Behavioral Data Mining

Lecture 3
Measurement

Rhine Paradox*

Joseph Rhine was a parapsychologist in the 1950's (founder of the *Journal of Parapsychology* and the *Parapsychological Society, an affiliate of the AAAS*).

He ran an experiment where subjects had to guess whether 10 hidden cards were red or blue.

He found that about 1 person in 1000 had ESP, i.e. they could guess the color of all 10 cards.

* Example from Jeff Ullman/Anand Rajaraman

Rhine Paradox

He called back the "psychic" subjects and had them do the same test again. They all failed.

He concluded that the act of telling psychics that they have psychic abilities causes them to lose it…(!)

This is funny, but less so because something similar happens almost every day in the experimental sciences.

Significance and Publication Bias

Most scientific tests are probabilistic (due to measurement error, sampling etc), and are only true to some "significance" value, which is the false positive rate, e.g. 0.05 or 1/20.

Both experimenters and journals have a strong bias to publish only results that "succeed", i.e. which were significant at 0.05

Suppose 20 experimenters try to show "coffee causes dementia"

Significance and Publication Bias

Assuming the hypothesis is false, nevertheless, on average about 1/20 experimenters get a positive result by chance.

The other 19 experimenters will probably never submit their results, nor the fact that they tried the experiment. If they did, unless there were other results in the paper, most journals would not publish it.

The one lucky experimenter will certainly want to publish and will probably succeed.

The research community concludes the hypothesis holds.

Significance and Publication Bias

No-one really knows how strong the effects of publication bias are*.

Its also difficult to measure because of a second publication and funding bias against replication of scientific results.

Workarounds: Documenting Trials: Many medical journals require public registration of trials before they will publish the results. That gives a more accurate count of tries/successes.

* May be as high as 10:1 in some fields.

But…

The New York Times

Research

Search All NYTimes.com

Go

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION | ARTS | STYLE | TRAVEL | JOBS | REAL ESTATE | AUTOS

RESEARCH | FITNESS & NUTRITION | MONEY & POLICY | VIEWS | HEALTH GUIDE

Search Health 3,000+ Topics

Go

For First Time, AIDS Vaccine Shows Some Success

By DONALD G. MCNEIL Jr.
Published: September 24, 2009

Scientists said Thursday that a new AIDS vaccine, the first ever declared to protect a significant minority of humans against the disease, would be studied to answer two fundamental questions: why it worked in some people but not in others, and why those infected despite vaccination got no benefit at all.

The vaccine — known as RV 144, a combination of two genetically engineered vaccines, neither of which

SIGN IN TO RECOMMEND

TWITTER

COMMENTS (33)

SIGN IN TO E-MAIL

PRINT

REPRINTS

SHARE

SHARE

Well
Tara Parker-Pope on Health

Tips for Navigating Medicare
October 16, 2009

Show Off Your Vegetables With Pasta
October 16, 2009

High-Deductible Health Plans: Better for You or Your Employer?
October 16, 2009

The Roving Runner: Prospect Park
October 16, 2009

Alternative Medicine Cabinet: Thyme for Toenails
October 15, 2009

TicketWatch - Theater Offers by E-Mail

Sign up for ticket offers from Broadway shows

Maintaining Significance

Bonferroni Discounting: A simple but conservative approach when running k tests is to reduce the significance threshold for each test to 0.05/k. This guarantees that the probability of any the reported results occurring by chance will be less than 0.05.

You can also partition the total significance into unequal-sized bins applied to different tests. E.g. if you have a test that you think is important but may not succeed, you could assign 0.025 out of 0.05 to it, and distribute the remaining 0.025 among other tests. All of this should happen before you run any experiments.

Why you should care

Data mining technologies and large datasets make it incredibly easy to ask questions and test hypotheses. i.e. you can effectively run hundreds of experiments in a few hours.

You will get many "positives" by chance.

Its very important to know what the false positive rate is, and whether a result you see is really "unusual" relative to chance.

Report everything you tried, not just successes.

Testing Hypotheses

- Construct a hypothesis, e.g. "gamers have less active social networks than non-gamers."
- Define an experiment to test the hypothesis (modify H if needed).
- Choose a population and sampling method.
- Create a null hypothesis H0.
- Construct a test statistic.
- Choose a significance level, sample size.
- Run the experiment.
- Report all the results.

Testing Hypotheses

- Construct a hypothesis, e.g. "gamers have less active social networks than non-gamers."
- Define an experiment to test the hypothesis (modify H if needed).
  - Use wall posts on Facebook as measure of social activity.
  - Define "gamer" via self-reported weekly playing time.
  - Decide on control variables (age, gender, education,…)

Testing Hypotheses

- Construct a hypothesis, e.g. "gamers have less active social networks than non-gamers."
- …
- Choose a population and sampling method.
  - Facebook users
  - Recruitment: direct through email, through a snowball sample, piggyback a popular app, or your own app,

Testing Hypotheses

- Construct a hypothesis, e.g. "gamers have less active social networks than non-gamers."
- …
- Create a null hypothesis H0:
  - "gamers have identical normal distributions of activity in their social networks as non-gamers" (parametric) OR
  - "gamers have equal mean activity in their social networks as non-gamers" (non-parametric)

A null hypothesis allows you to reason "forward" to a "weak contradiction" (an unlikely outcome). Its rather like a mathematical proof by contradiction.

Testing Hypotheses

- Construct a hypothesis, e.g. "gamers have less active social networks than non-gamers."
- …
- Construct a test statistic. This usually follows a recipe and gives you a provably most-powerful test. E.g. for distinguishing means you would use:

$$S = \text{mean}(V_A) - \text{mean}(V_B)$$

Where V_A and V_B are the measurements of social activity for sets A (gamers) and B (non-gamers).

Sometimes median difference will give better results (less sensitive to outliers).

Testing Hypotheses

- Construct a hypothesis, e.g. "gamers have less active social networks than non-gamers."
- …
- Choose a significance level, sample size.
  - Usually its 0.05. Sample size is often hard to know without running the experiment. Try a pilot study. For public data though (which is abundant), this is less of an issue.
- Run the experiment.
  - In DM that means collect and process the data
- Report all the results.
  - Helps minimize the effects of publication bias.

Errors

- Type I: False Positive. Experimenter rejects the null hypothesis (supports the original hypothesis) when it is not true, i.e. when the null hypothesis actually holds.

These are the most serious because results that are not true become part of the scientific record.

- Type II: False Negative. Experimenter does not reject the null hypothesis even though it does not hold.

Arguable that this is an "error." Provides no real evidence against the original hypothesis. Often the sample size was just too small. Minimal effect on the scientific record (may not get reported).

