COMP5310: Principles of Data Science

W6: Hypothesis testing & evaluation

Presented by
Dr Ali Anaissi
School of Computer Science

THE UNIVERSITY OF SYDNEY

Overview of Week 6

Today: Hypothesis testing and evaluation

Objective

Overview of experimental design and learn Python tools for hypothesis testing and classifier evaluation.

Lecture

- Questions: Do two populations differ? Which approach is better?
- Significance: pairwise t-tests, non-normal data
- Evaluation: classifier evaluation

Readings

- Data Science from Scratch, Ch. 7
- Hypothesis testing (scipy lectures) http://goo.gl/HCf3nP
- Model evaluation (sklearn doco) http://goo.gl/Avj51Y

Exercises

- scipy: statistical tests
- sklearn: evaluation metrics

TODO in W6

- Submit project stage 1

Goal of this lecture

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

Where to find more details

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/

Imagine..

Bob is developing a new diet

- Bob tries it on a sample of friends and family
- The mean weight loss for the diet is lower than methods compared
- Bob invests his savings in his diet startup

BUT the result was not reliable

- Followers of the diet report actually gaining weight!

What was Bob's mistake?

Types of Statistical Studies

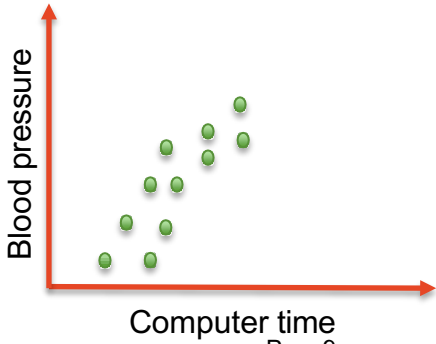
Types of Statistical Studies

- 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. e.g.
    - 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.
    - Study 2: Average Computer Time vs Blood Pressure
      - Enroll 100 individuals in the observational study.
      - Ask them about the average computer time they spent each day.
      - Measure their blood pressure.
  - Only establish correlation not causality



