Statistical methods in research

Ben Rubinstein

School of Computing and Information Systems
The University of Melbourne

Research Methods; Semester 2, 2018

Outline

1. Introduction
   - Introducing myself
   - Is this CS or Stats?
2. Experiments and statistics
   - Why experiments in science and engineering
   - Why statistics in experiments
3. Randomization and random sampling
   - Populations and samples
   - Random sampling
4. Methods of standard error calculation
   - Assume normality
   - Bootstrap

Outline

1. Introduction
    - Introducing myself
    - Is this CS or Stats?
2. Experiments and statistics
    - Why experiments in science and engineering
    - Why statistics in experiments
3. Randomization and random sampling
    - Populations and samples
    - Random sampling
4. Methods of standard error calculation
    - Assume normality
    - Bootstrap

Who Am I?

- Ben Rubinstein
- brubinstein@unimelb.edu.au
- Office 7.21
- Associate Professor in Computing & Information Systems

Formerly

- IBM Research Australia (Mathematical Sciences)
- Microsoft Research (Silicon Valley)
- Shorter stints at Intel Labs, Yahoo! Research, Google Research
- Berkeley PhD (2010); Melbourne BSc, BE, MCompSci

...I use stats in computing almost every day!

A Personal Philosophy: Stats vs CS...

...is that there's **no** “vs" to speak of.

Prof. Michael I. Jordan of UC Berkeley CS & Statistics
(One of the highest-cited researchers in either area)

> *“I personally don't make the distinction between statistics and machine learning that your question seems predicated on.*
>
> *“Also I rarely find it useful to distinguish between theory and practice; their interplay is already profound and will only increase as the systems and problems we consider grow more complex” — Reddit AMA, Sep 2014*

While CS/Stats now converging, ideas in today's lecture are Stats

Outline

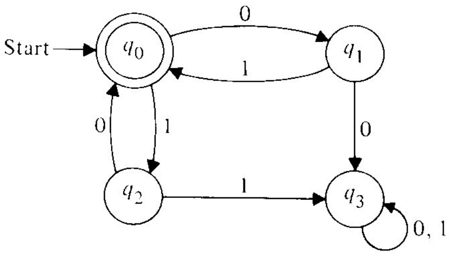
1. Introduction
   - Introducing myself
   - Is this CS or Stats?
2. Experiments and statistics
   - Why experiments in science and engineering
   - Why statistics in experiments
3. Randomization and random sampling
   - Populations and samples
   - Random sampling
4. Methods of standard error calculation
   - Assume normality
   - Bootstrap

Empiricism of science

Natural science:

- Reduction of natural world to mathematical description
- Needs experiments to do this
- (Are human created artifacts “natural”?)

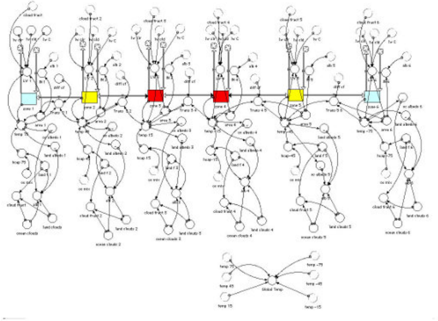
Formalism of computation/engineering



Computation and many engineering processes:

- Already mathematically described
- So why do we need experiments?

Complexity of our systems

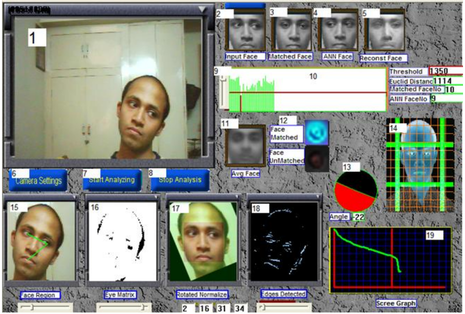


- Our systems are becoming incredibly complex
- Our systems are often not ‘closed’

...cannot be formally analysed as deterministic logical construct

- Behaviour must be assessed empirically

Interaction with outside world



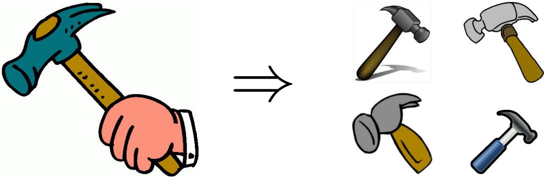
Threshold 1350
Euclid Distanc 1114
Matched FaceNo 10
ANN FaceNo 9
Angle -22

- Our systems have to interact with outside world
- . . . particularly with human users and their perceptions

Complexity and uncertainty

- Complexity means experiment outcome not precisely predictable
- Once experiment performed, the outcome is known
- If all we wanted to know was outcome of this instance of the experiment, then we're done
- Or if the outcome of this experiment perfectly predicts future outcomes of interest to use, we're also done (exhaustive testing)

Generalisation of outcomes



- Normally, we want to generalise the outcome of this experiment
- Use experiment to predict the behaviour of our system
- Question then arises: How accurate is our prediction?

Statistics as an approach to generalisation

Mean running time of program: 10.5 ± 0.8 seconds

Method A correctly detected 12% more grammatical errors than Method B, statistically significant at the 0.01 level.