Normal Distributions, Mean, Variance

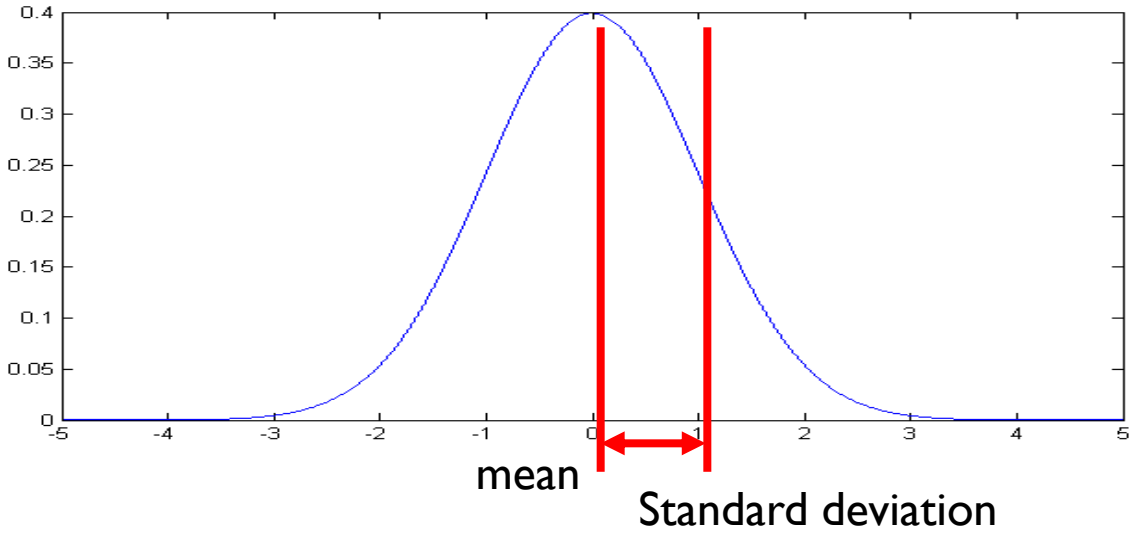
The mean of a set of values is just the average of the values.

Variance a measure of the width of a distribution. Specifically, the variance is the mean squared deviation of samples from the sample mean:

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

The standard deviation is the square root of variance.

The normal distribution is completed characterized by mean and variance.

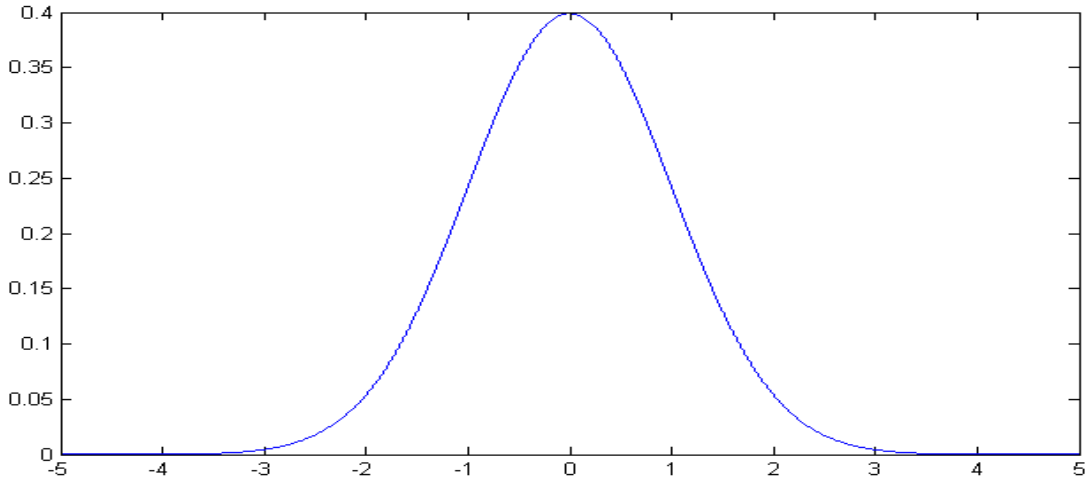


Central Limit Theorem

The distribution of the sum (or mean) of a set of n identically-distributed random variables Xi approaches a normal distribution as $n \rightarrow \infty$.

The common parametric statistical tests, like t-test and ANOVA assume normally-distributed data, but depend on sample mean and variance measures of the data.

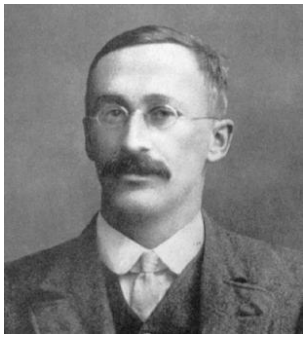
They typically work reasonably well for data that are not normally distributed as long as the samples are not too small.



First parametric test: t-test

The t-statistic was invented by William Sealy Gosset, A Chemist working for Guinness Breweries, in 1908.

Gosset had to publish under a pseudonym, for which he chose "Student", hence "Student's t-test"



The t-test compares means and variances of two samples. Data are assumed to be normal and independent and identically-distributed within each sample.

T-test

Assuming equal variance: $t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

where $S_{X_1X_2} = \sqrt{\frac{(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2}{n_1 + n_2 - 2}}.$

Assuming unequal variance: $t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$

where $s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$

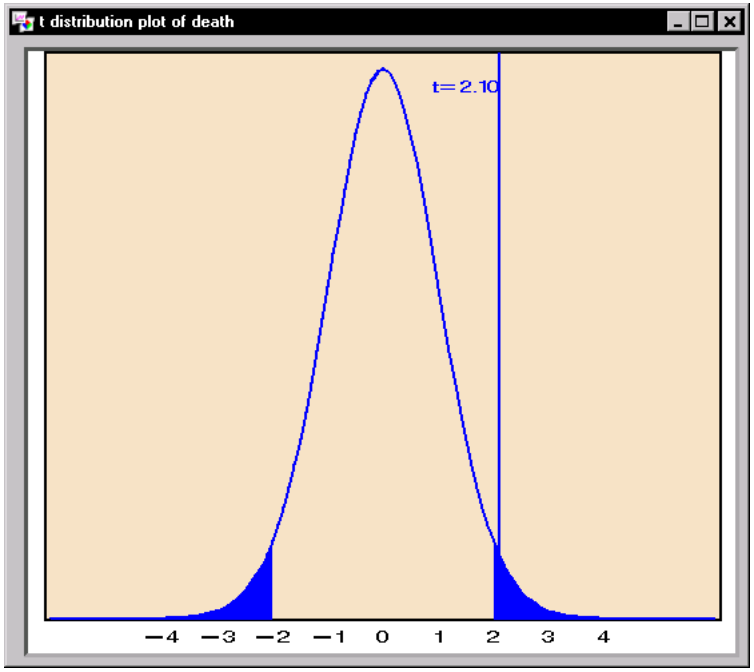
And S_i^2 is an unbiased variance estimator. (Check d.o.f.)

t-statistic and t-distribution

The t-statistic we just described, when applied to datasets that satisfy all the constraints (IID, normal, H0) will have a t-distribution.

From the distribution we can ask how "unusual" was the measurement.

A reasonable way to do this is to consider all values "at least as extreme" of the test statistic.