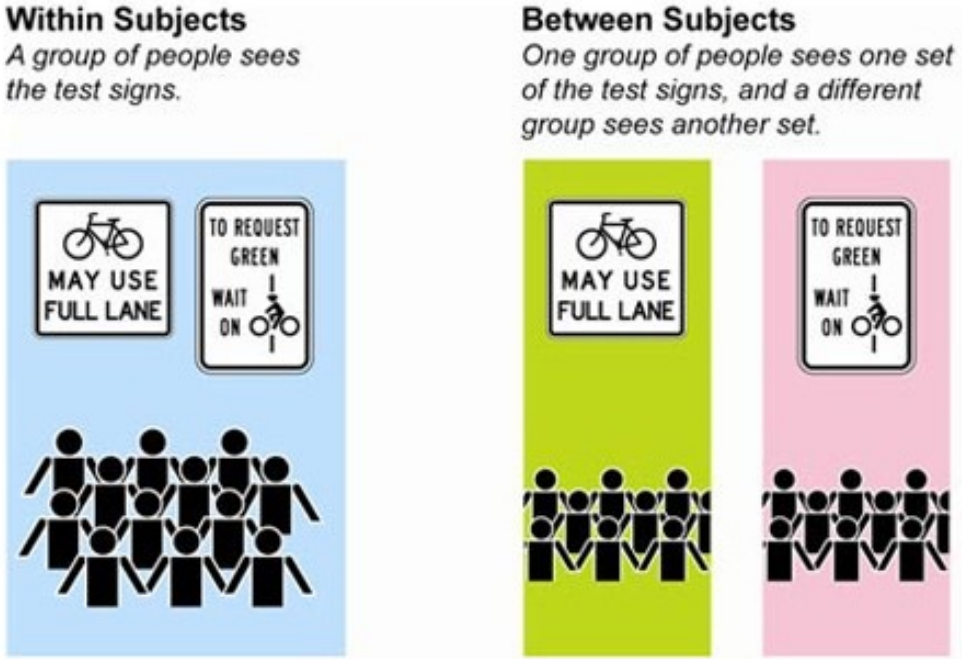
Experimental Studies

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

Definitions

Experiment design

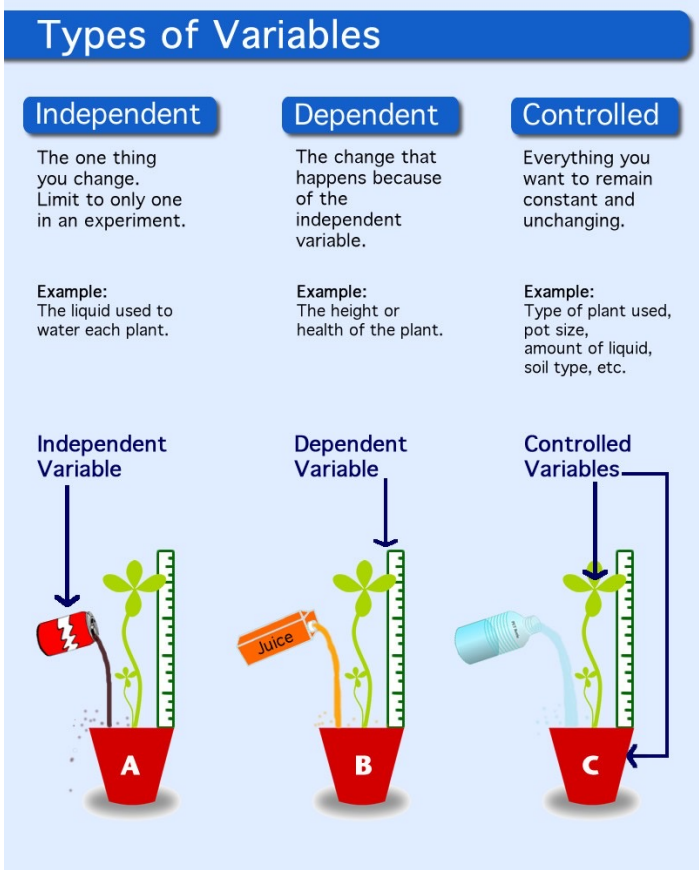
- ***Between subjects***: Each subject sees one and only one condition
- ***Within subjects***: Subjects see more than one or all conditions



http://www.fhwa.dot.gov/publications/research/safety/pedbike/11035/004.cfm

Types of variables

- ***Dependent variable*** is the measure of interest
- ***Independent variable*** is manipulated to observe the effect on dependent variable
- ***Controlled variables*** are materials, measurements and methods that don't change



http://edtech2.boisestate.edu/angelacovil/506/procedure.html

Research question

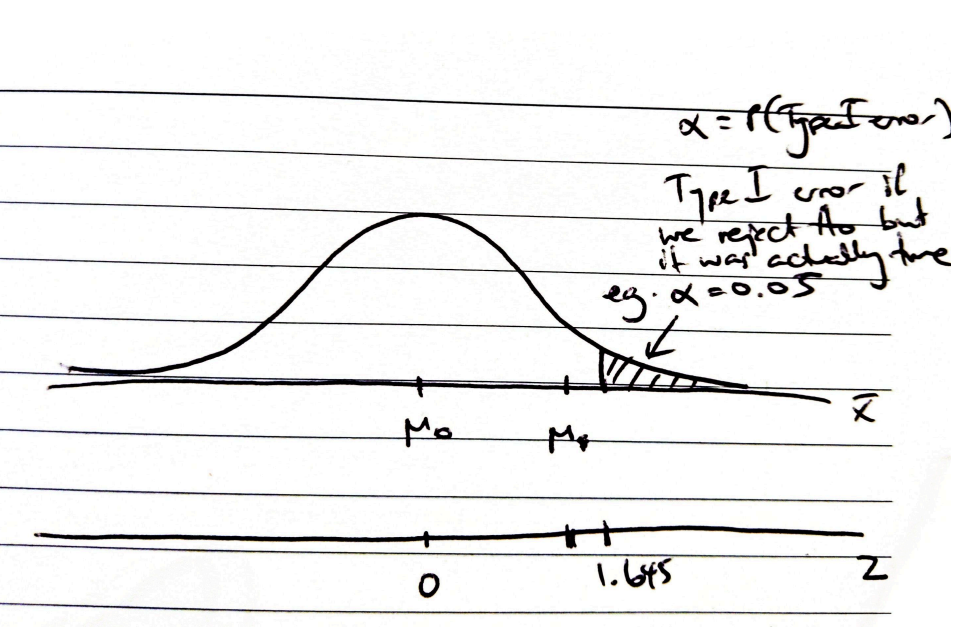
- Research question (Q):
  - Asks whether the independent variable has an effect
  - “If there is a change in the independent variable, will there also be a change in the dependent variable?”
- Null hypothesis (H_0):
  - The assumption that there is no effect
  - “There is no change in the dependent variable when the independent variable changes.”