- Purpose of statistics in experiments is to *generalise*
- In particular, to say *how confident* we are in a result, or *how precise* that result is
- Statistics quantifies future uncertainty—by directly harnessing that uncertainty through randomness; it is an 'inverse' to probability

Example: single-sample confidence interval

11.05	10.29	9.84	10.66	8.80
10.63	9.60	10.13	9.92	10.61

- We run a batch program 10 times, under identical conditions
- We get the above 10 timings in seconds
- The mean execution time observed is 10.15 seconds
- What should we report as the $\pm$?

Example: two-sample hypothesis test

14 17 16 15 15 19   9 16 11 10 13 13
17 8 16 16   11 12 13 14

- We have e-learning systems A and B
- Ten students study with A, ten with B
- Students sit a test, and get the above scores (out of 20)
- A achieves a mean score of 15.3, B of 12.2
- Is A significantly better than B?

System evaluation: a primer

System could be: Information retrieval, databases, AI, UX...

- System tested, compared on *test queries*
- Score per query averaged to give system score
- Systems compared using system scores
- How confident are we in a comparison?

Aside - I use the R Project for Statistical Computing
`https://www.r-project.org`
Hands down best package support for stats, data science, etc.
Python also v good particularly for production.

Example: comparing retrieval system scores

Query	1	2	3	4	5	6	7	8	9
pirc	0.16	0.06	0.41	0.09	0.72	0.17	0.52	0.52	0.57
uog	0.08	0.03	0.58	0.14	0.51	0.24	0.05	0.57	0.00

Query	10	11	12	13	14	15	16	17	18
pirc	0.42	0.64	0.33	0.54	0.15	0.72	0.12	0.09	0.38
uog	0.23	0.53	0.33	0.50	0.17	0.46	0.02	0.08	0.28

Query	19	20	21	22	23	24	25	**Mean**
pirc	0.25	0.71	0.33	0.54	0.58	0.32	0.74	**0.40**
uog	0.14	0.00	0.02	0.43	0.66	0.62	0.67	**0.29**

- System pirc, mean score 0.40, on 25 topics
- System uog, mean score 0.29, on 25 topics
- Is pirc really better than uog? Or was it just *lucky*?

Paired observations

We start by noting that system scores are *paired* by query.

- We can look at score differences on each query
- This “controls” for variability in query difficulty

Paired experiments are available if: the same subject is given two different, independent treatments

Quiz: paired experiments

Paired experiments are generally easier to arrange when computers are involved more than in many other sciences.

- Why?
- Think of an experiment from another scientific field that can be paired.
- Think of an engineering/computing experiment that cannot be paired.

Pairing scores, taking deltas

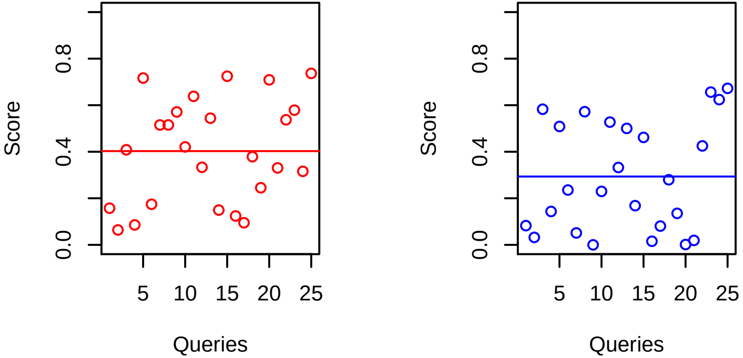
Query	1	2	3	4	5	6	7	8	9
pirc	0.16	0.06	0.41	0.09	0.72	0.17	0.52	0.52	0.57
uog	0.08	0.03	0.58	0.14	0.51	0.24	0.05	0.57	0.00
Delta	0.08	0.03	−0.17	−0.05	0.21	−0.07	0.47	−0.05	0.57

Query	10	11	12	13	14	15	16	17	18
pirc	0.42	0.64	0.33	0.54	0.15	0.72	0.12	0.09	0.38
uog	0.23	0.53	0.33	0.50	0.17	0.46	0.02	0.08	0.28
Delta	0.19	0.11	0.00	0.04	−0.02	0.26	0.10	0.01	0.10

Query	19	20	21	22	23	24	25	**Mean**
pirc	0.25	0.71	0.33	0.54	0.58	0.32	0.74	**0.40**
uog	0.14	0.00	0.02	0.43	0.66	0.62	0.67	**0.29**
Delta	0.11	0.71	0.31	0.11	−0.08	−0.30	0.07	**0.11**

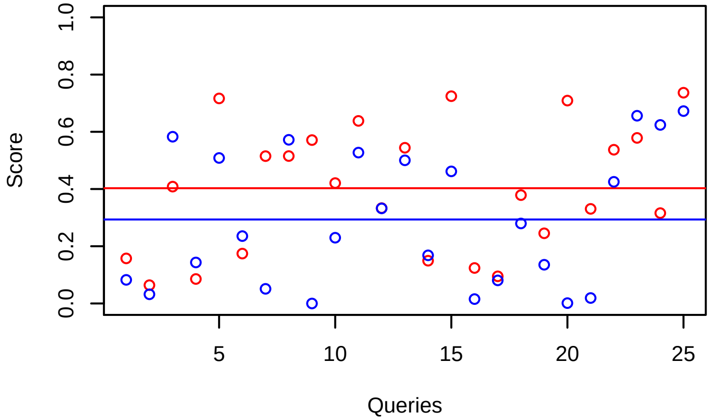
- Observed values are now deltas
- Summary value is delta between means
- Statistically, gone from a two-sample to a one-sample test