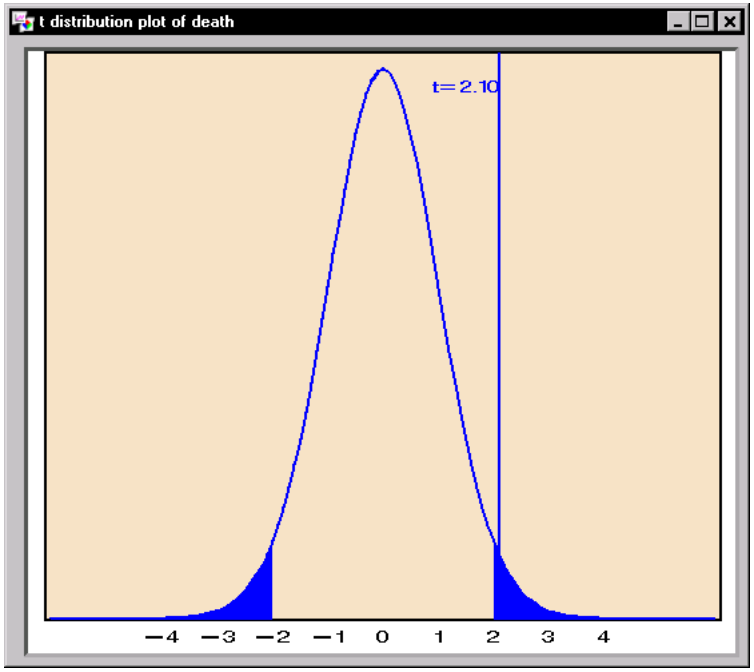


t-statistic and t-distribution

The cumulative probability of a statistic value greater than or equal to the measured value (in one or both tails of the t-distribution) is the p-value.

This is the estimated false positive probability of the test.

If p is less than the significance (say 0.05), we declare that the test succeeded and reject the null hypothesis.



One-Way ANOVA

ANOVA (ANalysis Of VAriance) allows testing of multiple differences in a single test. Suppose our experiment design has an independent variable Y with four levels:

Y

Primary School	High School	College	Grad degree
4.1	4.5	4.2	3.8

The table shows the mean values of a response variable (e.g. avg number of Facebook posts per day) in each group.

We would like to know in a single test whether the response variable depends on Y, at some particular significance such as 0.05.

ANOVA

Doing this with t-tests requires all paired (6) tests, and with Bonferroni discounting we would have to set the significance thresholds to 0.05/6 = 0.008.

In ANOVA we compute a single statistic (an F-statistic) that compares variance between groups with variance within each group.

$$F = \frac{VAR_{between}}{VAR_{within}}$$

The higher the F-value is, the less probable is the null hypothesis that the samples all come from the same population.

F-statistic

$$F = \frac{VAR_{between}}{VAR_{within}}$$

$VAR_{between}$ measures the variance across groups, using the mean of each group as a sample.

VAR_{within} is simply an average of the variances within each group.

The F-distribution depends on the degrees of freedom of numerator and denominator, and these should always be quoted when reporting F-values. Normally the d.o.f. is one less than the number of squares in each sum.

e.g. F(3,28) = 3.4, p = 0.033

Post-hoc tests

If the F-test is significant, ANOVA analysis is normally followed by individual paired t-tests called post-hoc tests.

Those tests establish the relative order of group means.

Primary School	High School	College	Grad degree
4.1	4.5	4.2	3.8

T-tests shouldn't be done if the F-test fails – any results are likely to be due to noise.

Factorial ANOVA

Factorial ANOVA deals with several independent variables.

Suppose our experiment design has variables X with two levels and Y with three levels:

X	Y: High-School	Y: College	Y: Grad degree
Male	4.5	4.2	3.8
Female	5.1	4.8	4.0

There are now four dependencies that can be measured:

- All-independent-groups ("corrected model" in SPSS)
- Main effect of X
- Main effect of Y
- Interaction effect of XY

Factorial ANOVA

The first test (corrected model) should be done before all the others. If it fails, then other tests probably shouldn't be done – just as with post-hoc t-tests.

If the model error test is skipped, then Bonferroni correction needs to be applied to main effects and interactions.

Main effects and interactions on the same data are still distinct tests, and need to be corrected.

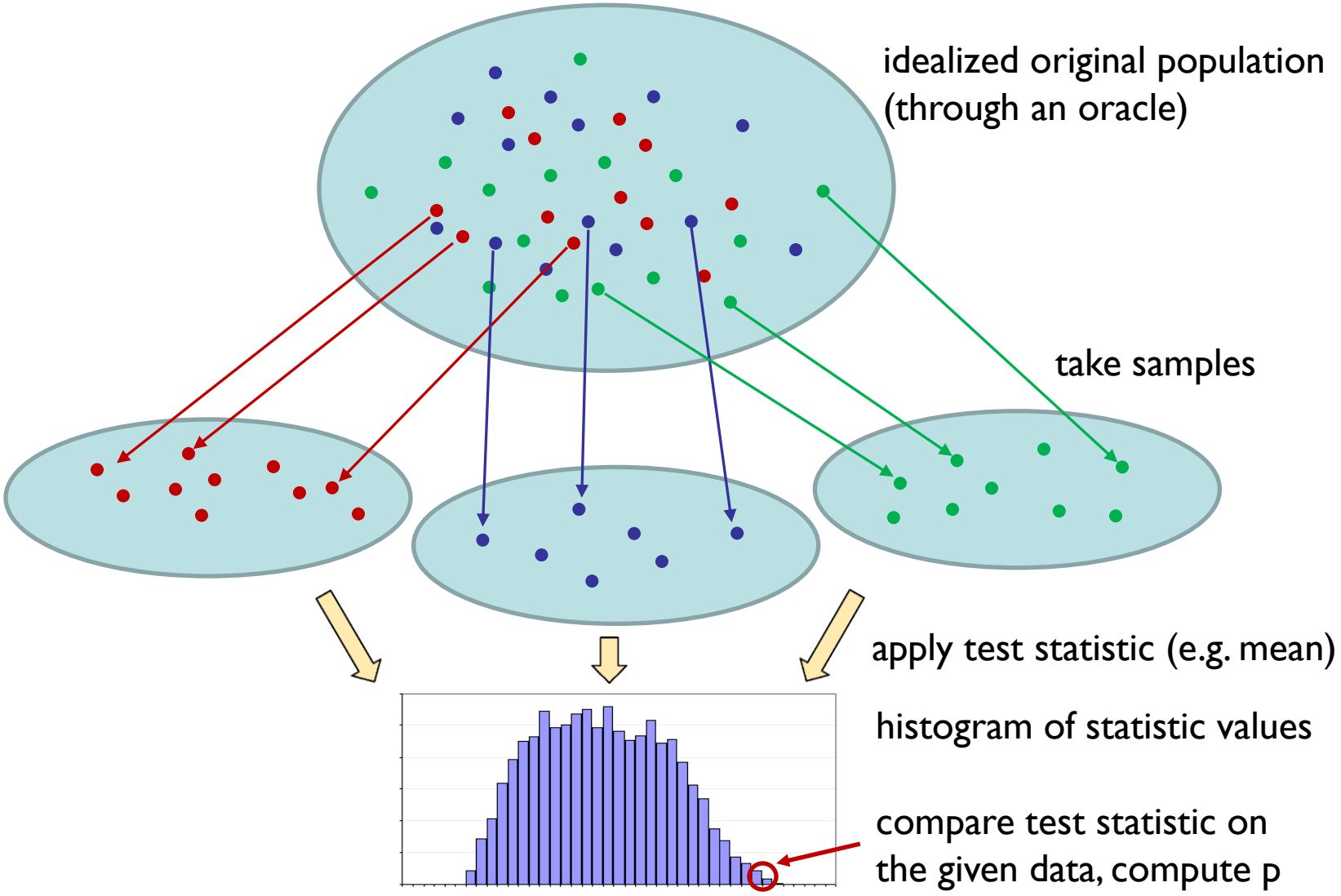
Non-parametric tests

Non-parametric tests use the data themselves to make inferences without assuming a probability distribution for the data.

