COMP30120

Evaluation in Machine Learning
Part 1

Derek Greene

School of Computer Science and Informatics

Autumn 2015

Overview

- Part 1
  - Objectives of Evaluation
  - A/B Testing
  - Basic Statistics Reminder
  - Statistical Significance
    - Student’s t-test
    - Tests for proportions
- Part 2
  - Evaluation Measures
  - Overfitting
  - Experimental Setup

Objectives of Evaluation

Q. Is machine learning algorithm *A* better than algorithm *B*?

- **Supervised Learning**
  - Does classifier *A* have better accuracy than *B* on a given dataset?
  - Does classifier *A* have better accuracy across many different datasets?
  - What is the difference in *generalisation performance* on new data not seen in training?
- **Unsupervised Learning**
  - Does clustering algorithm X provide more useful or interpretable results than algorithm Y?

Q. Is the difference between the results statistically significant?

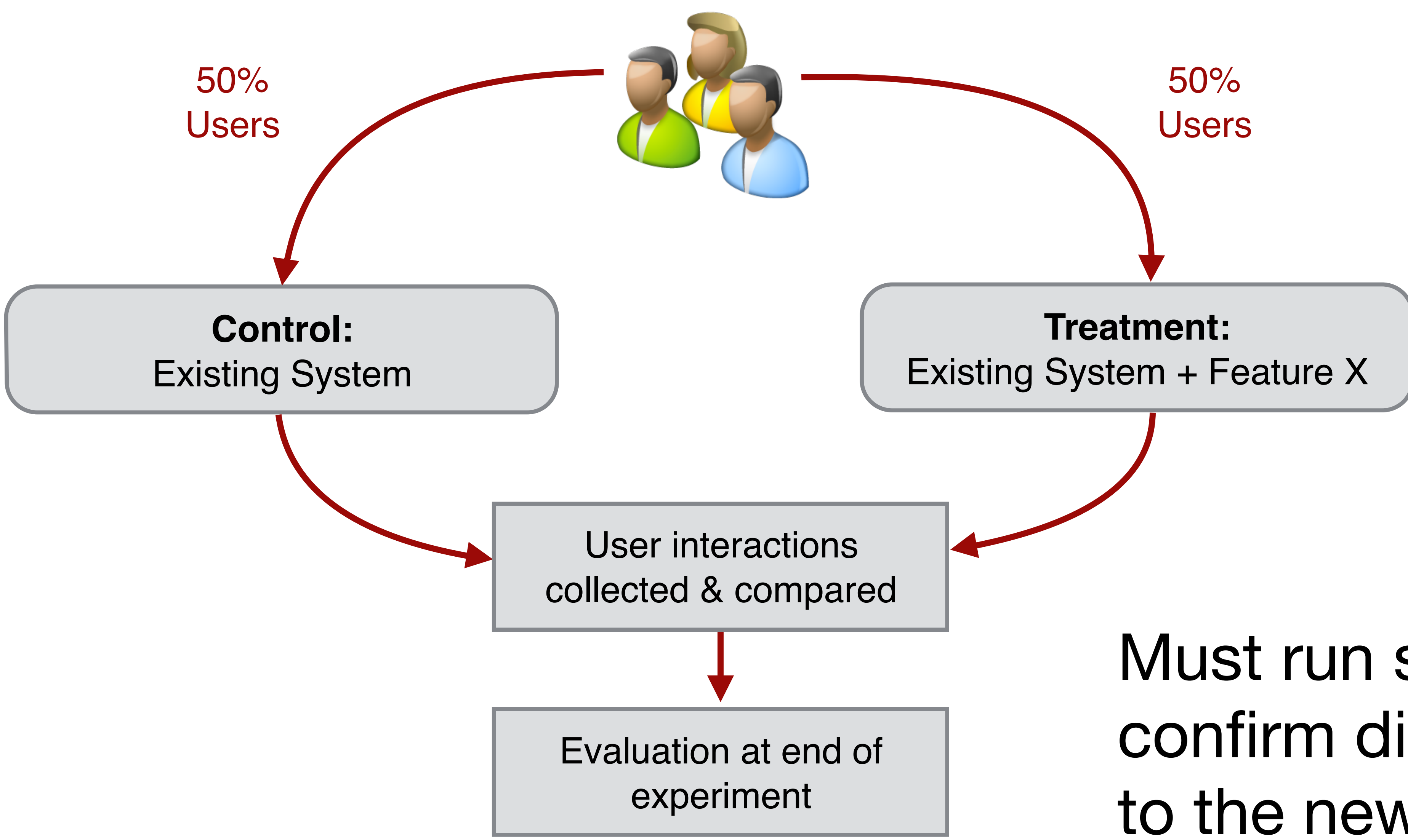
A / B Testing

- **Example: Amazon Shopping Cart Recommendations**
  - Add an item to your shopping cart at a website, most sites then show cart to the user.
  - At Amazon, Greg Linden suggested showing the user recommendations based on cart items instead.
  - What are the possible effects of this website change?
    - ✓ Pro: cross-sell more items (increase average basket size)
    - X Con: distract people from checking out (reduce conversion)
  - Evaluation: Simple user experiment was run, change was wildly successful.

http://glinden.blogspot.com/2006/04/early-amazon-shopping-cart.html

Simple Controlled Experiments

1. Randomly split traffic between two or more versions
   e.g. (A) Control, (B) Treatment
2. Collect and analyse metrics of interest



Must run statistical tests to confirm differences are due to the new feature, not due to chance!

Basic Statistics Reminder

