

# COMP5310: Principles of Data Science

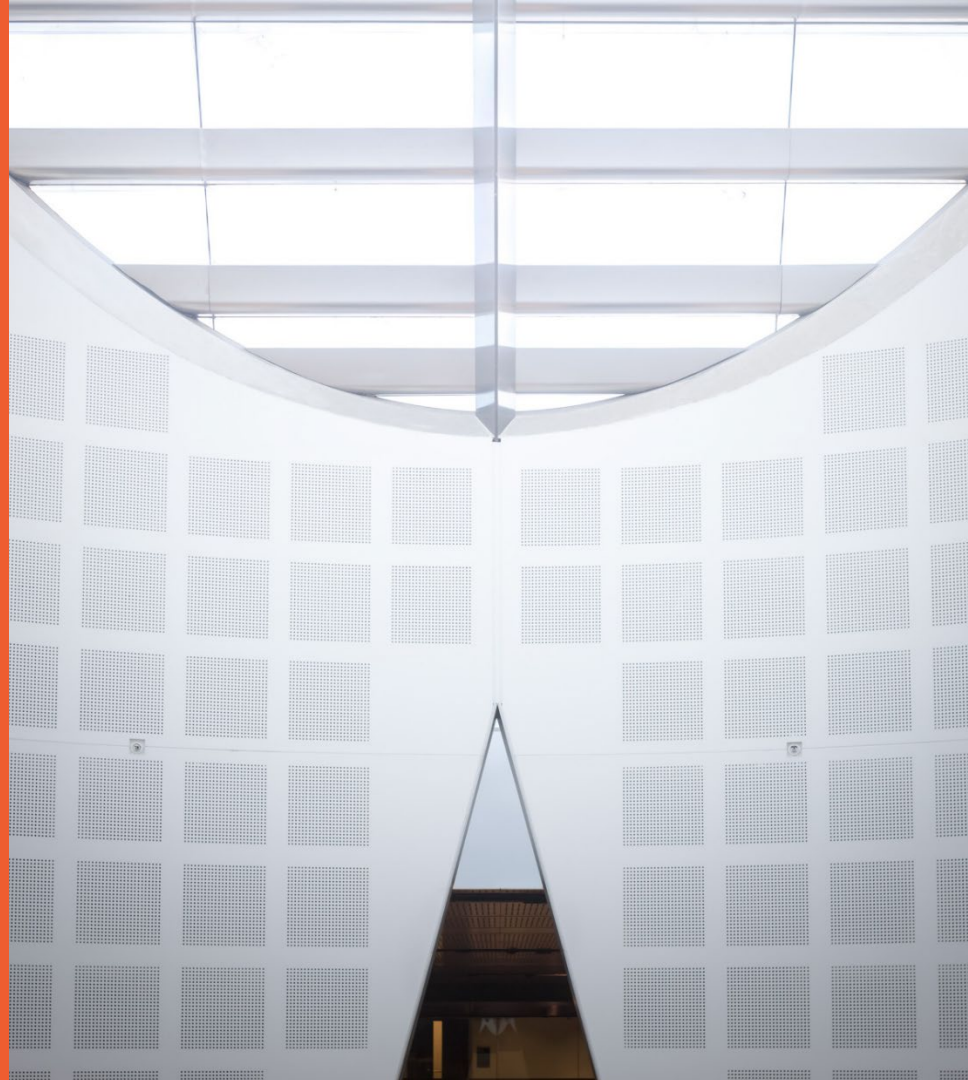
## W6: Hypothesis Testing and Evaluation

**Presented by**

Maryam Khanian

School of Computer Science

Based on slides by previous lecturers of this unit of study



# Last week: Querying and summarising data

## Objective

- To be able to extract a data set from a database, as well as to leverage on the SQL capabilities for in-database data summarisation and analysis.

## Lecture

- Data gathering reprise.
- SQL querying.
- Summarising data with SQL.
- Statistic functions support in SQL.

## Readings

- Data Science from Scratch: Ch 24.

## Exercises

- Data Loading.
- SQL Querying.
- Python DB Querying.
- Data Summarization using SQL.

## TO-DO in W5

- Finish Ed Lessons Python modules.
- Finish Ed Lessons SQL modules.

# Goal of today's lecture

- High-level overview of statistical tests (not a deep dive)
- Provide some guidance on selecting appropriate statistical tests for evaluating a predictive model, and justifying the choice of tool, in Assignment Stage 2A
- Help you seek details of how to use a statistical method or tool in the data analytic process

# TYPES OF STATISTICAL STUDIES

# Types of statistical studies

## Observational Study

- Simply observing what happens.
- Records information about subjects without applying any treatments to subjects (passive participation of researcher).

## Experimental Study

- Records information about subjects while applying treatments to subjects and controlling study conditions to some degree (active participation of researcher).

# Observational studies

## Sample survey

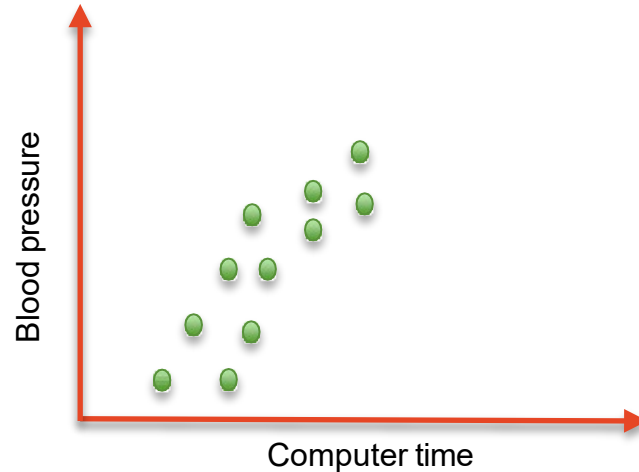
- Provide information about a population based on a sample at a specific point in time.
- Only establish correlation not causality!
  - **Study 1:** Tanning and Skin Cancer.
    - The observational study involved 1,500 people.
    - Selected a group of people who had skin cancer and another group of people who did not have skin cancer.
    - Asked all participants whether they used tanning beds.
    - Wanted to see if there was an association between tanning beds and skin cancer prevalence.