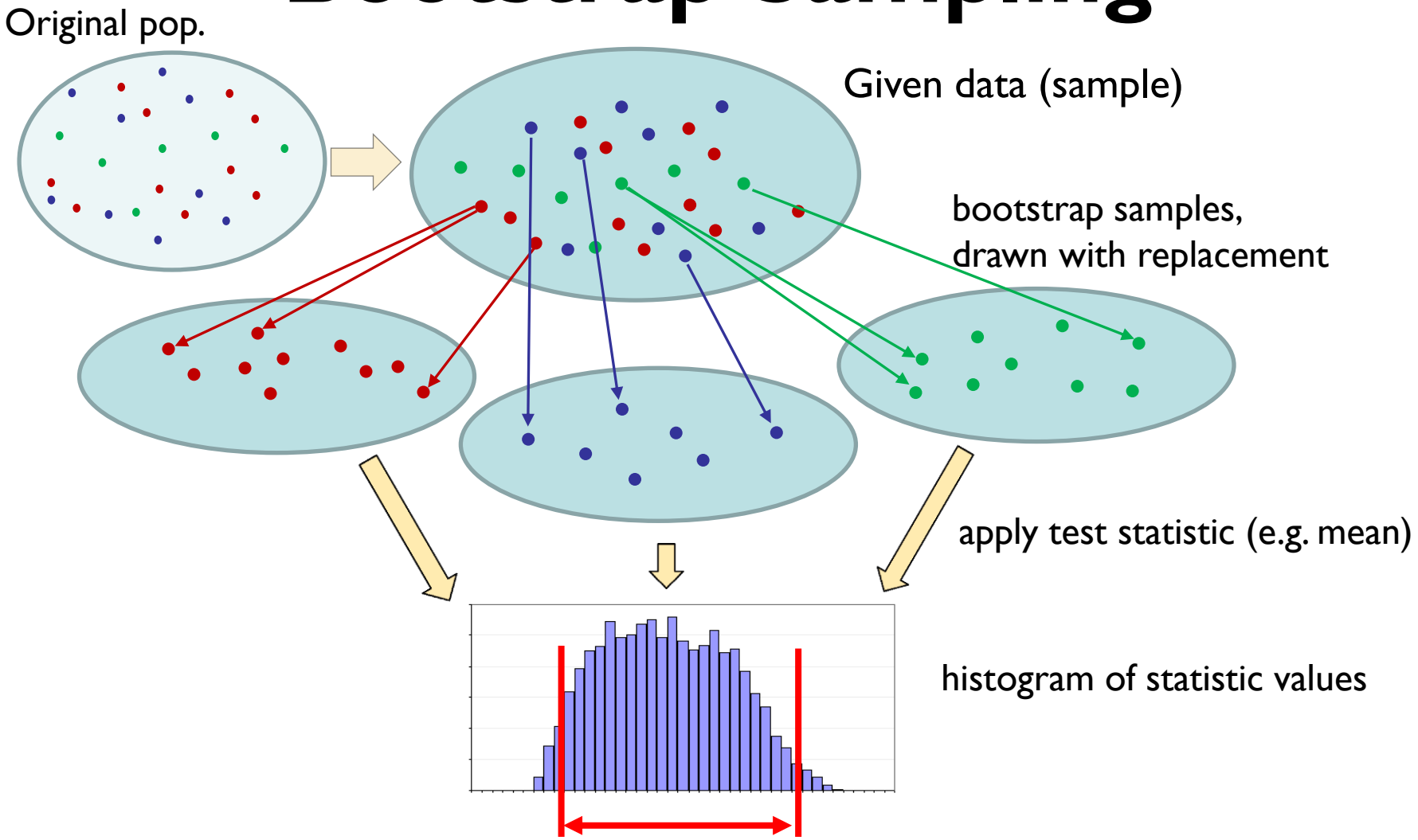
Two common kinds:

- Bootstrap: generate new samples by resampling the given data
- Permutation tests: mix data for different conditions as permitted under the assumptions of the null hypothesis.

Idealized Sampling



Bootstrap Sampling



The region containing 95% of the samples is a 95% confidence interval (CI)

Bootstrap CI tests

Then a test statistic outside the 95% CI would be considered significant at 0.05, and probably not drawn from the same population.

e.g. Suppose the data are differences in running times between two algorithms. If the 95% bootstrap CI does not contain zero, then original distribution probably has a mean other than zero, i.e. the running times are different.

We can also test for values other than zero. If the 95% CI contains only values greater than 2, we conclude that the difference in running times is significantly larger than 2.

Bootstrap CI tests

Simple bootstrap tests exhibit bias, but most tools (e.g. Matlab) provide "bias-corrected bootstrap" interval estimates. A related method called BCA* (Bias-Corrected Accelerated) is very popular.

Pros of bootstrap tests:

- Simple, fast, support complex statistics and/or distributions
- Support small samples
- Can be used to estimate power (how likely it is that another test will succeed)

Cons

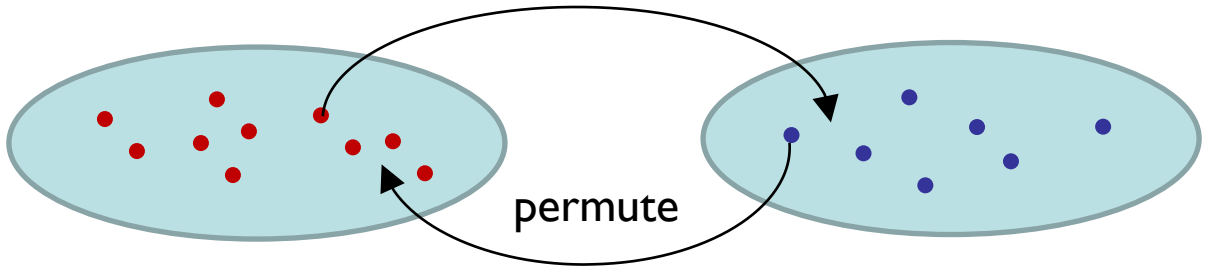
- Can be overly optimistic (type I errors).