- Let $(x_1, x_2, \ldots, x_n)$ be the values of some variable (data) *X*, for a sample of size *n*.
- The arithmetic mean of the data *X* is calculated as:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Measures of dispersion characterise how spread out the distribution of the sample is - i.e. how variable the data are.
- The variance is the arithmetic mean of the squared deviations from the sample mean.

$$var(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$$

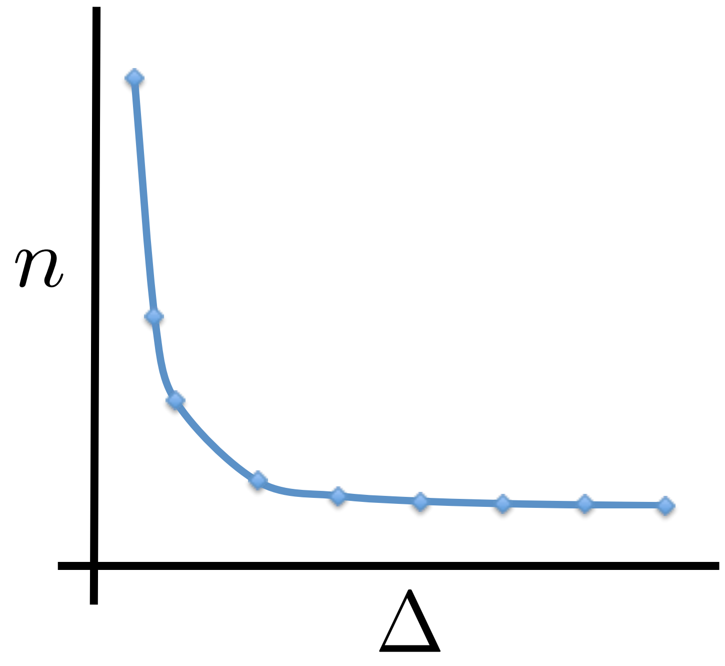
- The standard deviation is the square-root of the variance.

$$\sigma(X) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}}$$

Hypothesis Testing

- For statistical significance, the most important relationship is between the difference (delta) and sample size *n*.

$$n \propto \frac{1}{\Delta^2}$$



Inverse square relationship

- The smaller the difference, the more data required to test the hypothesis.
- In the past, getting enough data to test a hypothesis was the problem.
- Now we often (but not always) have to deal with an overabundance of data.

Example: Statistical Significance

- Have cases of two different treatment for broken wrists. Two groups:
  1. *Control:* Plaster Cast
  2. *Treatment:* Surgery (Pins) + Cast
- Want to test for difference in proportions. Is there a significant difference between the control and the treatment?

	Control	**Treatment**
Size	50	50
Cured	10	20

Contingency Table

- Difference between two groups is statistically significant (p≅0.04).
- If only 18/50 patients in treatment group had been cured (instead of 20/50), this difference would not be significant.
- Small sample size has a substantial impact on significance here.

➡ Good News: Significant effects do not always require big data

Example: Statistical Significance

- Report on deaths after surgery surveyed over one week in 2011.
- Is there a significant difference between death rates in UK and Ireland?