Obs

# Observational studies

ob

- **Study 2:** Average Computer Time vs Blood Pressure.
  - Enrol 100 individuals in the observational study.
  - Ask them about the average computer time they spent each day.
  - Measure their blood pressure.



# Experimental studies

- Strong hypotheses, sample size for desired power and controlled data collection per specified protocols.
- Establish causality.
- **Example:** randomized control trials.
  - 100 subjects.
  - Factor: Average Computer Time.
  - Treatments:
    - Control group (computer time: max. 30 minutes).
    - Treatment group (computer time: 2 hours).
  - 50 subjects randomly assigned to each treatment.
  - Response: we measure the blood pressure for each group.



# which statistical study looks more suitable to apply in your assignment



# Experimental vs observational

- Main difference between observational studies and experiments
  - Most experiments use random assignment while observational studies do not.
- Observational studies typically only establish correlation but not causality
- Experimental studies establish causality

# STATISTICAL SIGNIFICANCE TESTING

# Types of variables

## Dependent variable

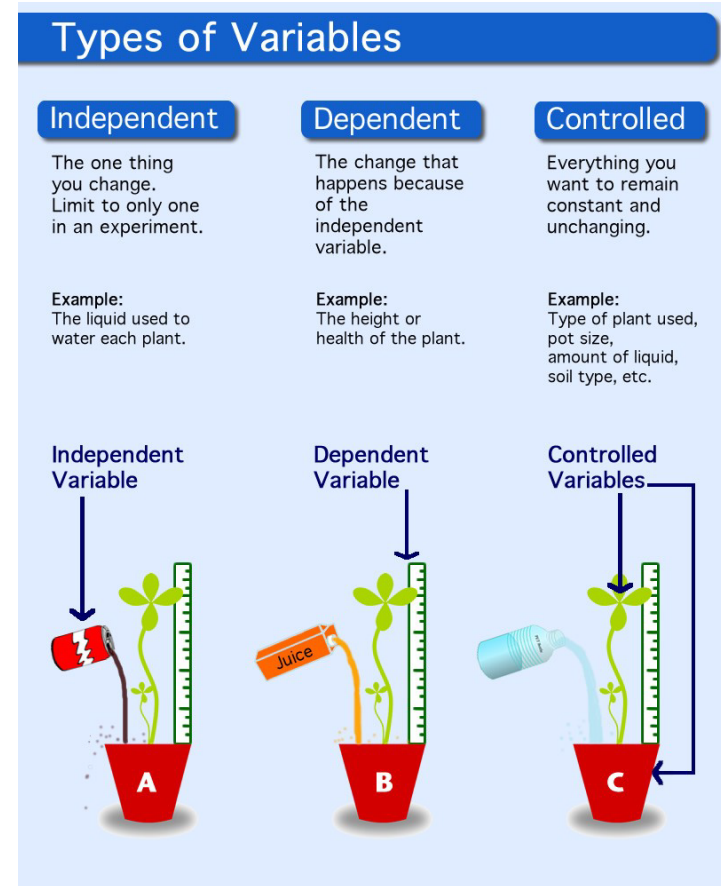
- Measure of interest.

## Independent variable

- Manipulated to observe the effect on dependent variable

## Controlled variables

- Materials, measurements and methods that don't change.



# Research question

## Research question (Q)

- Asks whether the independent variable has an effect.
- “If there is a change in the independent variable, will there also be a change in the dependent variable?”

## Null hypothesis ( $H_0$ )

- The assumption that there is no effect.
- “There is no change in the dependent variable when the independent variable changes.”

# Hypothesis testing

- We use hypothesis testing to specify whether to accept or reject a claim about a population depending on the evidence provided by a sample of data.
- A hypothesis test examines two opposing hypotheses about a population parameter (e.g. the mean):
  - The **null hypothesis** and the **alternative hypothesis**.
  - The null hypothesis represents our **initial assumption** about the parameter, and we **collect evidence** to possibly **reject the null hypothesis** in favour of the **alternative hypothesis**.
- **Example:** Determine whether the mean of a population differs significantly (this has a special meaning) from a specific value or from the mean of another population.

# Testing reliability with p-values

- Most tests calculate a p-value for measuring observation extremity, to measure whether or not the Null Hypothesis ( $H_0$ ) is correct.
- Compare to significance level threshold  $\alpha$ .
  - $\alpha$  is the probability of (wrongly) rejecting  $H_0$  given that it is true (Type I error rate, i.e., false positive).
  - Commonly use  $\alpha$  of 5% or 1%.