* Efron, B. (1987). "Better Bootstrap Confidence Intervals". *Journal of the American Statistical Association*, (82) 397

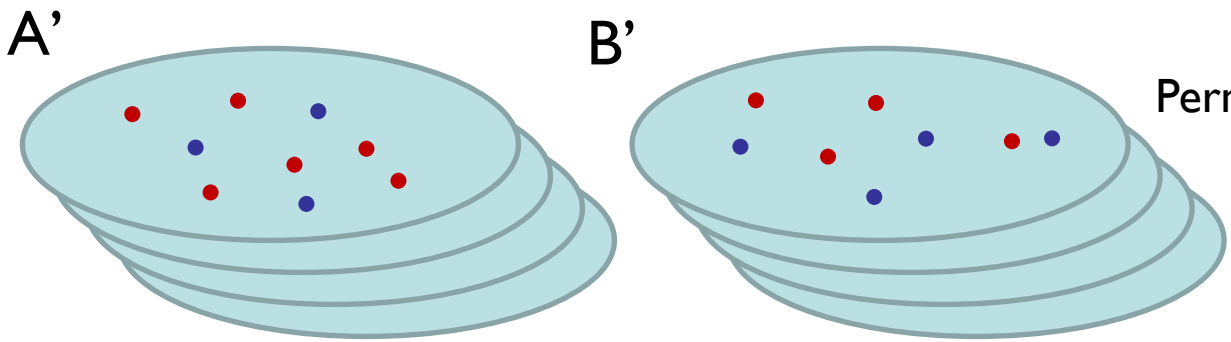
Permutation Tests

Another very simple, non-parametric hypothesis test is the permutation test, performed on two or more sets of data.

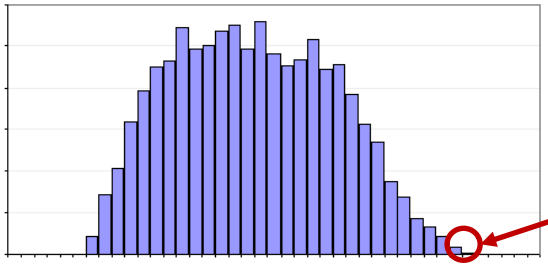
Under the null hypothesis, samples in A and B are exchangeable:



Test statistic = differences in means



Permutation samples A' B'



Evaluate test statistic on permutation samples
➔ Permutation distribution

compare test statistic on the original data, compute p

Monte-Carlo Permutation Tests

Enumerating all the permutations of two sets A, B is expensive. The total number of permutations is $(n_A + n_B)!/n_A!\, n_B! =$

$\binom{n_A + n_B}{n_B}$ which approaches $2^{(n_A+n_B)}$

For large $n = n_A+n_B$, we approximate the permutation distribution by Monte-Carlo sampling. i.e. by generating random permutations of (1,…,n) and using them to generate A', B' sets.

We construct the histogram of statistic values on these samples as before, and use it to estimate p for the given data.

Permutation Test Exactness

Permutation tests are generally "exact," meaning the actual false positive rate is exactly what the test predicts.

Parametric tests often rely on approximations, and at the very least "exactness" depends on the fit of the data to assumptions (IID normal input data).

They are often "most powerful" tests when data satisfy an assumed distribution (lowest false negative rate).

Permutation Tests

More complex permutation tests exist for many kinds of design (e.g. Factorial ANOVA, MANOVA, repeated measures).

They are not widely available in off-the-shelf tools, but code exists for Matlab and R *.

They have many virtues and only a few downsides:

- Computation time
- Sensitivity to random number artifacts

* "Permutation Tests for Complex Data" Pesarin and Salmaso, Wiley 2010.

Cluster update

iCluster memory/disk upgrade not complete yet

MarkLogic seems to have installed OK, but cluster config needs to wait on the RAM.

Precision and Recall

When evaluating a search tool or a classifier, we are interested in at least two performance measures:

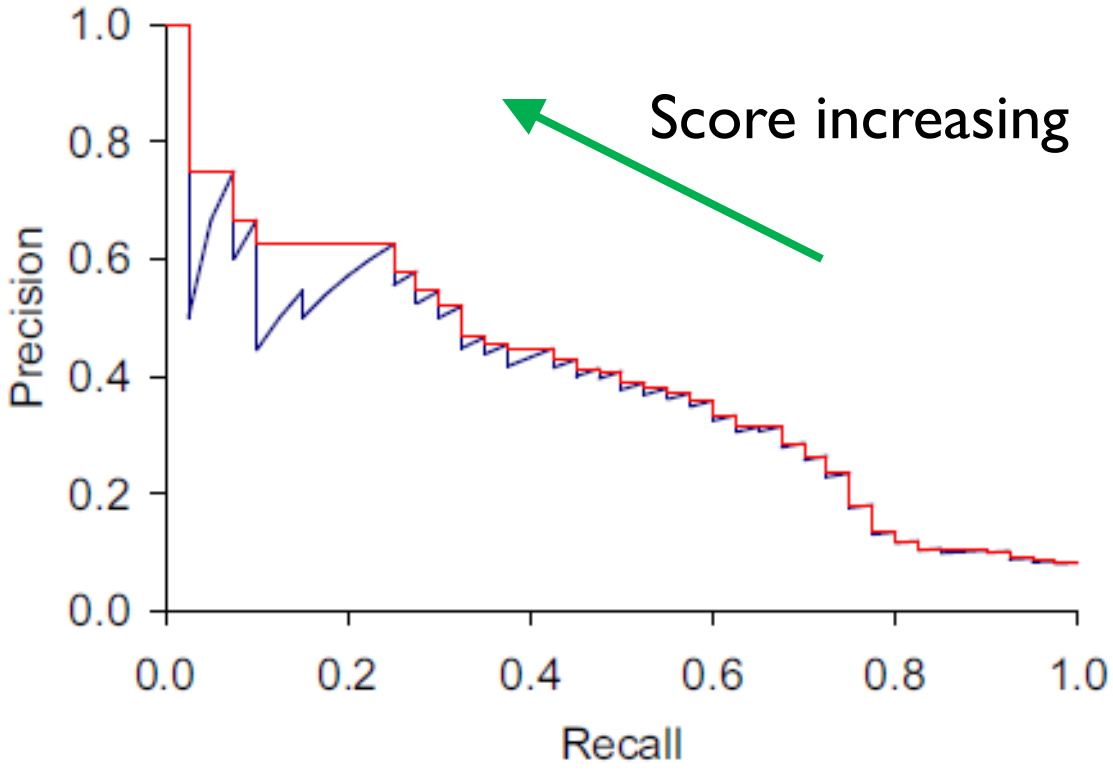
Precision: Within a given set of positively-labeled results, the fraction that were true positives = tp/(tp + fp)

Recall: Given a set of positively-labeled results, the fraction of all positives that were retrieved = tp/(tp + fn)

Positively-labeled means judged "relevant" by the search engine or labeling in the class by a classifier. tp = true positive, fp = false positive etc.