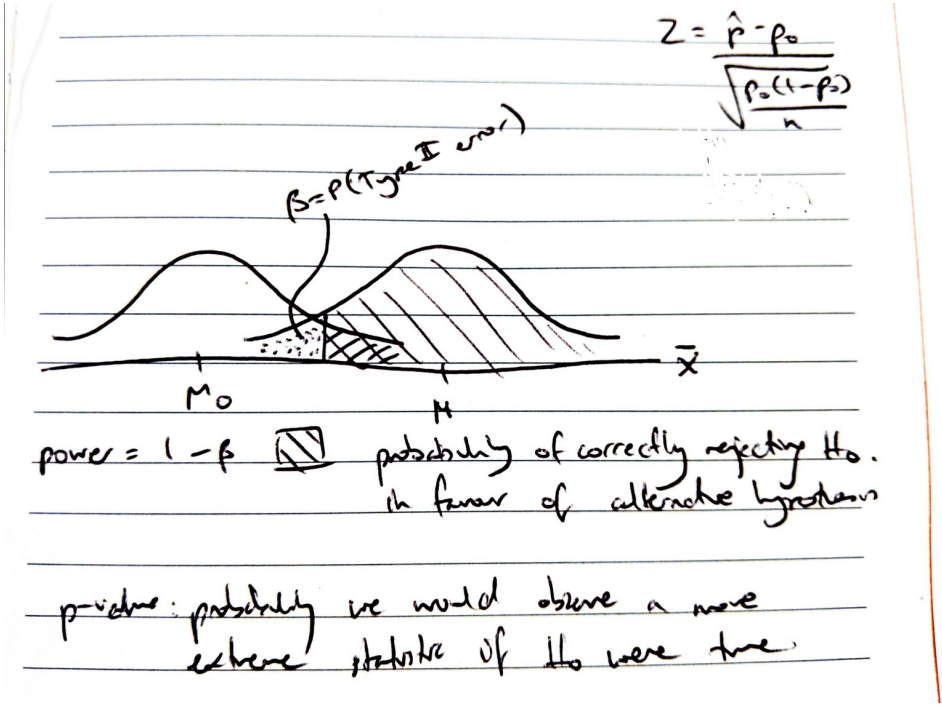
Statistical significance testing

Hypothesis testing

- We use it 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**
    - **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.

Hypothesis testing





Testing reliability with p-values

- Most tests calculate a p-value for measuring observation extremity: more extreme values
- Compare to significance level threshold α
  - α is the probability of (wrongly) rejecting H_0 given that it is true
  - aka Type I error rate (false positive)
  - Commonly use α of 5% or 1%

Truth \ Decision	Accept H_0	Reject H_0
H_0 (No difference)	Right Decision	Type I error
H_1 (Difference exists)	Type II error	Right Decision

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

Not every test result is correct

- P=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
- http://faculty.washington.edu/dwhm/2016/03/09/the-arbitrary-magic-of-p-0-05/

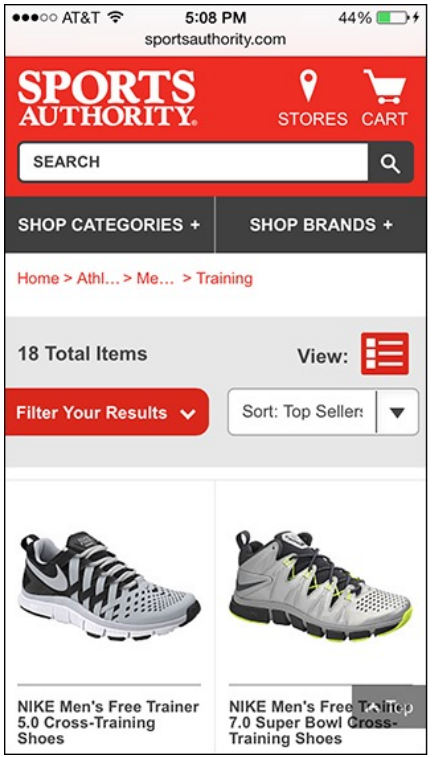
Increase the power of a significance test

- Obtain a larger sample
- Larger N means more reliable statistics
- Less likely to have errors
  - Type I: Reject true H_0
  - Type II: Fail to reject false H_0

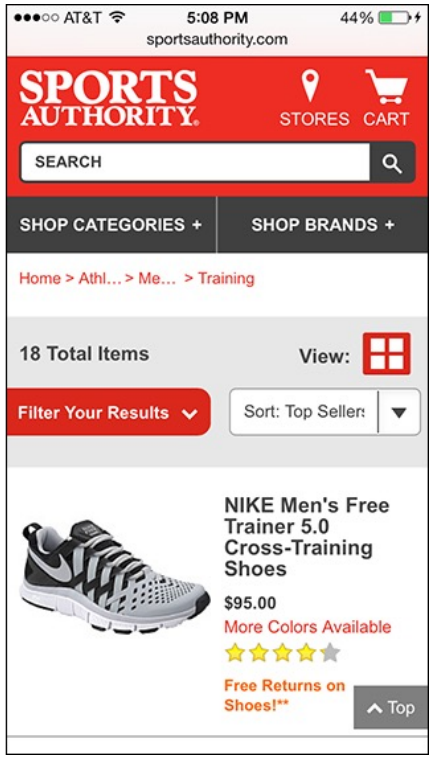
Testing which approach is better between subjects

Scenario: Comparing visual layouts

Grid view



List view



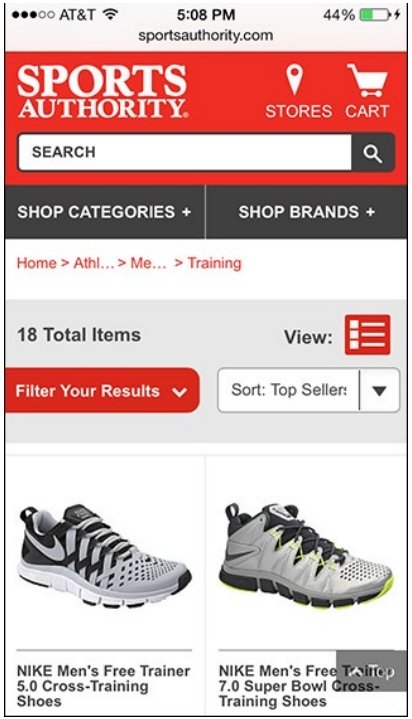
https://www.nngroup.com/articles/image-vs-list-mobile-navigation/

Research question

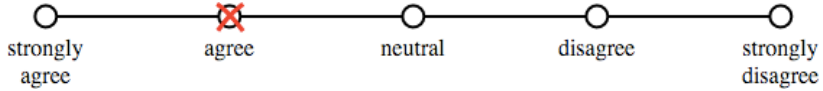
Do users prefer grid or list view?