Actual Condition

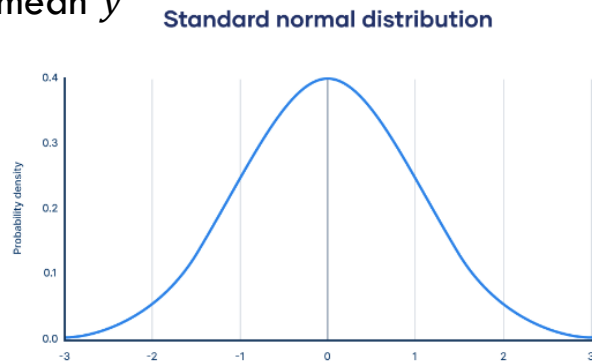
Test results

		Accept $H_0$	Reject $H_0$
Actual Condition	$H_0$ ( $H_0$ is True) - No difference	Right Decision	Type I error
	$H_1$ ( $H_0$ is False) - Difference exists	Type II error	Right Decision

P-value	Indicates	Reject $H_0$ ?
$<\alpha$	Strong evidence against the null hypothesis	Yes
$>\alpha$	Weak evidence against the null hypothesis	No
$=\alpha$	Marginal	NA

# General idea

- Suppose we have two normal random variables  $X$  and  $Y$  with the same known variance, we want to test whether they have the same mean
  - Sample 100 random numbers from  $X$ , and calculate its mean  $\bar{x}$
  - Sample 100 random numbers from  $Y$ , and calculate its mean  $\bar{y}$
  - Let  $\mu_x$  and  $\mu_y$  be the mean of  $X$  and  $Y$ , respectively
  - How do we conclude whether  $\mu_x = \mu_y$  from the value of  $\bar{x} - \bar{y}$
- **null hypothesis:**  $\mu_x = \mu_y$
- **alternative hypothesis:**  $\mu_x \neq \mu_y$
- P-value: probability of generating two sets of samples of 100 each such that the difference between their empirical means is at least  $|\bar{x} - \bar{y}|$ , under the assumption that the null hypothesis is true



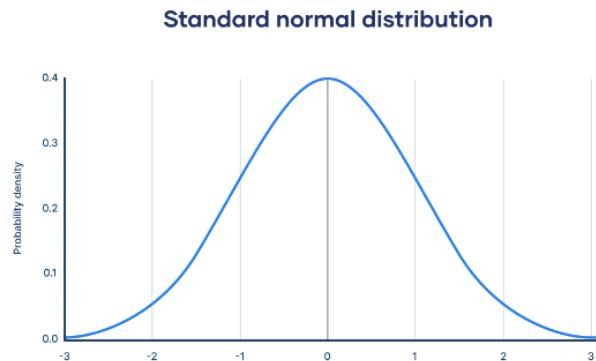


# Not every test result is correct

- $\alpha=0.05$  will erroneously reject  $H_0$  5% of the time
- Perform enough tests and you will get a false result (p-hacking)
- Good science:
  - Determine hypotheses before looking at data
  - Perform hypothesis-agnostic data cleaning
  - Remember that p-values do not replace common sense
- <https://sites.uw.edu/stlab/2016/03/09/the-arbitrary-magic-of-p-0-05/>

# One-side test

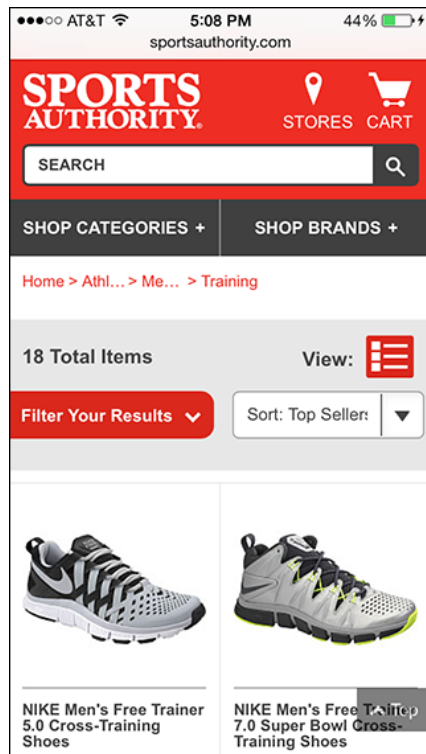
- Suppose we want to test whether  $\mu_x > \mu_y$ 
  - **null hypothesis:**  $\mu_x = \mu_y$
  - **alternative hypothesis:**  $\mu_x > \mu_y$
  - P-value: probability of generating two sets of samples of 100 each such that the difference between their empirical means is at least  $\bar{x} - \bar{y}$ , under the assumption that the null hypothesis is true
- P-value (in general):
  - $P(\text{observed or more extreme outcome} \mid H_0 \text{ true})$