	UK	Ireland
Cohort	10630	856
Died	378	55
	3.56%	6.43%

- Difference between two groups is statistically significant.
- If only 41/856 patients in Ireland had died (instead of 55/856), difference would not be significant. Difference could be due to chance.
- Small sample size has a substantial impact on significance here.

Mortality after surgery in Europe
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3493988/

Hypothesis Testing

- The goal of hypothesis testing is to formally examine two opposing hypotheses H_0 and H_A. These two hypotheses are mutually exclusive, so one is true to the exclusion of the other.

Definitions

- Null Hypothesis H_0: States the assumption to be tested.
  e.g. There is no difference between the performance of two machine learning algorithms.
- *p*-Value: The probability of obtaining a result equal to or "more extreme" than what was actually observed, assuming that the hypothesis H_0 is true.
- *Type I error*: Rejecting H_0 when it is in fact true.
  i.e. “false alarm” - detecting a difference, when none exists.
- *Type II error*: Failing to reject H_0 when it is in fact false.
  i.e. concluding there is no difference, when there is.
- *Power of a test*: The potential of a statistical test to correctly reject a false null hypothesis H_0 (i.e. not commit a Type II error).

Type I and Type II Errors

- *Type I error*: Rejecting H_0 when it is in fact true.
  i.e. "false alarm" - detecting a difference, when none exists.
- *Type II error*: Failing to reject H_0 when it is in fact false.
  i.e. concluding there is no difference, when there is.

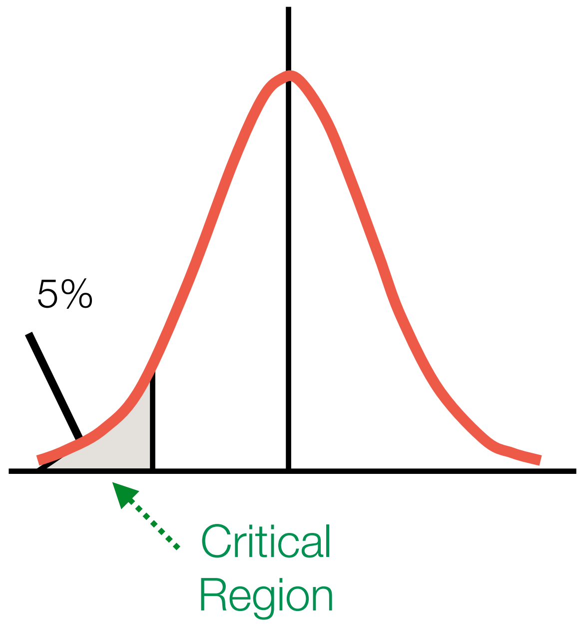
Statistical Test Result

Real World	H_0 Rejected	H_0 Not Rejected
There is a real difference	**Correct** A Hit	**Type II Error** Missed a real difference
There is in fact no difference	**Type I Error** False alarm	**Correct** Right to be sceptical of H_A

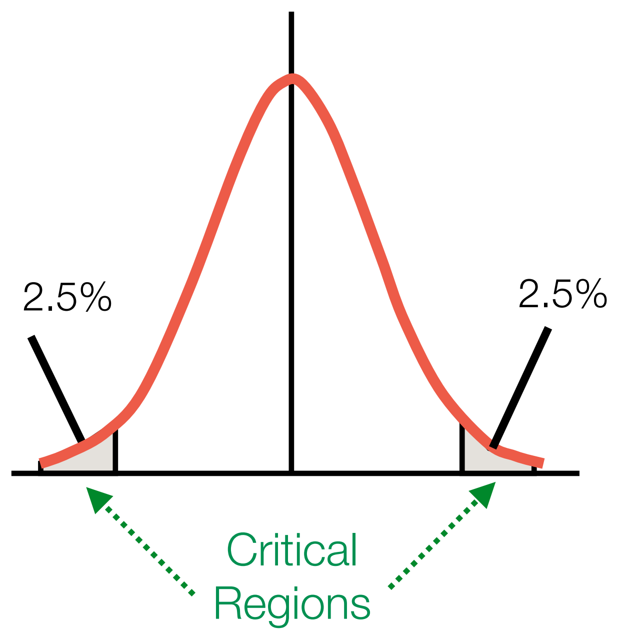
Two Tail vs One Tail

Before testing, we need to decide if we are interested in a *one-tailed* or a *two-tailed* statistical test.