Data/Measurement: User ratings of layouts

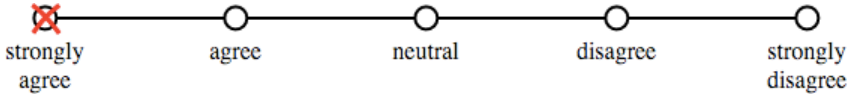
Example response from User Group A



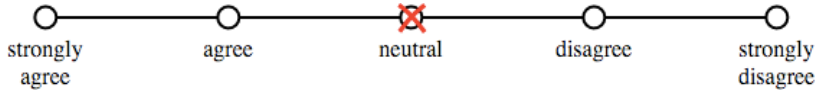
Page is easy to use.



Page gives good overview.

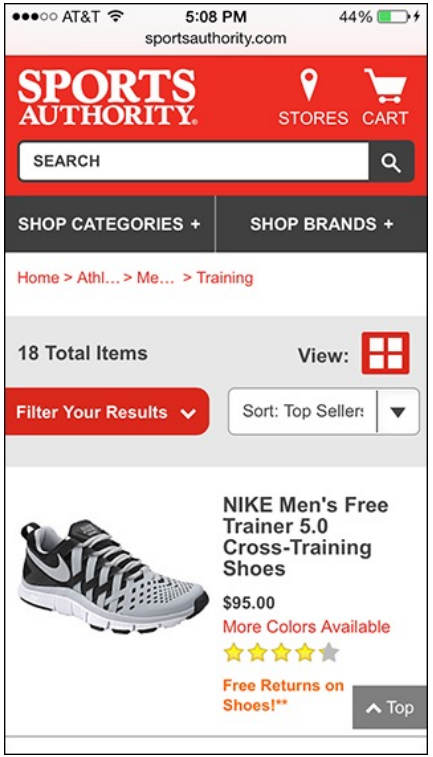


Page gives sufficient detail.

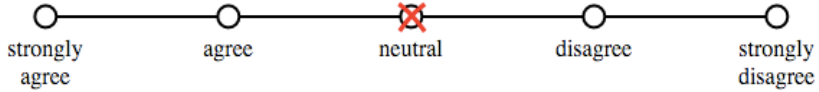


Data/Measurement: User ratings of layouts

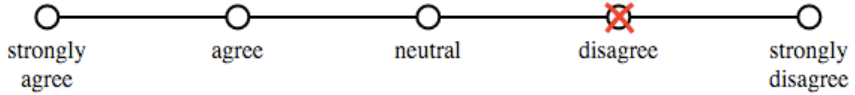
Example response from User Group B



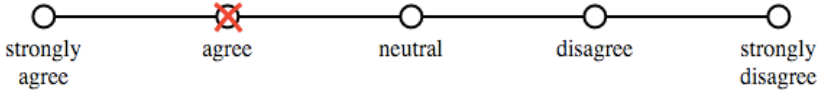
Page is easy to use.



Page gives good overview.



Page gives sufficient detail.



Generate ratings data

- We assume different subject groups for each conditions.
- 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?

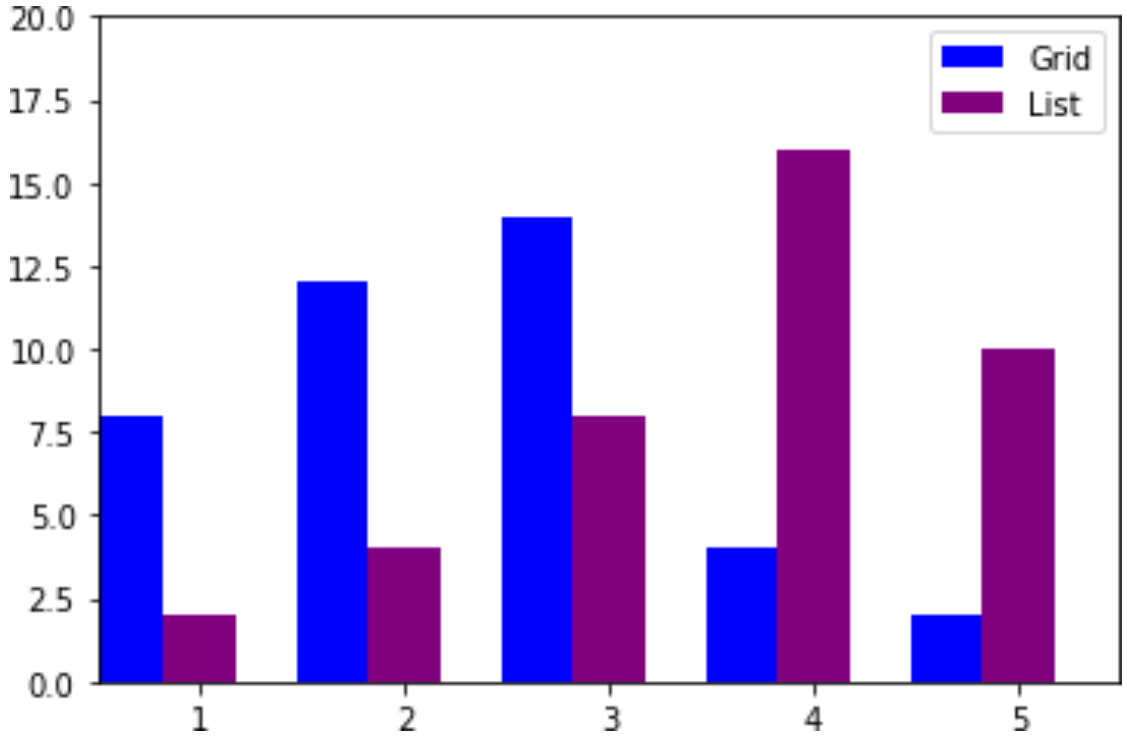
  1=strongly agree; 2=agree; 3=neutral; 4=disagree; 5=strongly disagree

