing an A/B test and drawing conclusions based on its results. To conclude, here are some common ones to consider.

* Novelty effect and change aversion when existing users first experience a change
* Sufficient traffic and conversions to have significant and repeatable results
* Best metric choice for making the ultimate decision (eg. measuring revenue vs. clicks)
* Long enough run time for the experiment to account for changes in behavior based on time of day/week or seasonal events.
* Practical significance of a conversion rate (the cost of launching a new feature vs. the gain from the increase in conversion)
* Consistency among test subjects in the control and experiment group (imbalance in the population represented in each group can lead to situations like [**Simpson's Paradox**](https://en.wikipedia.org/wiki/Simpson%27s_paradox))

Khan’s academy