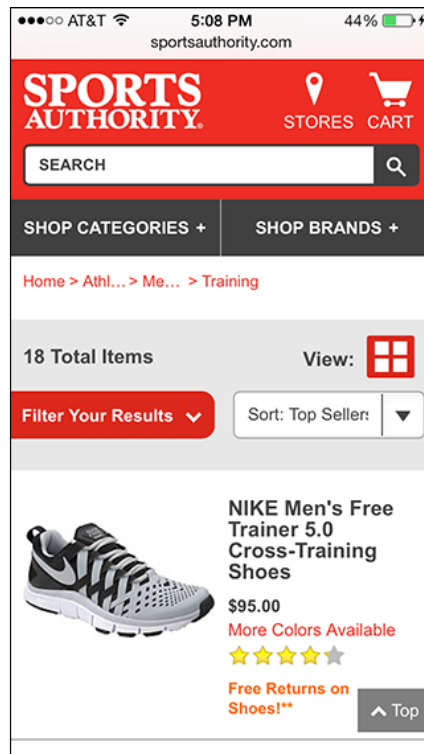
# TESTING WHICH APPROACH IS BETTER BETWEEN SUBJECTS

# Scenario: Comparing visual layouts

Grid view



List view



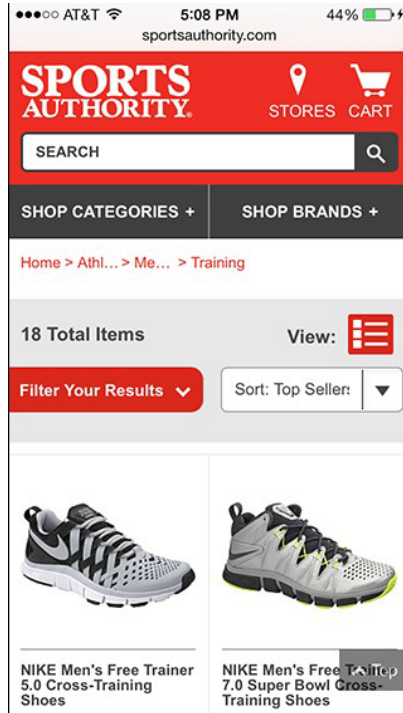
<https://www.nngroup.com/articles/image-vs-list-mobile-navigation/>

## Research question

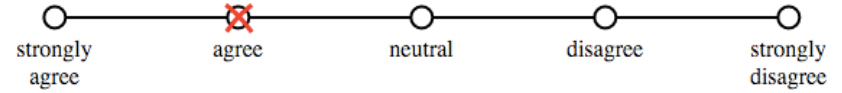
**Do users prefer grid view?**

# Data/Measurement: User ratings of layouts

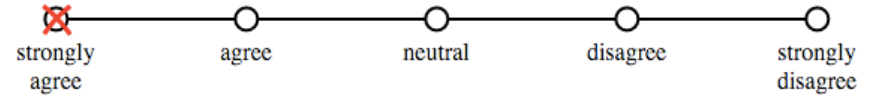
Example response  
from User Group A



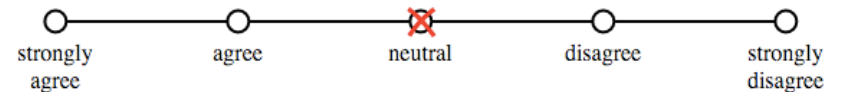
Page is easy to use.



Page gives good overview.

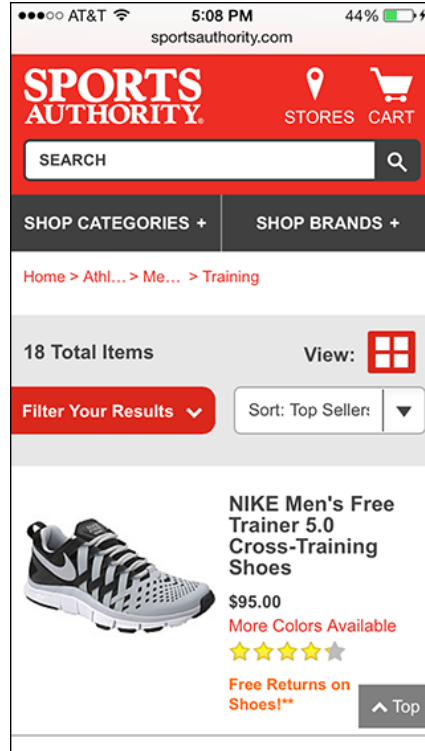


Page gives sufficient detail.

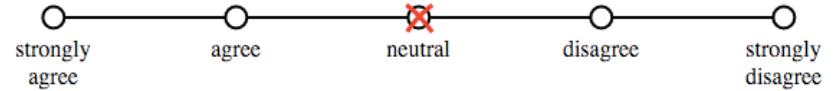


# Data/Measurement: User ratings of layouts

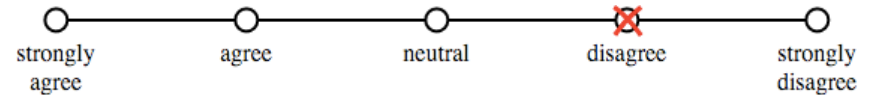
Example response  
from User Group B



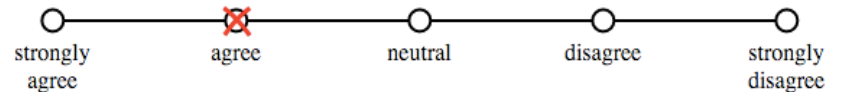
Page is easy to use.



Page gives good overview.



Page gives sufficient detail.