System scores, unpaired



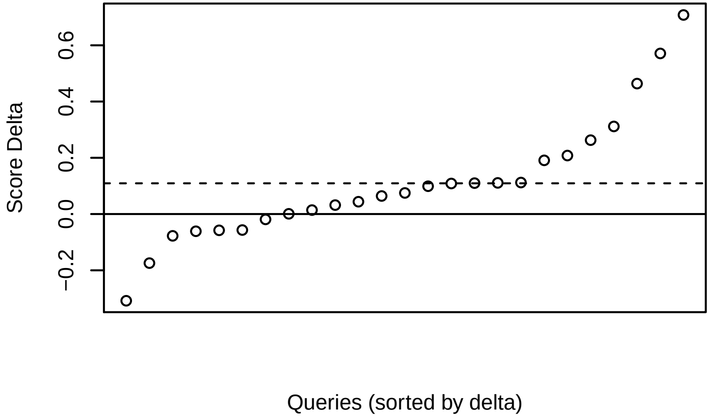
Comparing unpaired scores is tricky.

System scores, paired



Pairing results make comparison somewhat clearer.

System score deltas



Taking deltas (and sorting) makes comparison much clearer.

- Seven queries have negative score delta
- Eighteen queries have positive score delta

Quiz: sampling proportions

We have seen that, for the previous experiment, `pirc` beat `uog` on 18 of the 25 queries. Let's say we've picked these queries randomly, from some large set of queries, *Q*.

- If `pirc` were only better that `uog` on half of the queries in *Q*, how likely would we be to choose *exactly* 18 at random in which `pirc` was better?
- How likely would we be to choose *at least* 18 in which `pirc` was better?
- This analysis forms a simple kind of *significance test* (known as the sign test). What information have we thrown away in performing this analysis?

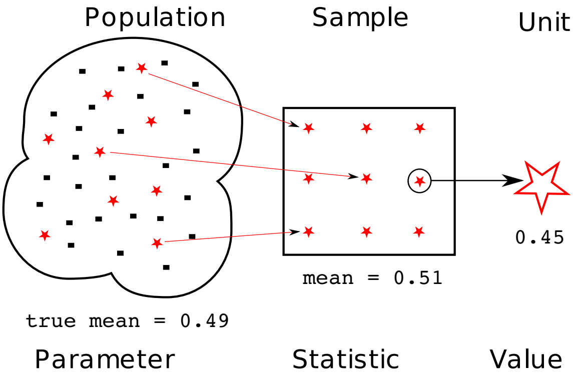
Outline

1. Introduction
   - Introducing myself
   - Is this CS or Stats?
2. Experiments and statistics
   - Why experiments in science and engineering
   - Why statistics in experiments
3. **Randomization and random sampling**
   - **Populations and samples**
   - **Random sampling**
4. Methods of standard error calculation
   - Assume normality
   - Bootstrap

Generalisation, revisited

- Observations on *this particular set of subjects* . . .
- . . . generalized to *all possible subjects*
- "this particular set of subjects" = *sample*
- "all possible subjects" = *population*

Population, sample, unit



- There is a large (possibly infinite) *population* of items.
- Of this population, the experiment examines a *sample*.
- Each element of the sample is a *unit*

Example: mice

In a medical experiment on laboratory mice:

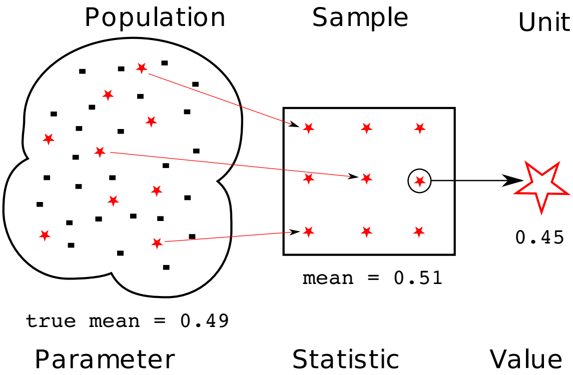
- The unit is the mouse
- The sample is the set of mice in the experiment
- The population is all mice

Example: Search System

In a search experiment:

- The unit is the query
- The sample is the set of queries in the test collection
- The population is all possible queries

Value, statistic, parameter



- For each unit we measure some *value*
- Function (e.g. mean, median) on values of sample is *statistic*
- Function (same, other) on values of population is *parameter*

Example: mice

In the lab experiment on mice:

- The value is the length of time the mouse survives after an injection
- The statistic is the median survival time of the experimental mice
- The parameter is the median survival time of all mice

Example: Search System

In a search experiment:

- The value is:
- The statistic is:
- The parameter is:

Example: Search System

In a search experiment:

- The value is: the score delta between systems on a query
- The statistic is:
- The parameter is:

Example: Search System

In a search experiment:

- The value is: the score delta between systems on a query
- The statistic is: the mean score delta on the query set
- The parameter is:

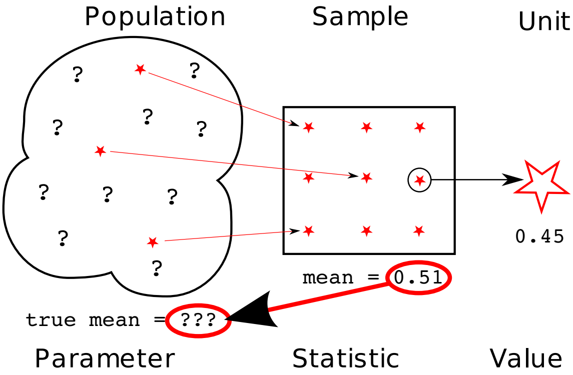
Example: Search System

In a search experiment:

- The value is: the score delta between systems on a query
- The statistic is: the mean score delta on the query set
- The parameter is: the mean score delta on all queries

This population mean score is sometimes known as the *true score*

Inference from statistic to parameter



Aim of inferential statistics

- Infer parameter on population, from statistic on sample
- Statistic is the *estimator*
- Calculate error bounds on estimate

Quiz: estimators and expected values

We randomly sample n values from a population with mean μ. We perform this sampling r times. Each time we calculate the mean m_i of the sample.

- What is:

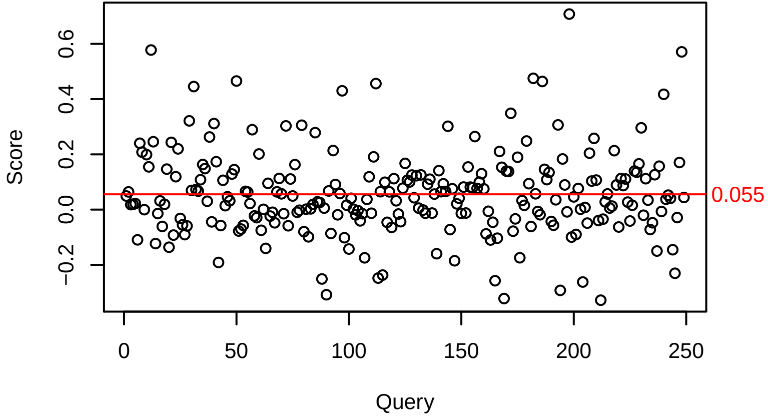
$$\lim_{r\to\infty} \sum_{i=1}^{r} \frac{m_i}{r} \ ?$$

The fact of the answer is known as the *Law of Large Numbers*

- What is:

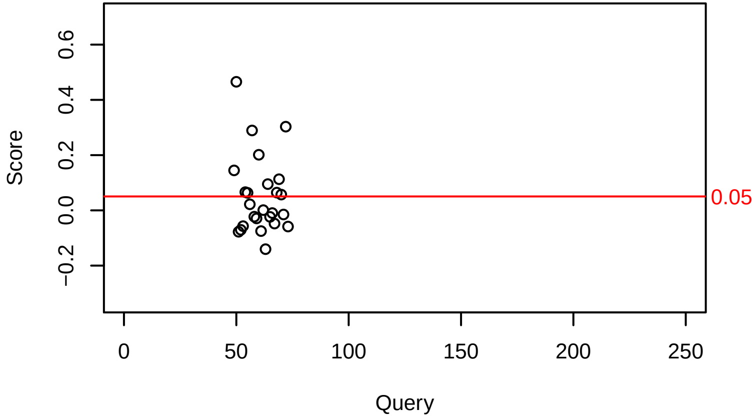
$$\lim_{r\to\infty} \sum_{i=1}^{r} \frac{|m_i - \mu|}{r} \ ?$$

Estimates (on samples) have errors



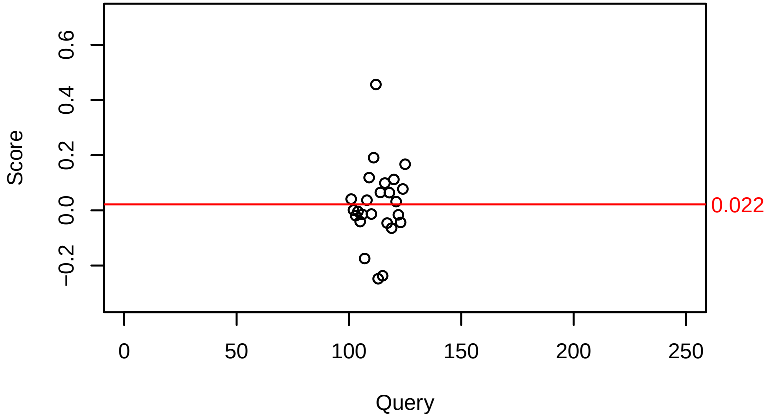
Let the above figure represent the values of our population (which, of course, we can’t generally see). Each dot is the score delta between two systems. (Our population here is finite.)