- One-tailed: We decide in advance of looking at the data that one *mean value* will be larger than the other.
  e.g. "Did a generic drug work <u>better than</u> a brand name drug?"



- Two-tailed: We have no strong belief on whether the sample mean is likely to be higher or lower than the mean in the null hypothesis.
  e.g. "Did a generic drug work <u>better than or worse than</u> a brand name drug?"



P-Value Testing

General Approach for Testing

1. Calculate a test statistic on the sample data that is relevant to the hypothesis being examined.
2. Convert the result to a *p*-value by comparing its value to the distribution of test statistics under the null hypothesis.
3. Decide, for a specific level of significance, if we should reject or not reject the null hypothesis, based on the *p*-value:

$$p \leq \alpha \implies \text{reject } H_0 \text{ at level } \alpha$$
$$p > \alpha \implies \text{do not reject } H_0 \text{ at level } \alpha$$

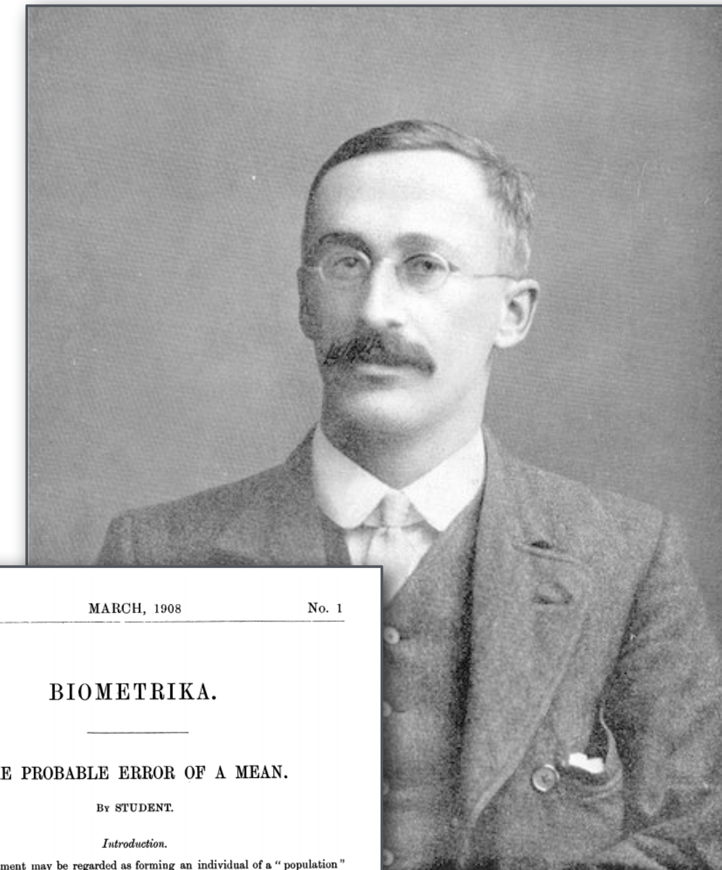
"Is it low enough to be significant?"

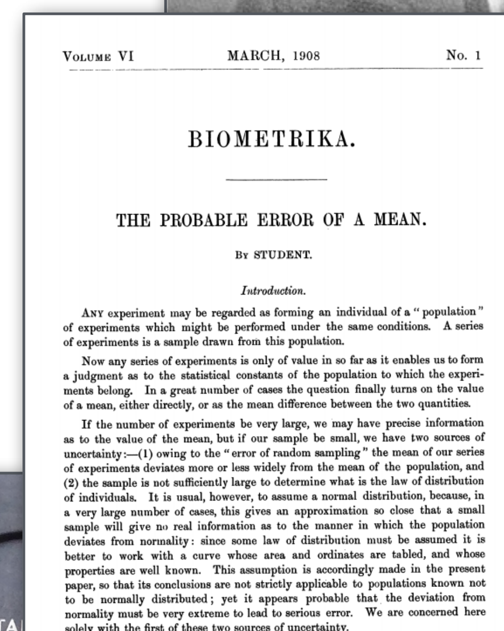
- The actual *p*-value threshold (α) depends on the problem, but 0.05 or 0.01 are often chosen "by default".
- The choice controls the Type I Error rate: "How serious is it to believe that something is true when it is in fact false?"