  - 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 : Users do not prefer Grid view.

Significance: Unpaired Student's t-test

- Tests the null hypothesis that two population means are equal
- Assumes
  - The samples are independent
  - Populations are normally distributed
  - Standard deviations are equal
- Note
  - Multiply two-tailed p-value by 0.5 for one-tailed p-value (e.g., to test A>B, rather than A>B OR A<B)
- http://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
- Assumes
  - The samples are independent
- Note
  - N should be at least 20

The Mann-Whitney U test is a nonparametric test of the null hypothesis that the distribution underlying sample x is the same as the distribution underlying sample y.
It is often used as a test of difference in location between distributions.

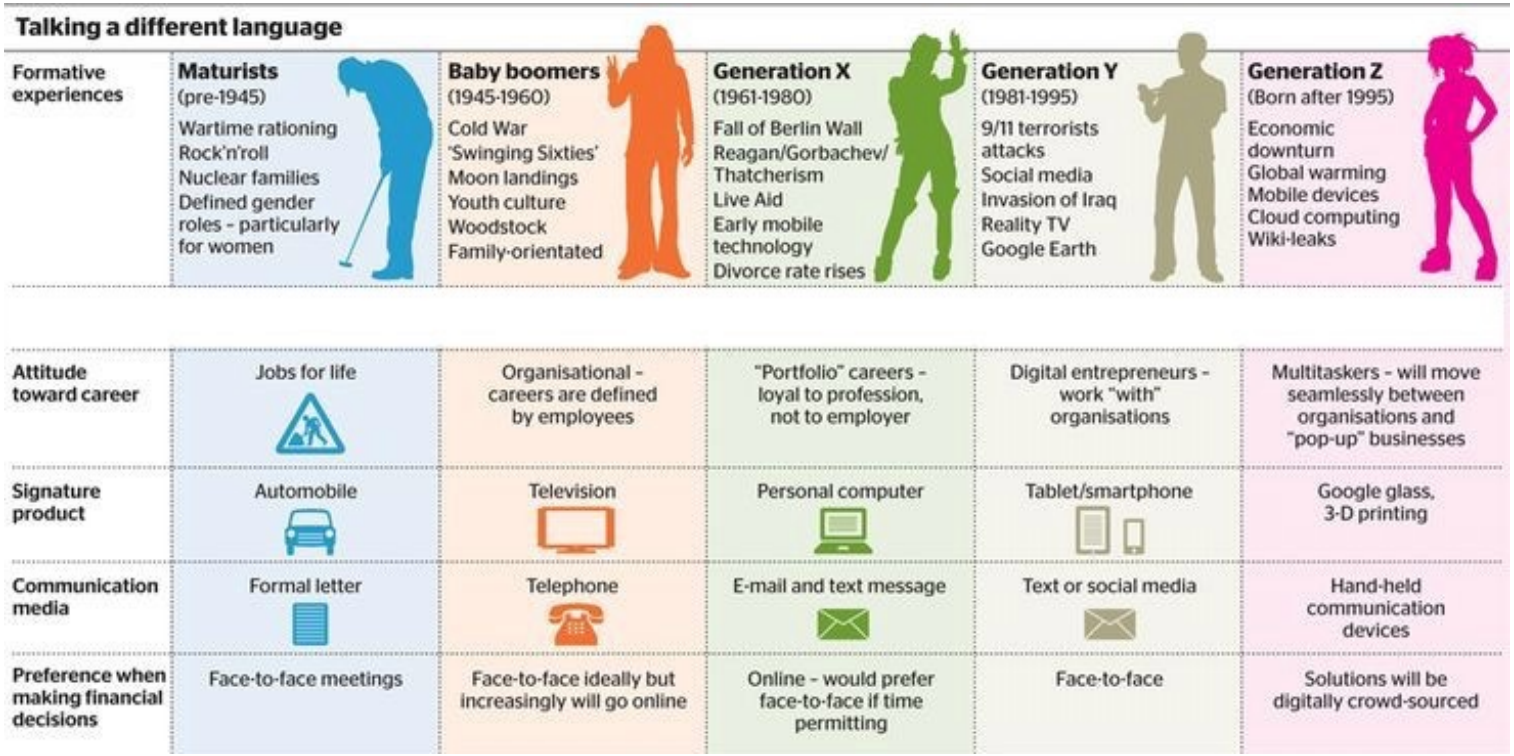
- http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html#scipy.stats.mannwhitneyu