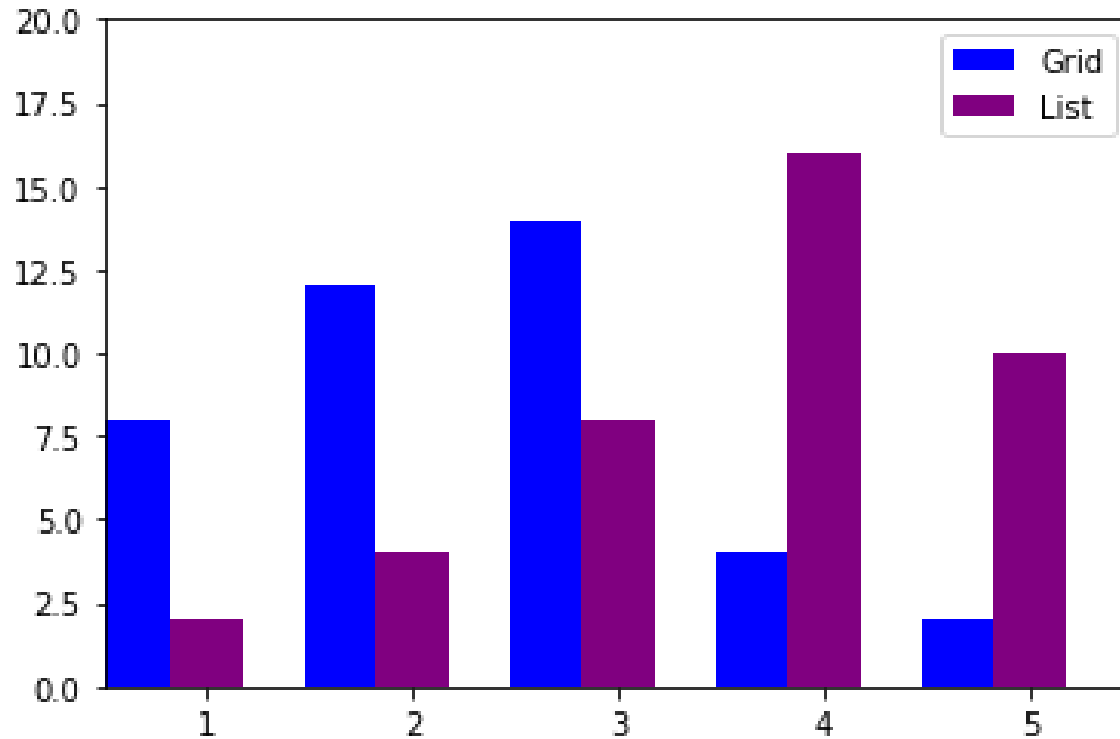


# Generate ratings data

- We assume different subject groups for each condition.
- Each subject sees one of the layouts and is asked to rate on a 5-point Likert scale how strongly he agree or disagree with the statement:
- Question to subjects: Page gives a good overview?
- 1=strongly agree; 2=agree; 3=neutral; 4=disagree; 5=strongly disagree.
  - **G\_data** = [1, 3, 3, 2, 4, 2, 3, 3, 1, 5, 2, 3, 4, 2, 1, 3, 2, 2, 1, 3, 2, 3, 4, 2, 1, 3, 2, 2, 1, 3, 1, 3, 3, 2, 4, 2, 3, 3, 1, 5]
  - **L\_data** = [4, 5, 2, 4, 4, 3, 5, 4, 3, 5, 1, 4, 5, 3, 4, 4, 2, 3, 4, 5, 1, 4, 5, 3, 4, 4, 2, 3, 4, 5, 4, 5, 2, 4, 4, 3, 5, 4, 3, 5]
- **G\_data** corresponds to ratings from users that see the **grid view**.
- **L\_data** corresponds to ratings from users that see the **list view**.



# Visualise ratings data



# Setup: Comparing two versions of a display

- Subjects are users of the display (or summary, interface, etc).
  - **Dependent variable** is user rating (or comprehension, etc).
  - **Independent variable** is the version of the display.
- **Problem:** Find out which version of a display is better.
- **Question:** Do users prefer Grid view?
- **Null hypothesis ( $H_0$ ):** there is no difference between Grid view and List view.

# Significance: Unpaired Student's t-test

- Tests the null hypothesis that two population **means** are equal.
- **Assumptions:**
  - The samples are *independent*.
  - Populations are *normally distributed*.
  - Standard deviations are *equal (by default)*.
- [https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest\\_ind.html#scipy.stats.ttest\\_ind](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html#scipy.stats.ttest_ind)

# Significance: Mann-Whitney U test















- Nonparametric version of unpaired t-test.
  - Test the null hypothesis that the **distribution** underlying sample x is the same as the **distribution** underlying sample y.
- **Assumptions:**
  - The samples are *independent*.
- **Note**
  - N should be at least 20.
- <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html#scipy.stats.mannwhitneyu>

# Exercise: Comparing visual layouts

- Test for difference
  - Run the code cell under “Test whether grid view is preferred”
  - Do users prefer grid view?

# TESTING WHETHER MULTI- GROUPS DIFFER

# Scenario: Mobile use by generation

Talking a different language					
Formative experiences	<b>Maturists</b> (pre-1945) Wartime rationing Rock'n'roll Nuclear families Defined gender roles – particularly for women 	<b>Baby boomers</b> (1945-1960) Cold War 'Swinging Sixties' Moon landings Youth culture Woodstock Family-orientated 	<b>Generation X</b> (1961-1980) Fall of Berlin Wall Reagan/Gorbachev/ Thatcherism Live Aid Early mobile technology Divorce rate rises 	<b>Generation Y</b> (1981-1995) 9/11 terrorists attacks Social media Invasion of Iraq Reality TV Google Earth 	<b>Generation Z</b> (Born after 1995) Economic downturn Global warming Mobile devices Cloud computing Wiki-leaks 
Attitude toward career	Jobs for life 	Organisational - careers are defined by employees	"Portfolio" careers - loyal to profession, not to employer	Digital entrepreneurs - work "with" organisations	Multitaskers - will move seamlessly between organisations and "pop-up" businesses
Signature product	Automobile 	Television 	Personal computer 	Tablet/smartphone 	Google glass, 3-D printing
Communication media	Formal letter 	Telephone 	E-mail and text message 	Text or social media 	Hand-held communication devices
Preference when making financial decisions	Face-to-face meetings	Face-to-face ideally but increasingly will go online	Online - would prefer face-to-face if time permitting	Face-to-face	Solutions will be digitally crowd-sourced

<https://ihumanmedia.com/2015/09/14/gen-x-millennials-vs-baby-boomer-real-estate-baby-work-travel-politics-shopping/>

## Research question

**Does mobile use differ across generations?**



# Data/Measurement: Survey of mobile use

- May be collected by survey or user data.
- **Dependent variable:**
  - Number of texts per day.
- **Independent variable:**
  - Generation {B,G,M}.