Student’s t-Test

- William Sealy Gosset - an English statistician who was employed as a chemist by Arthur Guinness & Son in Dublin.
- Wrote papers in his spare time under the pen name “Student”.
- Most noteworthy achievement is called Student’s t-test (1908), designed to compare small samples from quality control experiments in brewing.

➡ Are the means of two groups statistically different from each other?





Volume VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

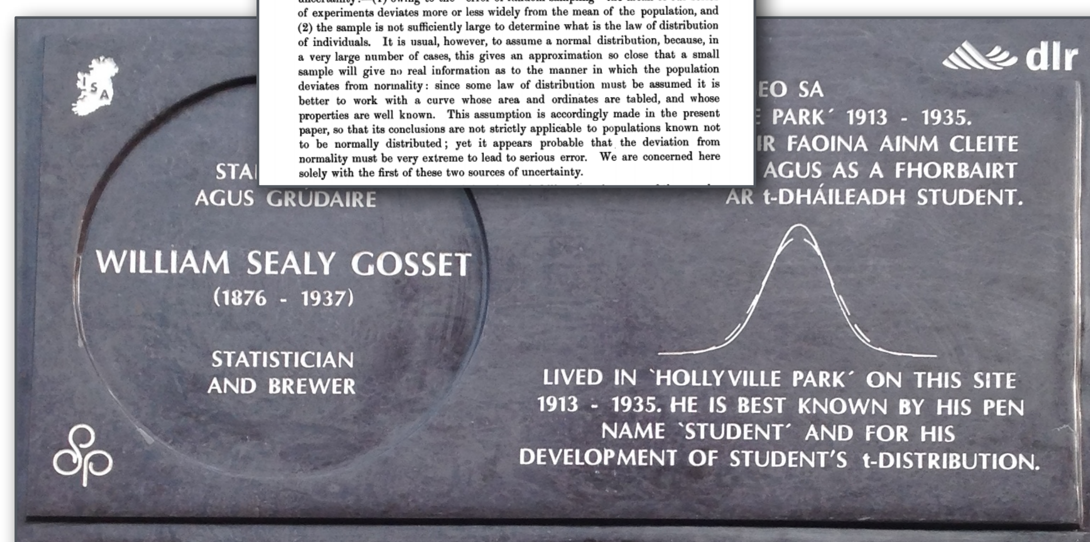
By STUDENT.

Introduction.

Any experiment may be regarded as forming an individual of a “population” of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty:—(1) owing to the “error of random sampling” the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in a very large number of cases, this gives an approximation so close that a small sample will give no real information as to the manner in which the population deviates from normality: since some law of distribution must be assumed it is better to work with a curve whose area and ordinates are tabled, and whose properties are well known. This assumption is accordingly made in the present paper, so that its conclusions are not strictly applicable to populations known not to be normally distributed; yet it appears probable that the deviation from normality must be very extreme to lead to serious error. We are concerned here solely with the first of these two sources of uncertainty.