Estimates (on samples) have errors



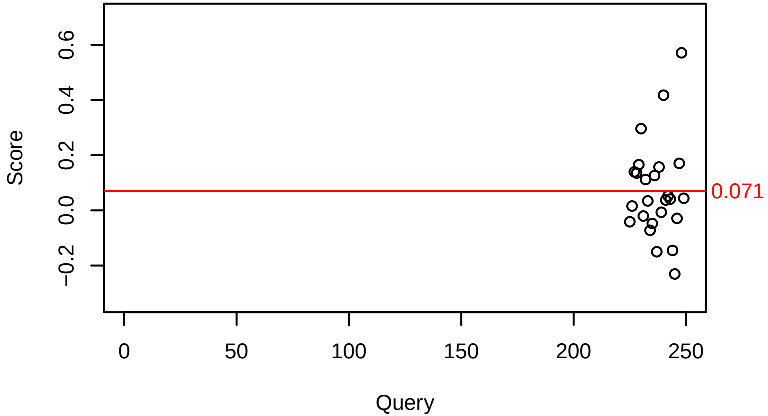
Some samples will give a close estimate

Estimate (on samples) have errors



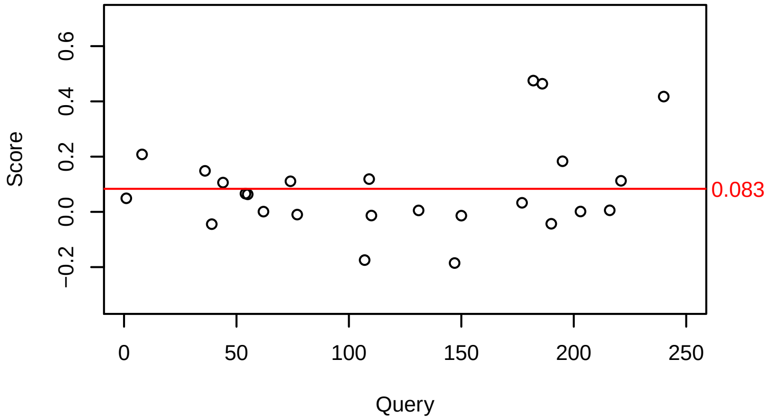
Others will give an under-estimate

Estimates (on samples) have errors



Others will give an over-estimate

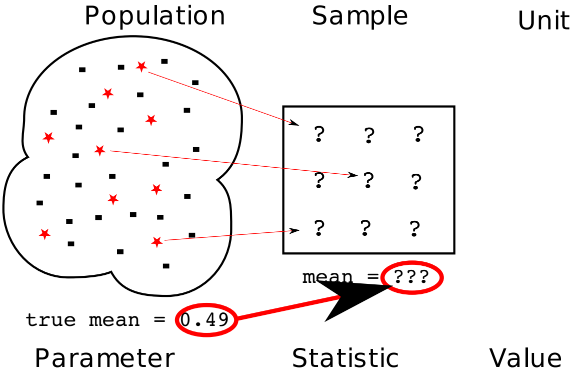
Estimates (on samples) have errors



In an actual experiment, all we see is the sample, and all we can directly calculate is the statistic. We do not know whether the statistic, as an estimator of the parameter, is low or high.

- But we can estimate the likely error of the sample statistic

Random sampling



Assume we know the population. How do we estimate the likely degree of error of a sample statistic?

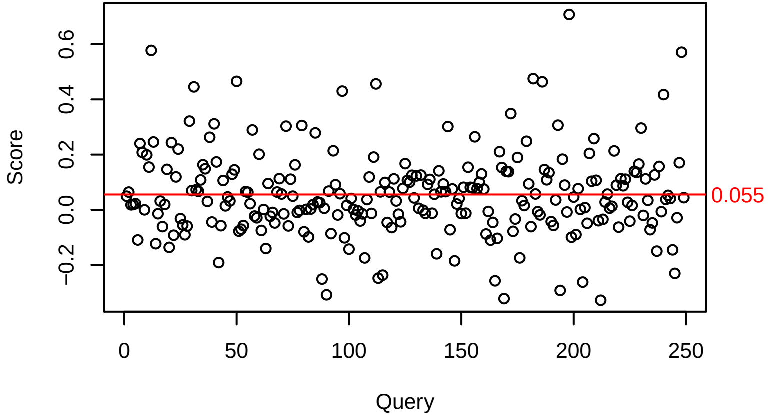
- Ensure (or assume) the sample is random
- Then the error will be random
- And the distribution of the error can be estimated

Determining statistic distribution from population

We want to know “what is the distribution of the mean of 25 values randomly sampled from our population?”