## Texting survey

1. What year were you born?
2. How many texts do you send per day?

# Significance: Analysis of variance (ANOVA)

- Tests the null hypothesis two or more groups have the same population mean.
- **Assumptions:**
  - The samples are *independent*.
  - Populations are *normally distributed*.
  - Standard deviations are *equal*.
- [https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f\\_oneway.html#scipy.stats.f\\_oneway](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html#scipy.stats.f_oneway)

# Significance: Kruskal-Wallis H-test

- Nonparametric version of ANOVA.
  - Test the null hypothesis that the population **median** of all of the groups are equal
- **Assumptions:**
  - Samples are *independent*.
- **Note:**
  - Not recommended for samples smaller than 5.
  - Not as statistically powerful as ANOVA.
  - Both ANOVA and Kruskal-Wallis H-test are extensions of the Unpaired Student's t-test and Mann-Whitney test used to compare the means of more than two populations.
- <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kruskal.html#scipy.stats.kruskal>

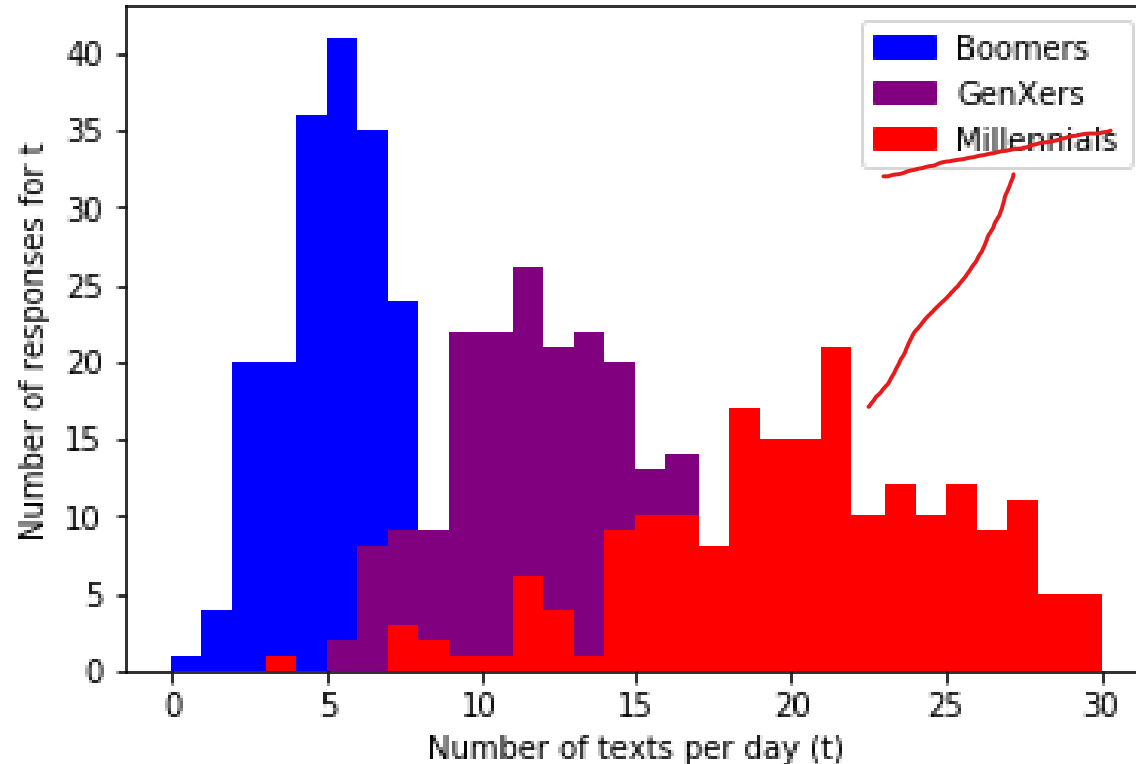
# Setup: Comparing behaviour across groups

- Subjects are rows of data.
  - **Dependent variable** is number of texts per day.
  - **Independent variable** is generation {B,G,M}.
- **Q:** Is there any difference between groups?
- **H<sub>0</sub>:** Group means (or medians, for nonparametric methods) are the same

# Generate generation data

- Imagine we conducted a survey of **200 baby boomers** (born 1945-1960), **200 generation Xers** (born 1961-1980) and **200 millennials** (born 1981-1995).
- For the purposes of this exercise, let's generate some simulated samples. We assume:
  - **Baby Boomers** send 5 texts per day on average with standard deviation 2.
  - **GenXers** send 12 texts per day on average with standard deviation 3.
  - **Millennials** send 20 texts per day on average with standard deviation 5.