Precision and Recall

Search tools and classifiers normally assign scores to items.
Sorting by score gives us a precision-recall plot.



Why not to use “accuracy”

The simplest measure of performance would be the fraction of items that are correctly classified, or the “accuracy” which is:

$$\frac{tp + tn}{tp + tn + fp + fn}$$

But this measure is dominated by the larger set (of positives or negatives) and favors trivial classifiers.

e.g. if 5% of items are truly positive, then a classifier that always says “negative” is 95% accurate.

The weighted "F" measure

A measure that naturally combines precision and recall is the β-weighted F-measure:

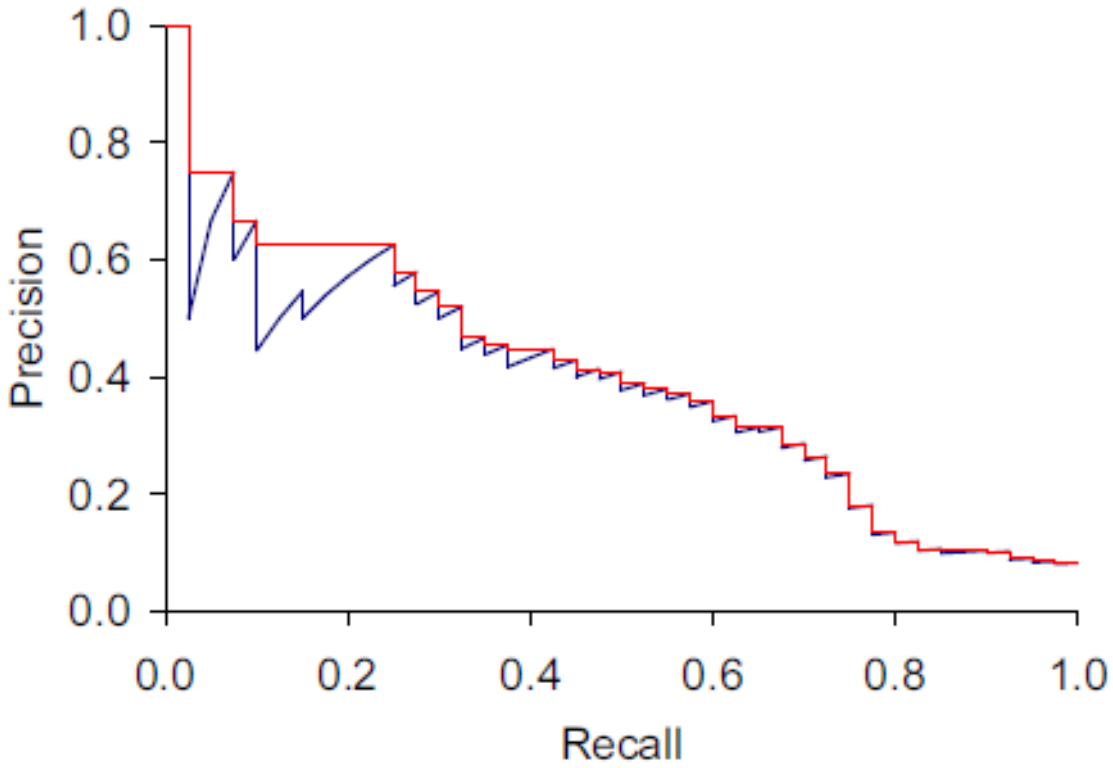
$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

Which is the weighted harmonic mean of precision and recall. Setting $\beta = 1$ gives us the F_1 – measure. It can also be computed as:

$$F_{\beta=1} = \frac{2PR}{P + R}$$

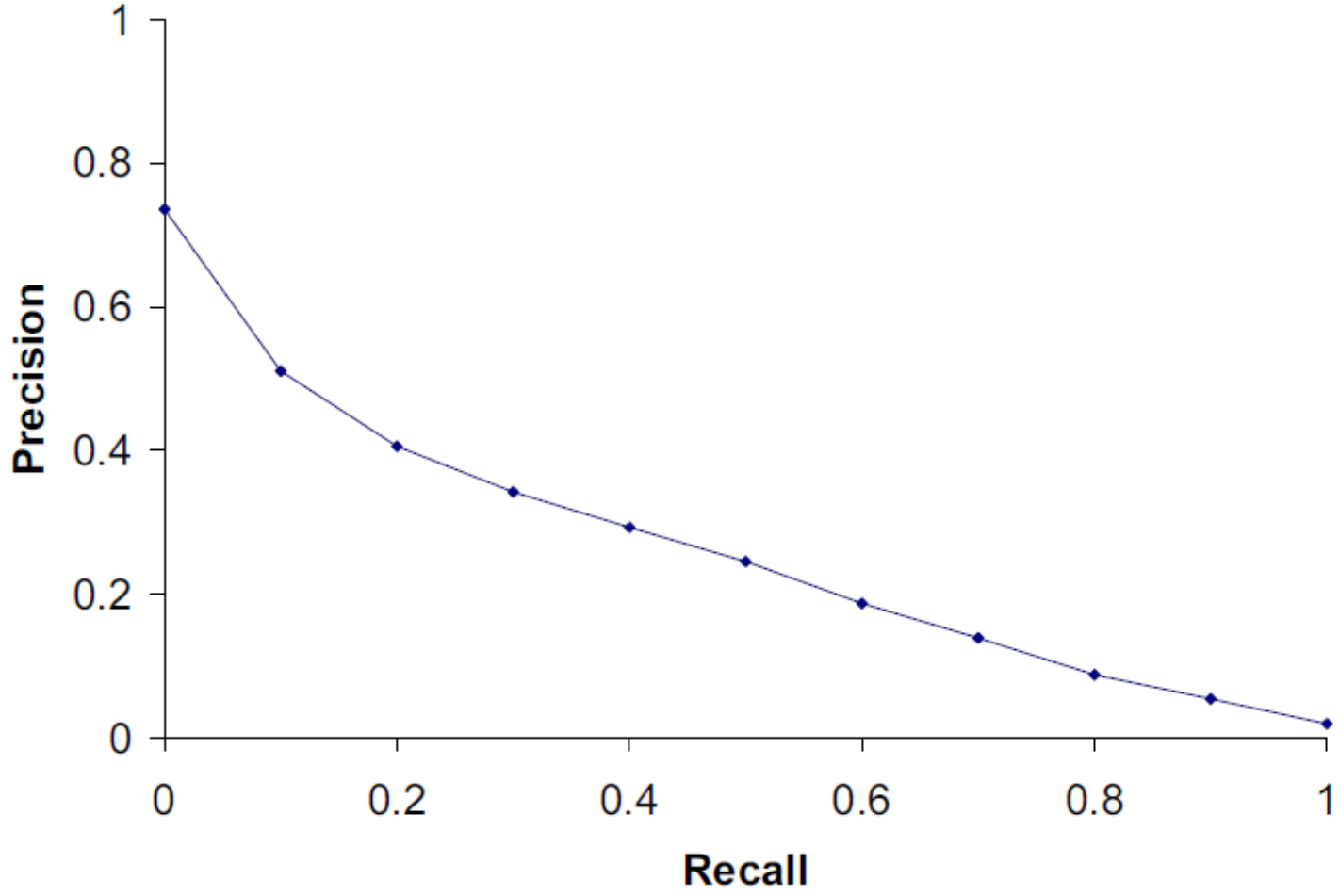
Interpolated Recall

The true precision plot (blue) necessarily dips at high precision each time a fp appears in the item ordering. These can be removed by using "interpolated precision" which is defined as the max precision at any recall value r' > the current r. An interpolated precision-recall curve is non-increasing.