Exercise: Comparing mobile behaviour

- Create data
  -  code cell under “Create ratings data”
  -  code cell under “Visualise ratings data”
- Test for differences
  -  code cell under “Test whether list is preferred”
  - Do users prefer list view?
- Further exercises
  -  p-hacking example

Testing whether groups differ

Scenario: Mobile use by generation



Talking a different language

	Maturists (pre-1945)	Baby boomers (1945-1960)	Generation X (1961-1980)	Generation Y (1981-1995)	Generation Z (Born after 1995)
Formative experiences	Wartime rationing Rock'n'roll Nuclear families Defined gender roles - particularly for women	Cold War 'Swinging Sixties' Moon landings Youth culture Woodstock Family-orientated	Fall of Berlin Wall Reagan/Gorbachev/ Thatcherism Live Aid Early mobile technology Divorce rate rises	9/11 terrorists attacks Social media Invasion of Iraq Reality TV Google Earth	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

What year were you born?

_____

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
- Assumes:
  - The samples are independent
  - Populations are normally distributed
  - Standard deviations are equal
- http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html#scipy.stats.f_oneway

Significance: Kruskall-Wallis H-test

- Nonparametric version of ANOVA
  - The test doesn't assume your data comes from a particular distribution such as normal distribution.
- Assumes samples are independent
- It is sometimes called the one-way ANOVA on ranks
  - as the ranks of the data values are used in the test rather than the actual data
- Note:
  - Not recommended for samples smaller than 5
  - Not as statistically powerful as ANOVA
  - Both ANOVA and Kruskall-Wallis H-test are extension of the Mann-Whitney test and Unpaired Student's t-test used to compare the means of more than two populations.

The Kruskal-Wallis H-test tests the null hypothesis that the population median of all of the groups are equal.

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kruskal.html#scipy.stats.kruskal