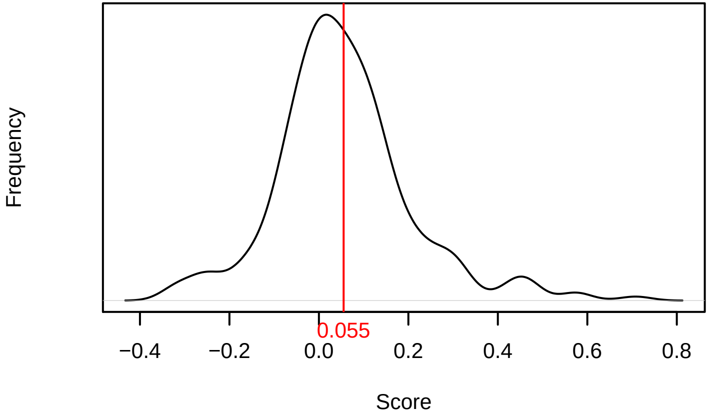
1. For *N* times:
    1. Randomly sample 25 values from the population
    2. Calculate and record the mean of the sample
2. Use the distribution of the means as our estimate of the distribution of the sample mean

Population distribution



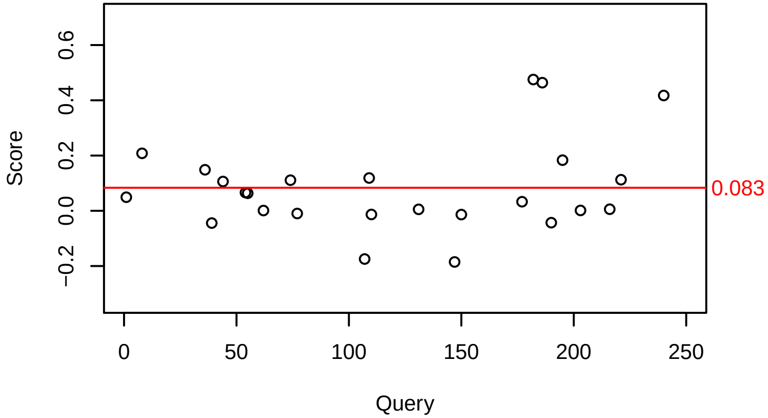
- If we knew the distribution of the population . . .

Population distribution



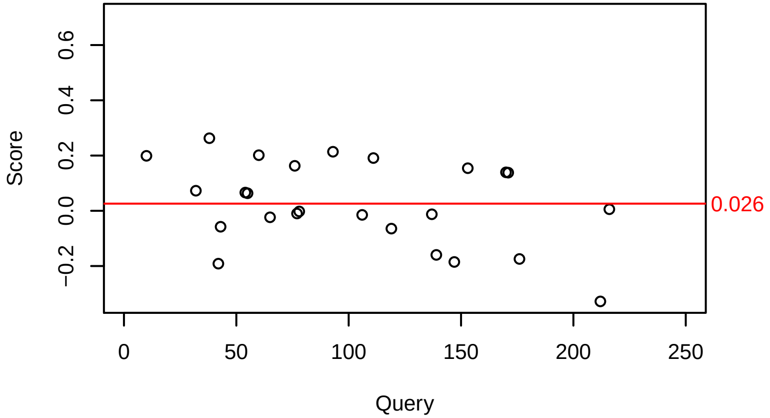
- If we knew the distribution of the population . . .

Sampling from population



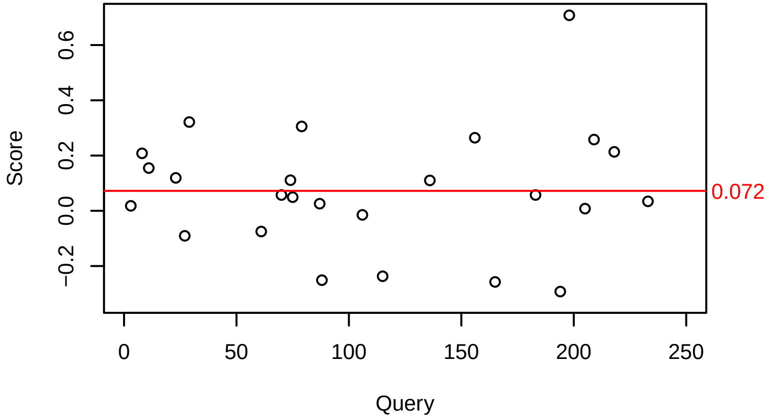
- . . . we could randomly sample from it, and calculate the value of the statistic on each sample . . .

Sampling from population



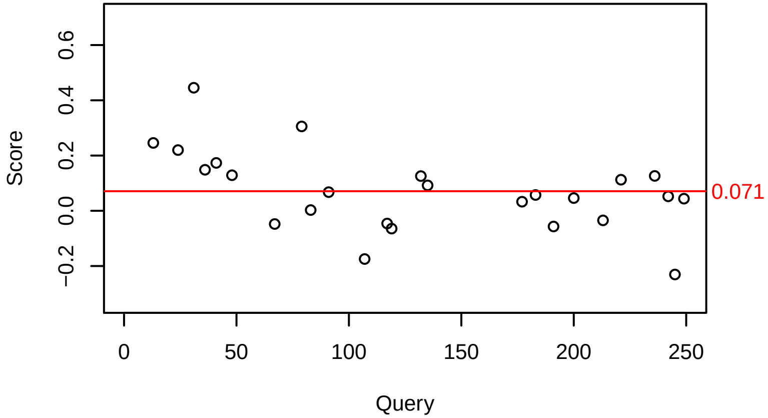
- . . . we could randomly sample from it, and calculate the value of the statistic on each sample . . .