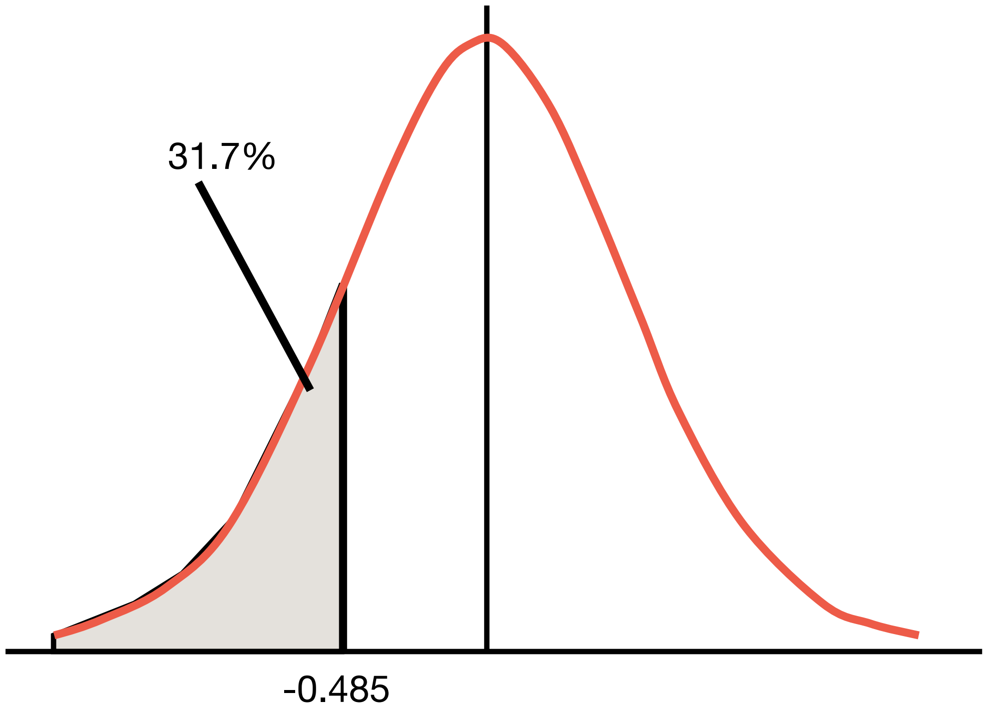
Student's t-Test

- Comparing scores for 2 teams. Is Team A <u>better than</u> Team B?

Team A	Team B
23	26
12	15
14	17
54	57
34	45
12	15
9	12
9	18
18	9
21	24

	Team A	Team B
N	10	10
Mean	20.600	23.800
Std Dev	14.017	15.455
Variance	196.489	238.844

Test statistic	t = -0.4850
P-Value	P(T≤t) one tail = 0.317



- For a given *t*-statistic value you can look up the confidence.
- There is a 31.7% chance that this difference is due to chance (according to this test).

➡ Difference between Team A and Team B is unlikely to be statistically significant.

Student's t-Test

- More observations and/or greater difference more likely to give statistical significance.

Team A	Team B
23	29
12	20
14	17
23	26
34	45
12	15
9	12
9	18
18	9
21	24
12	15
12	15
14	17
33	36
34	45
12	15
9	12
9	18
18	21
12	15

	Team A	Team B
N	20	20
Mean	17.000	21.200
Std Dev	8.423	10.288
Variance	70.947	105.853

Test statistic	$t = -1.413$
P-Value	$P(T \leq t)$ one tail = 0.083

➡ There is now a 8.3% chance that this difference is due to chance (according to this test).

Paired t-Tests

- Scores can be paired. e.g. Compare results achieved against the same teams: *Team A v Team C* & *Team B v Team C*
- Interested in the differences between each pair of scores.
- With paired data statistical significance can be determined using fewer observations.

Team A	Team B	Delta
23	26	-3
12	15	-3
14	17	-3
54	57	-3
34	45	-11
12	15	-3
9	12	-3
9	18	-9
18	9	9
21	24	-3

	Team A	Team B
N	10	10
Mean	20.600	23.800
Std Dev	14.017	15.455
Variance	196.489	238.844

Test statistic	$t = -1.945$
P-Value	$P(T \leq t)$ one tail = 0.042

➡ Lower P-value. We can now say with 95% confidence that Team B are better than Team A.

Student's t-Test: Formulae

- How are t-statistics calculated?
- Two unpaired samples, *A* and *B*:

$$t = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{\frac{var(A)}{n_A} + \frac{var(B)}{n_B}}}$$

Notation:

$\overline{X}_A$	Mean of sample A
$\overline{X}_B$	Mean of sample B
$var(A)$	Variance of sample A
$var(B)$	Variance of sample B
n_A	Number of observations in A
n_B	Number of observations in B

- What about paired data?
- Two paired samples, *A* and *B*:

$$t = \frac{\overline{X}_D \times \sqrt{n}}{\sigma_D}$$

Notation:

D	Difference in pairs in A and B
$\overline{X}_D$	Mean of differences D
σ_D	Standard Dev of differences B
n	Number of observations

Example: Unpaired t-Test

- Is Team A better than Team B, based on unpaired results?

Team A	Team B
23	26
12	15
14	17
54	57
34	45
12	15
9	12
9	18
18	9
21	24

	Team A	Team B
N	10	10
Mean	20.600	23.800
Std Dev	14.017	15.455
Variance	196.489	238.844

$$t = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{\frac{var(A)}{n_A} + \frac{var(B)}{n_B}}}$$

$$t = \frac{20.6 - 23.8}{\sqrt{\frac{196.489}{10} + \frac{238.844}{10}}}$$

$$t = -0.4850$$

Apply a one-tailed-test

Test statistic	t = -0.4850
P-Value	P(T≤t) one tail = 0.317

Example: Paired t-Test

- Is Team A better than Team B, based on paired results?