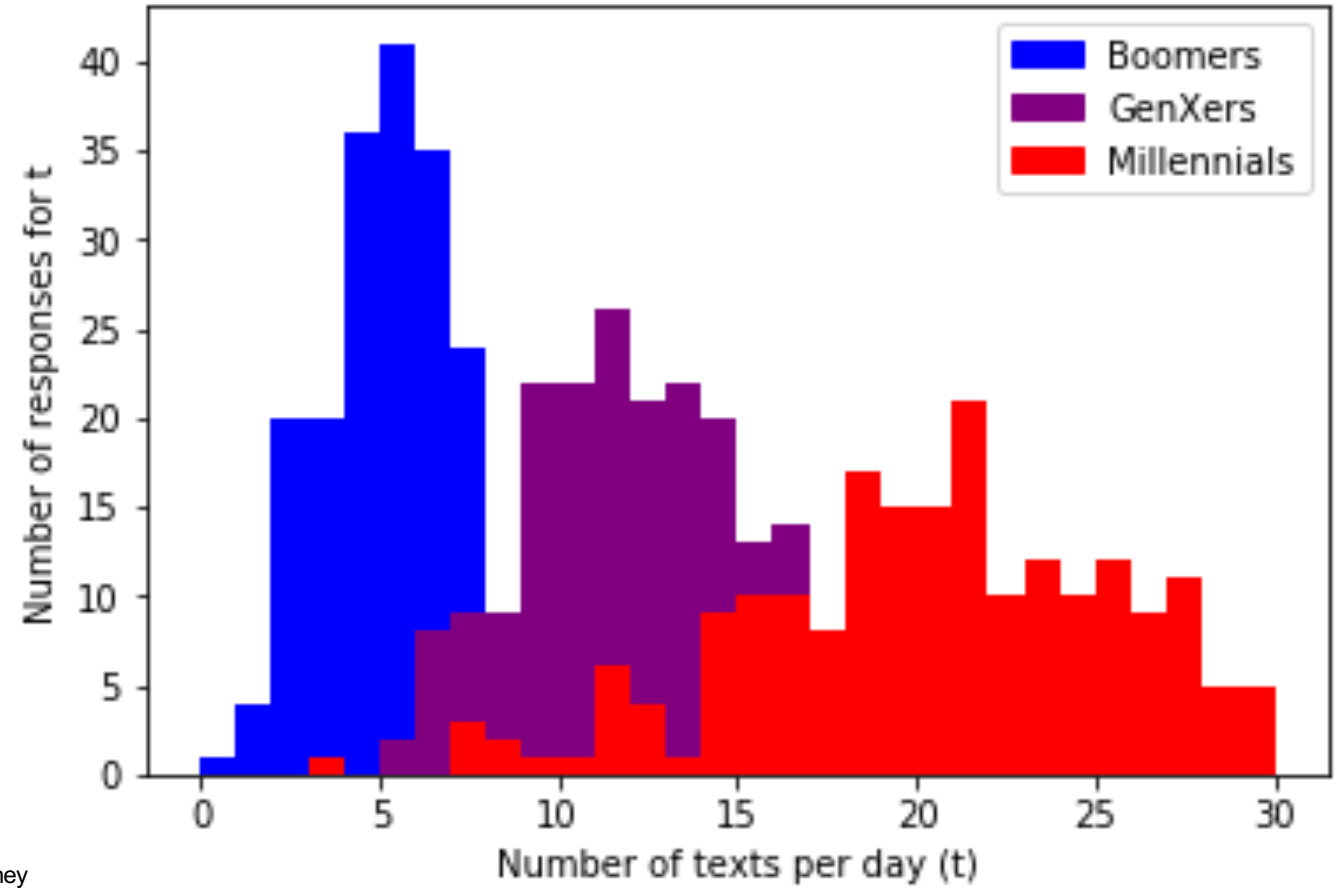
Setup: Comparing behavior across groups

- 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_0: 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

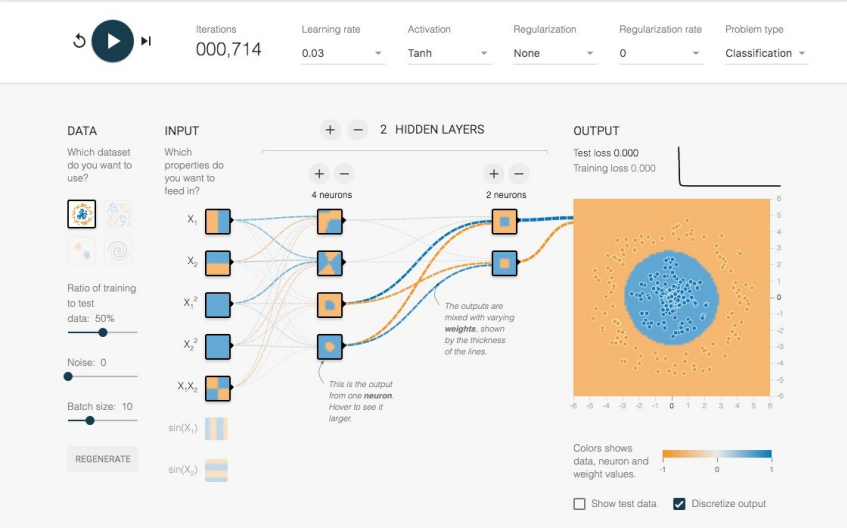


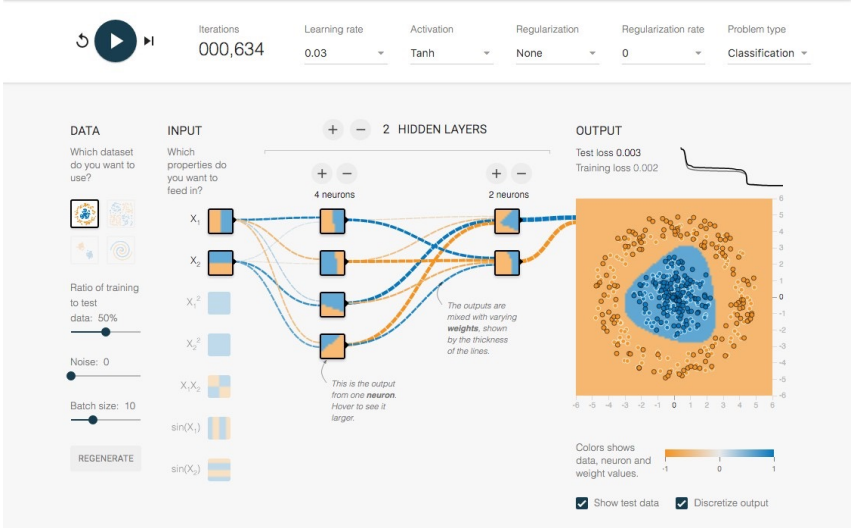
Exercise: Comparing mobile behaviour

- Generate data
  -  code cell under “Generate generation data”
  -  code cell under “Visualise generation data”
- Test for differences
  -  code cell under “Testing for differences”
  - Does the data satisfy ANOVA assumptions?
- Further exercises
  - Do millennials and generation Z differ?