# Visualise generation data



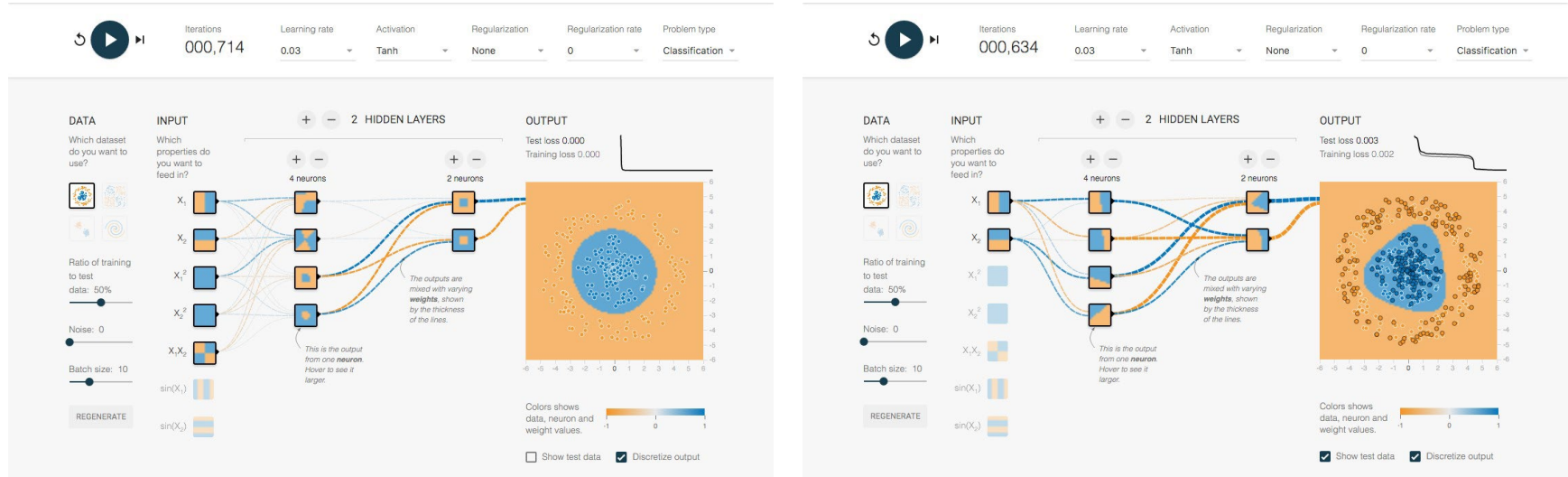
# Exercise: Comparing mobile behaviour

- Test for difference
  - Run the code cell under “Testing for differences”
  - Does the data satisfy ANOVA assumptions?

# TESTING WHICH APPROACH IS BETTER WITHIN SUBJECTS



# Example scenario: Comparing classifiers



<http://playground.tensorflow.org/>

## Research question

**Does my new model perform  
better?**

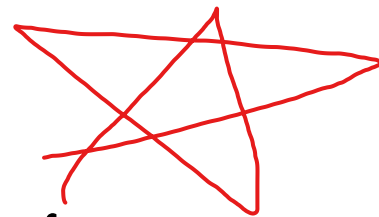
## Task: Spam/ham detection

- Let's assume our classifiers predict whether an email is:
  - 1: spam.
  - 0: ham.
- Features are words, e.g.:
  - .P.a.Y.p.a.l, bitcoin\_up, iphone.1 4.Pro, winner, Settlement4U.

# Measurement: Model evaluation

- Need to measure accuracy of system output  $S$ .
- Compare to gold-standard labelling  $G$ .
- Define evaluation measure:  $\text{score}(S, G)$ .
- [https://scikit-learn.org/stable/modules/model\\_evaluation.html](https://scikit-learn.org/stable/modules/model_evaluation.html)

# Measurement: Accuracy, precision, recall, f1



		Model prediction	
		Spam (s=1)	Ham (s=0)
Actual results	Spam (g=1)	<i>TP</i> (true positives)	<i>FN</i> (false negatives)
	Ham (g=0)	<i>FP</i> (false positives)	<i>TN</i> (true negatives)

- **Accuracy:** percentage of correct over all instances.
  - $(TP+TN) / N$
- **Precision:** percentage of correct system predictions.
  - $TP / (TP+FP)$
- **Recall:** percentage of correct gold labels.
  - $TP / (TP+FN)$
- **F1:** Harmonic mean of Precision and Recall.
  - $2PR / (P+R)$

## Confusion matrix for more than two classes

- E.g. iris data classification - confusion matrix:

a b c <-- classified as  
50 0 0 | a = Iris-setosa  
0 44 6 | b = Iris-versicolor  
0 3 47 | c = Iris-virginica

	<u>Setosa</u> +	Versicolor-	Virginica-
<u>Setosa</u> +	50 <u>tp</u>	0 <u>fn</u>	0 <u>fn</u>
Versicolor-	0 <u>fp</u>	44	6
Virginica-	0 <u>fp</u>	3	47

- accuracy =?

$$\text{accuracy} = \frac{(\text{tp} + \text{tn})}{(\text{tp} + \text{fn} + \text{fp} + \text{tn})}$$

- accuracy =  
 $(50 + 44 + 47) / (50 + 0 + 0 + 0 + 44 + 6 + 0 + 3 + 47)$   
 $= 141 / 150 = 94\%$