TREC Precision-Recall plots

We compute the interpolated precision values at ten values of recall, 0.1, 0.2,… 1.0.

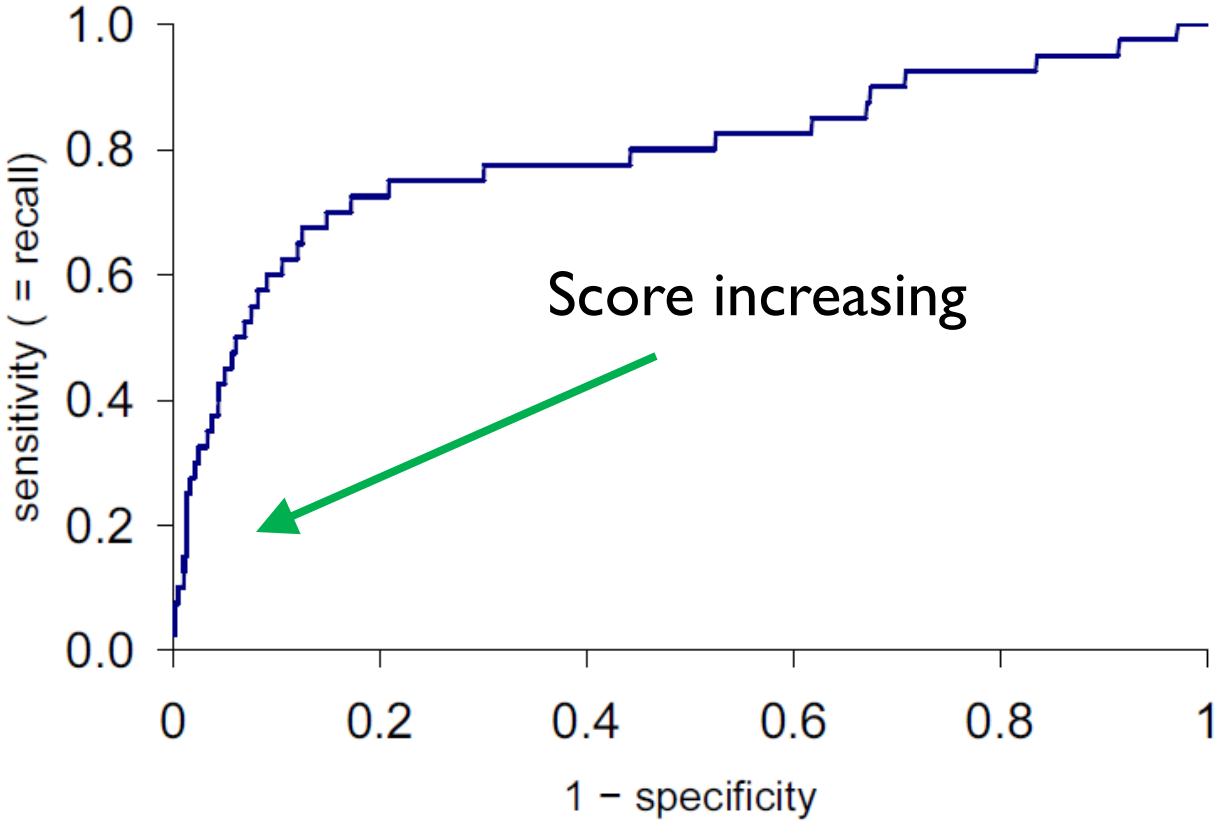


ROC plots

ROC is Receiver-Operating Characteristic. ROC plots

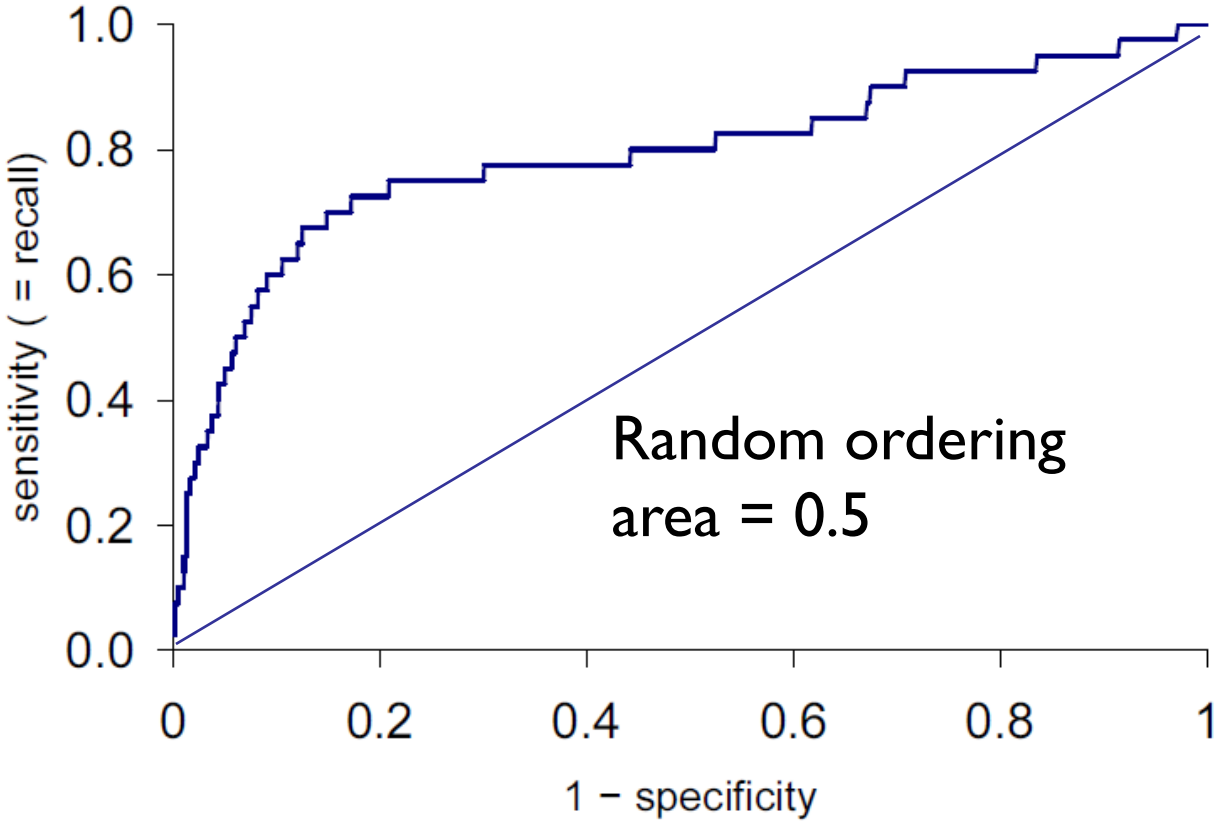
Y-axis: true positive rate = tp/(tp + fn), same as recall