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, eg:

  .P.a.Y.p.a.l, bitcoin_up, iphone.14.Pro, winner, Settlement4U

Measurement: Model evaluation

- Need to measure accuracy of system output `S`
- Compare to gold-standard labelling `G`
- Define evaluation measure: score(`S`, `G`)
- http://scikit-learn.org/stable/modules/model_evaluation.html#model-evaluation

Measurement: Accuracy, precision, recall, f1

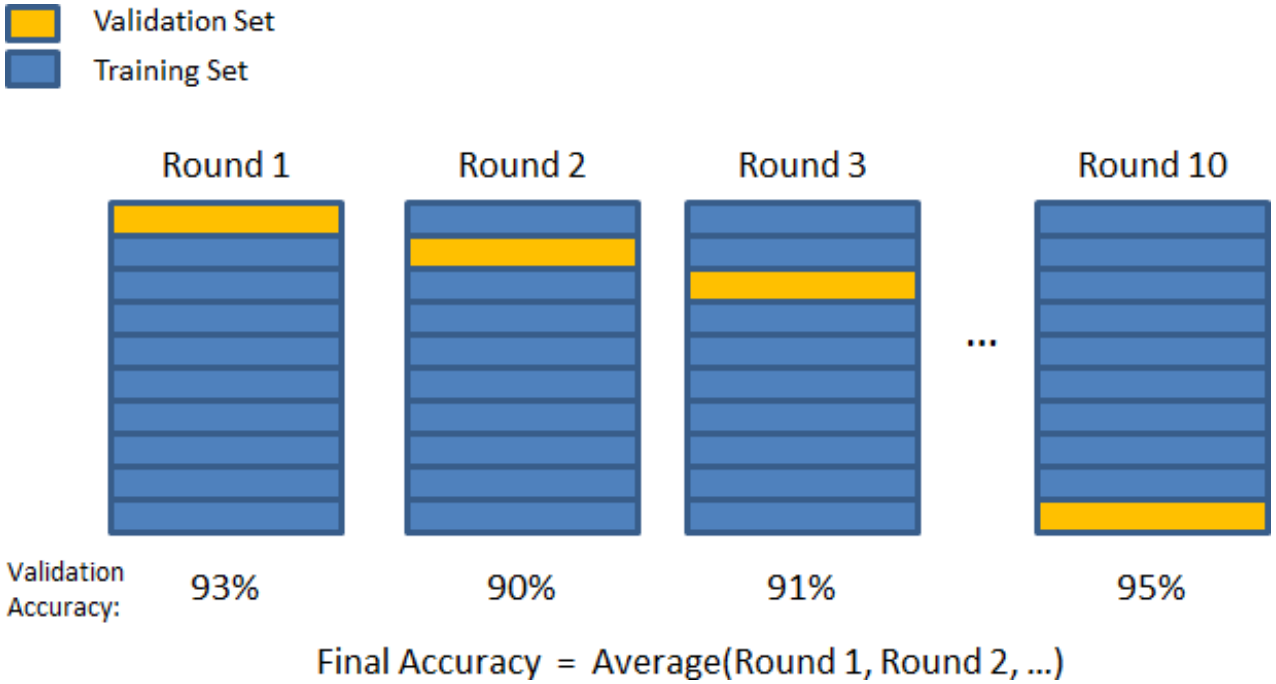
	s=1	s=0
g=1	*TP* *(true positives)*	*FN* *(false negatives)*
g=0	*FP* *(false positives)*	*TN* *(true negatives)*

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

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
  - Random sampling: a variation of holdout
    - 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



https://chrisjmccormick.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
- Assumes
  - ***The samples are paired (e.g. before and after a treatment)***
  - Populations are normally distributed
  - Standard deviations are equal
- Note
  - Multiply two-tailed p-value by 0.5 for one-tailed p-value (to test A>B, rather than A>B OR A<B)
- http://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
- Assumes
  - The samples are paired
- Note
  - Often used for ordinal data, e.g., Likert ratings
  - N should be large, e.g., ≥20
- http://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 a 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	Predicted 0
Actual	1	1600	400
	0	400	7600

System 2:

		Predicted 1	Predicted 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_0: Accuracy is not better for new model
- http://sci2s.ugr.es/keel/pdf/algorithm/articulo/dietterich1998.pdf

Exercise: Compare models

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

Review

THE UNIVERSITY OF SYDNEY

Today: Hypothesis testing and evaluation

Objective

Learn Python tools for exploring a new data set programmatically.

Lecture

- Questions: Do two populations differ? Which approach is better?
- Significance: pairwise t-tests, confidence intervals, non-normal data
- Evaluation: cluster evaluation, classifier evaluation, user evaluation

Readings

- Model evaluation (sklearn doco) http://scikit-learn.org/stable/modules/model_evaluation.html#model-evaluation
- Hypothesis testing (scipy lectures) http://www.scipy-lectures.org/packages/statistics/index.html#hypothesis-testing-comparing-two-groups

Exercises

- scipy: statistical tests
- sklearn: evaluation metrics

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)

- Your favourite statistics text book (or statistician)
- Montgomery. Design and analysis of experiments. http://opac.library.usyd.edu.au/record=3416341
- Robertson and Kaptein. Modern statistical methods for HCI. http://www.springer.com/gb/book/9783319266312
- Hartson. The UX book. Chapters 12-18. http://opac.library.usyd.edu.au/record=b4415045~S4
- Scott. Multi-armed bandit experiments. https://support.google.com/analytics/answer/2844870?hl=en

Next Time

Next week: Data Mining

Objective

Learn Python tools for data mining with a focus on clustering and association rule mining.

Lecture

– Association rule mining.

Readings

– Data Science from Scratch, Ch. 11, 19

Exercises

– Associations from scratch (Readings)

TODO in Week 7

– Start work on Assignment 2