Team A	Team B	Delta
23	26	-3
12	15	-3
14	17	-3
54	57	-3
34	45	-11
12	15	-3
9	12	-3
9	18	-9
18	9	9
21	24	-3

Observations (n)	10
Mean of differences (deltas)	-3.2
Std Dev of differences (deltas)	5.20

$$t = \frac{\overline{X}_D \times \sqrt{n}}{\sigma_D}$$

Look at the mean and standard deviation of the differences (deltas)

$$t = \frac{-3.2 \times \sqrt{10}}{5.2} = -1.946$$

Test statistic	t = -1.946
P-Value	P(T≤t) = 0.084

- Paired t-tests are often used for comparing classifiers if multiple test sets are available, and also in cross validation experiments.

Testing - Implementations

- Many libraries and packages are available for hypothesis testing e.g. SciPy for Python, Apache Commons Math for Java

Standard t-tests (two tail):

```
>>> from scipy import stats
>>> a = [23,12,14,23,34,12,9,9,18,21,12,12,14,33,34,12,9,9,18,12]
>>> b = [29,20,17,26,45,15,12,18,9,24,15,15,17,36,45,15,12,18,21,15]
>>> t, pvalue = stats.ttest_ind(a,b)
>>> print "The t-statistic is %.3f and the p-value is %.3f." % (t,pvalue)
The t-statistic is -1.413 and the p-value is 0.166.
```

Paired t-tests (two tail):

```
>>> from scipy import stats
>>> a = [23, 12, 14, 54, 34, 12, 9, 9, 18, 21]
>>> b = [26, 15, 17, 57, 45, 15, 12, 18, 9, 24]
>>> t, pvalue = stats.ttest_rel(a,b)
>>> print "The paired t-statistic is %.3f and the p-value is %.3f." % (t,pvalue)
The paired t-statistic is -1.945 and the p-value is 0.084.
```

Difference in Proportions

- A *t*-test is sometimes used to analyse differences in proportions e.g. comparison of conversion rates in A/B testing.
- Requires a number of assumptions about the population which are usually not true.

	Control	Treatment
Samples	n_1	n_2
Conversions	c_1	c_2

$$p = \frac{c1 + c2}{n1 + n2} \qquad p_1 = \frac{c_1}{n_1} \qquad p_2 = \frac{c_2}{n_2}$$

$$t \text{ statistic} = \frac{\text{Difference in proportions}}{\text{Standard error}}$$

$$t = \frac{p1 - p2}{\sqrt{p(1-p) \times (\frac{1}{n_1} + \frac{1}{n_2})}}$$

http://stattrek.com/hypothesis-test/difference-in-proportions.aspx

McNemar's Test

- Measure for comparing paired proportions.
  e.g. Which is better, classifier C2 or C3?
- Applied to 2x2 contingency tables.
- Test captures two key differences:

  n_{01}: number misclassified by 1st but not 2nd classifier.

  n_{10}: number misclassified by 2nd but not 1st classifier.

C1	C2	C3
✓	✓	✗
✓	✓	✓
✓	✓	✗
✓	✓	✓
✓	✓	✗
✗	✓	✗
✗	✓	✓
✗	✗	✗
✗	✗	✓
✗	✗	✓

Contingency for C2 v C1

3 n_{00}	2 n_{01}
0 n_{10}	5 n_{11}

McNemar C2 v C1

$$\chi^2 = 1/2 = 0.5$$

Contingency for C3 v C1

1 n_{00}	2 n_{01}
4 n_{10}	3 n_{11}

McNemar C3 v C1

$$\chi^2 = 1/6 = 0.1666$$

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$$

Note: For test to be applicable require $(n_{01}+n_{10}) > 10$

➡ $\chi^2 > 3.84$ required for statistical significance at 95%. So neither classifier significantly better!

Summary

- Objectives of Evaluation
- A/B Testing
- Hypothesis Testing
  - Student’s t-test
  - t-Test for paired data
  - Differences in proportions
  - McNemar’s test for proportions
- Next: Evaluation measures and setup for classification