X-axis: false positive rate = fp/(fp + tn) = 1 - specificity



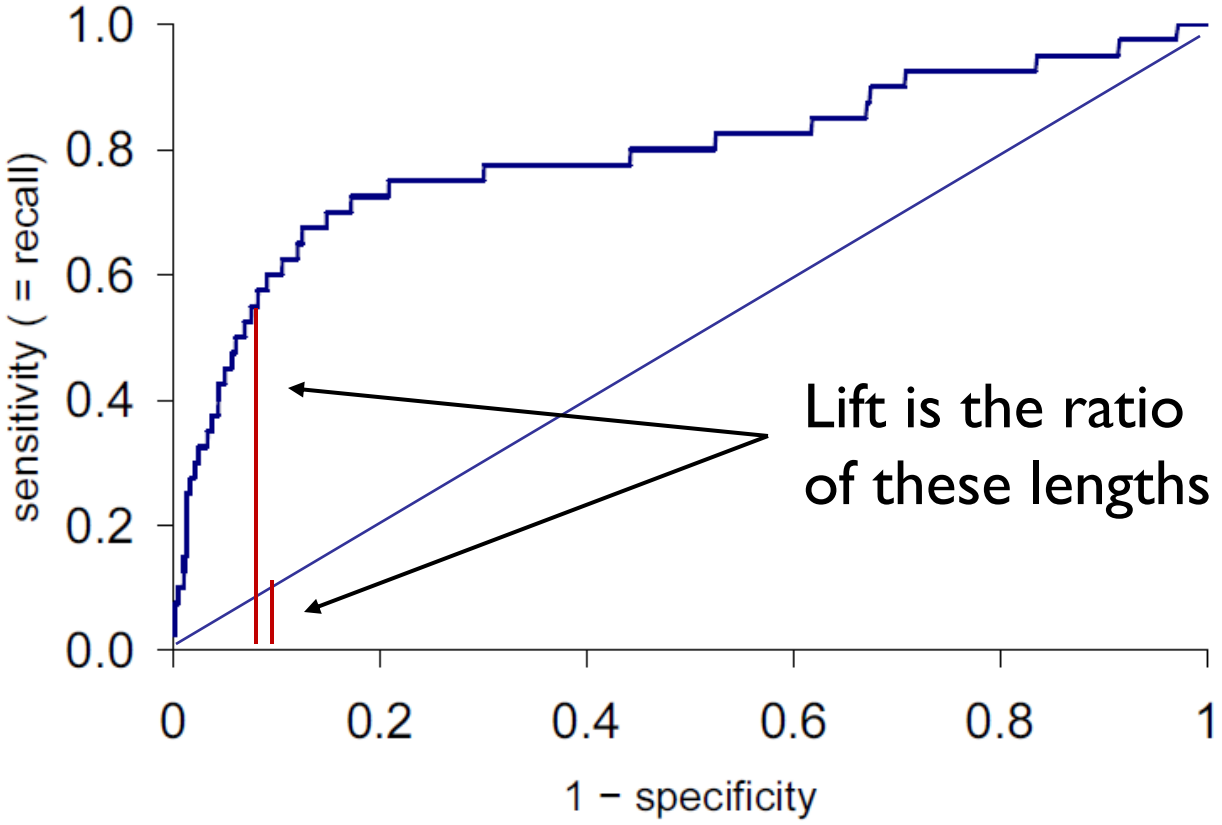
ROC AUC

ROC AUC is the "Area Under the Curve" – a single number that captures the overall quality of the classifier. It should be between 0.5 (random classifier) and 1.0 (perfect).



Lift Plot

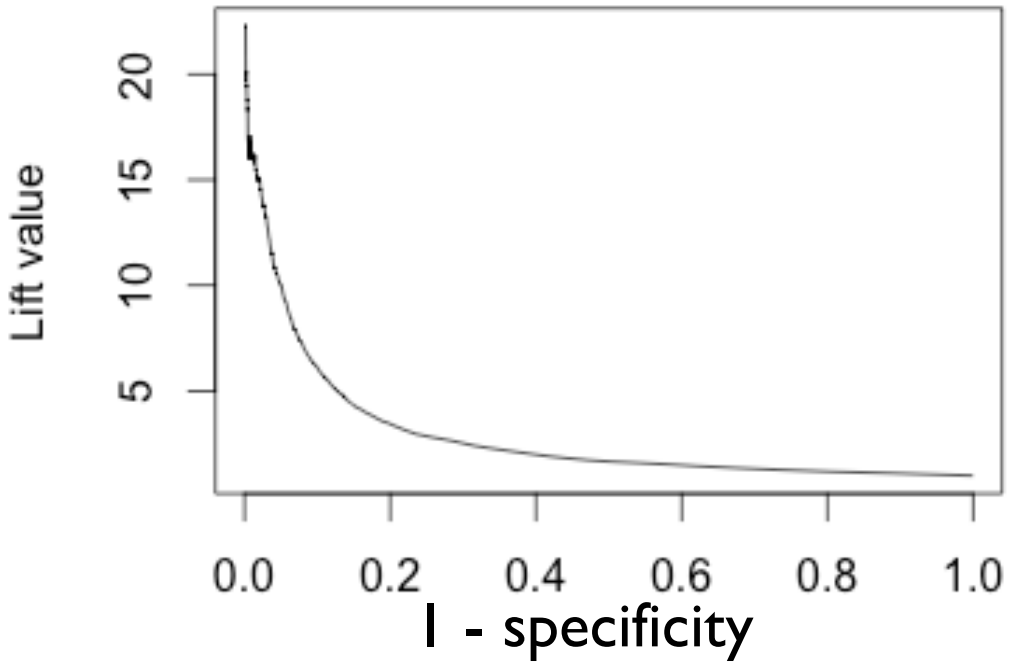
A derivative of the ROC plot is the lift plot, which compares the performance of the actual classifier/search engine against random ordering, or sometimes against another classifier.



Lift Plot

Lift plots emphasize initial precision (typically what you care about), and performance in a problem-independent way.

Note: The lift plot points should be computed at regular spacing, e.g. 1/00 or 1/1000. Otherwise the initial lift value can be excessively high, and unstable.



Summary

Hypothesis testing is easy and cheap to do with data mining, raising the risks of type-I errors.

Parametric tests (t-test, ANOVA) are simple and work well, even when they shouldn't.

ANOVA provides a way to test for several effects, without reducing significance.

Sampling methods such as the bootstrap and permutation tests deal with non-normal data at the expense of extra computation.

Precision-recall plots capture the performance of a search or classification tool across the ordering range.

ROC and Lift plots make it easy to compare performance across datasets and algorithms.