Sampling from population



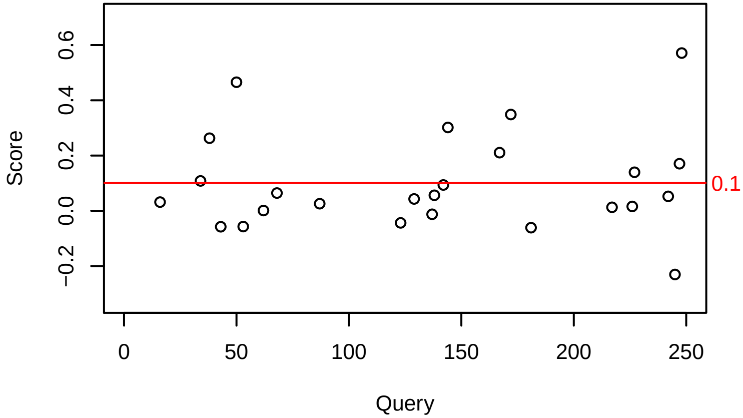
- . . . we could randomly sample from it, and calculate the value of the statistic on each sample . . .

Sampling from population



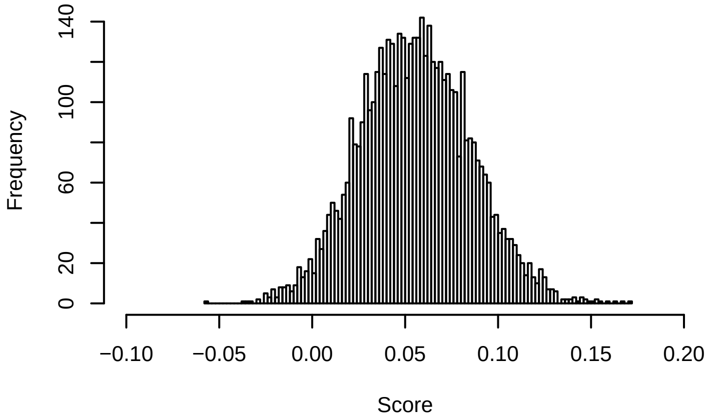
- . . . we could randomly sample from it, and calculate the value of the statistic on each sample . . .

Sampling from population



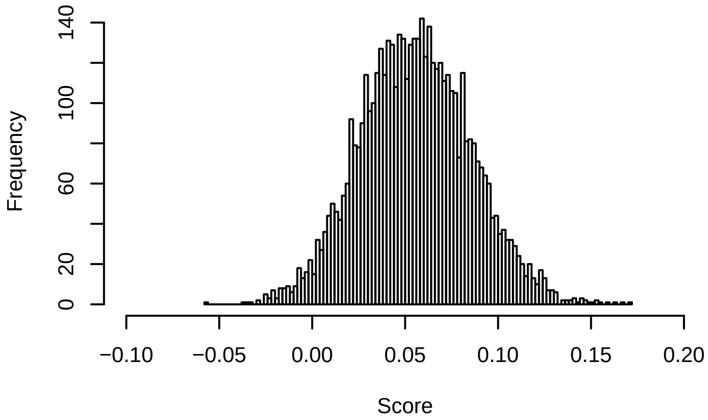
- . . . we could randomly sample from it, and calculate the value of the statistic on each sample . . .

Sampling distribution



- . . . to determine the distribution of the sample statistic
- Which allows us to estimate error of the sample statistic (i.e., confidence of the true parameter value estimated by statistic).

Problem: confidence interval



Given an individual sample statistic (say, 0.07), and given the *statistic's* distribution, how can we calculate a 95% *confidence interval* for true parameter value?

Wait, what's a confidence interval?

Misnomer:

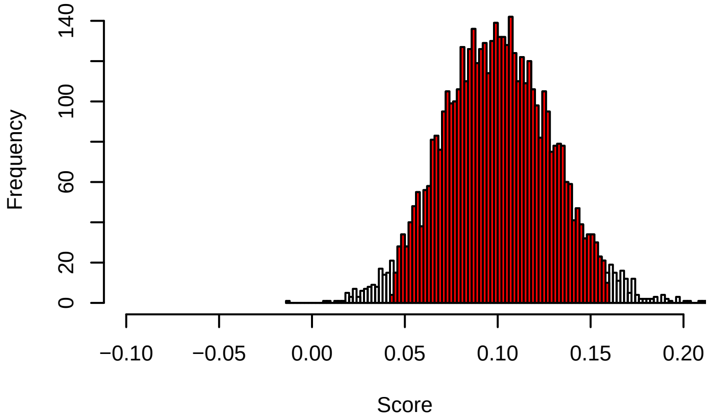
> *"95% confidence interval contains true population parameter with 95% probability"*

True parameters are not random; they are unknown, but fixed

Confidence interval $[A, B]$ for unknown parameter p at confidence 95%

- The interval (not p) is the statistic: A, B a function of sample.
- Compute intervals from 100 samples; we expect 95+ to contain p
- The probability that random $[A, B]$ contains p is ≥ 0.95

Answer: confidence interval



- Center the distribution on the sample statistic.
- Calculate the 2.5% and 97.5% *percentiles*.
- *A*, *B* given by these percentiles plus the sample statistic