# Evaluating classifier accuracy: Holdout & cross-validation methods

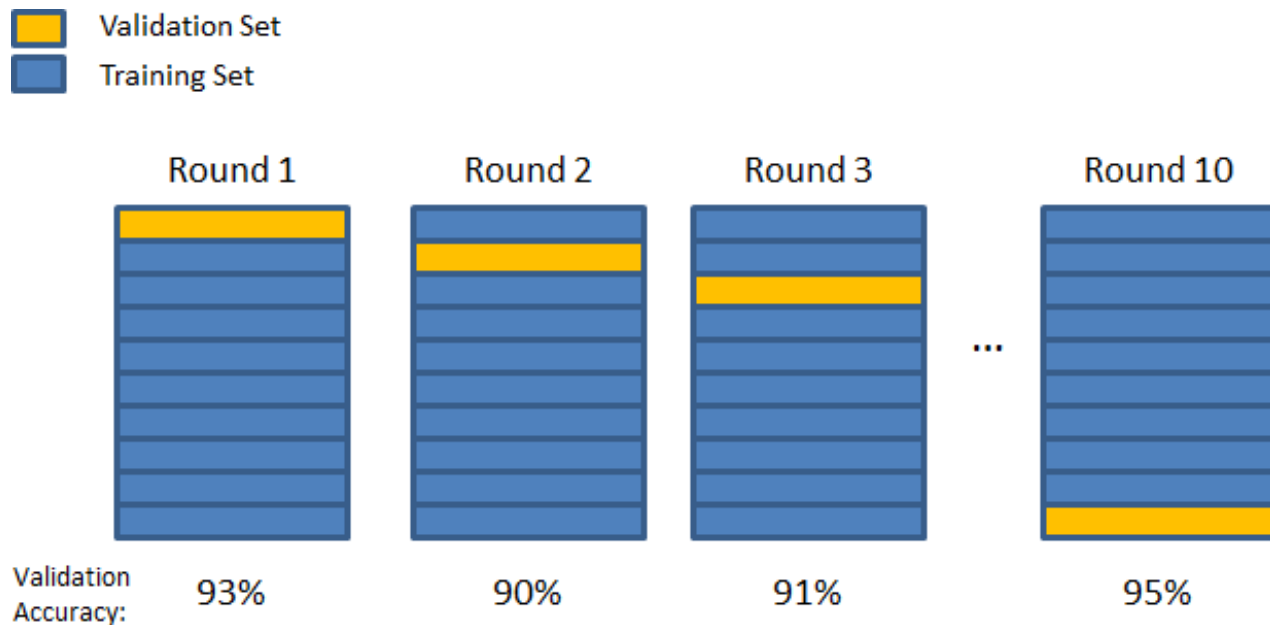
## Holdout method

- Splits the data randomly into two independent sets.
  - Training set (e.g.,  $2/3$ ) for model construction.
  - Test set (e.g.,  $1/3$ ) for accuracy estimation.
  - Repeat holdout  $k$  times, accuracy = avg. of the accuracies obtained.

## Cross-validation ( $k$ -fold, where $k = 10$ is most popular)

- Randomly partition the data into  $k$  mutually exclusive subsets, each approximately equal size.
- Leave-One-Out is a particular form of cross-validation:
  - $k$  folds where  $k = \#$  of tuples, for small sized data.

# Data: Cross validation



Final Accuracy = Average(Round 1, Round 2, ...)

<https://chrismccormick.wordpress.com/2013/07/31/k-fold-cross-validation-with-matlab-code/>



# Significance: Paired Student's t-test

- Tests the null hypothesis that two population means are equal.
- **Assumptions:**
  - The samples are *paired* (e.g. before and after a treatment).
  - Populations are *normally distributed*.
  - Standard deviations are *equal*.
- [https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest\\_rel.html#scipy.stats.ttest\\_rel](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html#scipy.stats.ttest_rel)

# Significance: Paired tests for non-parametric data

- Nonparametric version of paired t-test.
- **Assumptions:**
  - The samples are *paired*.
- **Note:**
  - Often used for ordinal data, e.g., Likert ratings.
  - N should be large, e.g.,  $\geq 20$ .
- <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html#scipy.stats.wilcoxon>

# Generate gold and classifier labellings

- We generate **10,000 gold labels**.
  - Marking approximately **20% as spam** (1) based on a random number generator and the rest as ham (0).
    - 0: 8000 and 1: 2000
- **System 1** incorrectly marks 5% of ham as spam and fails to detect 20% of actual spam.
- **System 2** incorrectly marks 10% of ham as spam and fails to detect 10% of actual spam.

System 1		PREDICTED	
		1	0
ACTUAL	1	1 600 (TP)	400 (FN)
	0	400 (FP)	7 600 (TN)

System 2		PREDICTED	
		1	0
ACTUAL	1	1800	200
	0	800	7200

## Setup: Comparing classifiers

- Subjects correspond to cross-validation folds.
  - **Dependent variable** is some measure of accuracy (precision, recall, f1, etc).
  - **Independent variable** is the algorithm, feature set, etc.
- **Q:** Is my shiny, new model better?
- **H<sub>0</sub>:** Accuracy is not better for the new model.

# Exercise: Comparing models

- Generate data
  - Run the code cell under “Generate gold and classifier labelling”
  - Run the code cell under “Split data into folds”
- Calculate accuracy
  - Run the code cell under “Calculate classifier accuracy”
  - Run the code cell under “Calculate scores across folds”
- Test for differences
  - Run the code cell under “Compute significance for sys1 and sys2”
  - How can we manage reliability?

# REVIEW

## Tips and tricks

- Statistical hypothesis testing ensures results are reliable.
- Experimental design includes:
  - Formulating a research question and null hypothesis.
  - Designing and running experiments.
  - Analysing results using appropriate statistics.
- Use textbooks and documentation to find the right stats.
- Sample representatively; Report p-value; Don't hack p-value.
- Report precision, recall, f-score and significance.

# Additional reading (not examinable)

Some great online resources:

- Hypothesis testing, power, sample sizes
  - <https://online.stat.psu.edu/stat415/>
- What does it all even mean?
  - <https://plato.stanford.edu/entries/statistics/>