Significance test

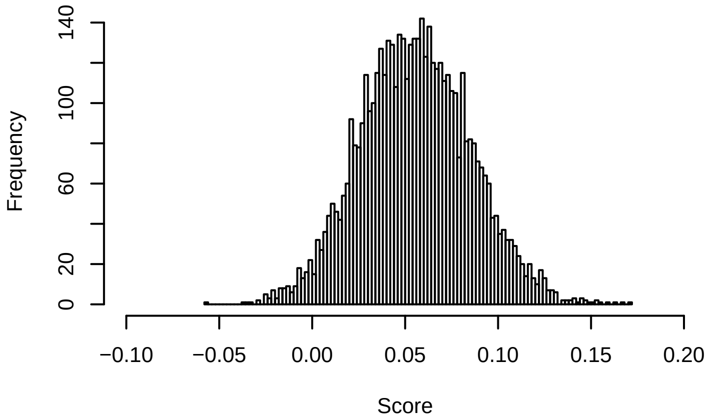
The simplest form of significance test tests this question:

> *I have randomly sampled n items, and observed a statistic of m. If the true parameter value were* 0*, what is the probability p that I would have achieved this high or higher a statistic?*

The answer to this question gives the *p value* of the significance test. If $p < \alpha \in \{0.05, 0.01, 0.001\}$, we declare the result significant (that is, very unlikely to have happened by chance if there were no true difference between the systems).

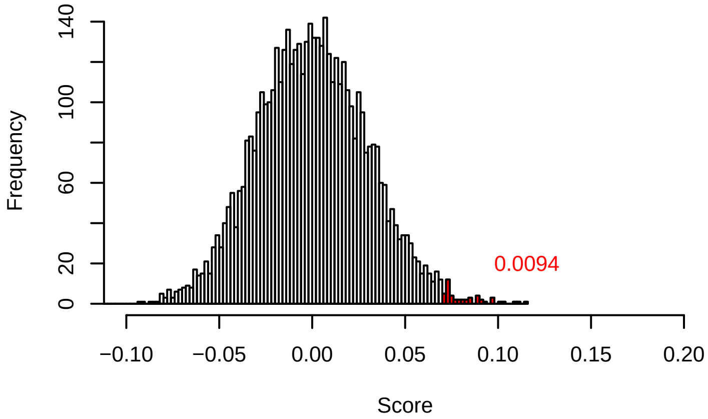
- We have seen this sort of reasoning before in this talk. Where?

Problem: significance test



Given an individual sample statistic (say, 0.1), and given the distribution of the sample statistic, how can we calculate the *p* value of the experiment?

Answer: significance test



- Center the distribution on 0 (for assumed true parameter being 0)
- Calculate the proportion of the distribution that equals or exceeds the observed sample statistic. That's the *p* value!

Quiz: what is the problem?

In a given experiment, what is the practical problem with using the above procedures to construct confidence intervals and significance test *p*-values?

Quiz: what is the problem?

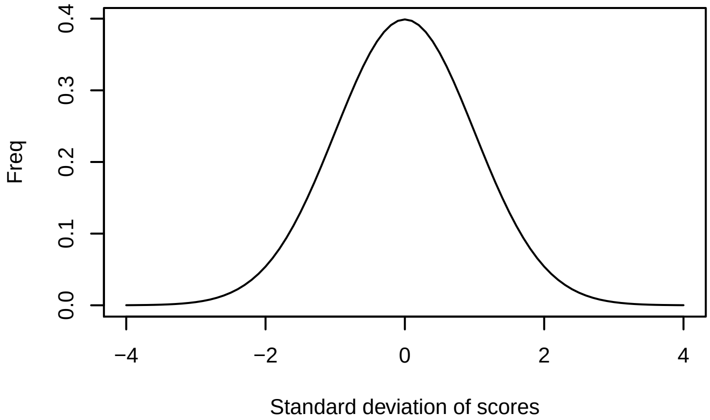
In a given experiment, what is the practical problem with using the above procedures to construct confidence intervals and significance test *p*-values?

- The problem is that, in an experiment, we only see the *sample*, we don't see the *population*.
- These calculations assumed knowledge of the *sample statistic distribution*, which comes from the population.

Outline

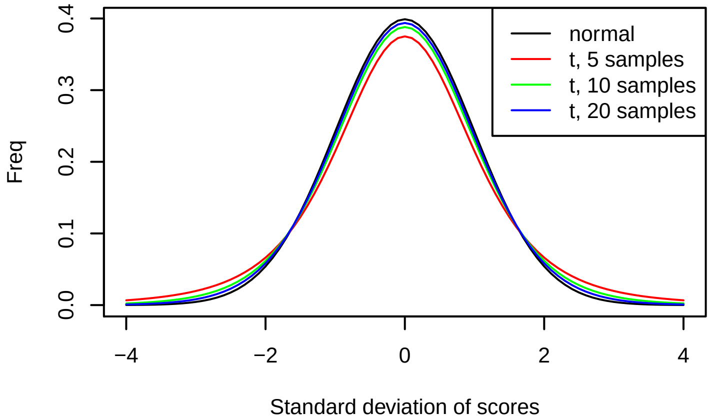
1. Introduction
   - Introducing myself
   - Is this CS or Stats?
2. Experiments and statistics
   - Why experiments in science and engineering
   - Why statistics in experiments
3. Randomization and random sampling
   - Populations and samples
   - Random sampling
4. Methods of standard error calculation
   - Assume normality
   - Bootstrap

Assume the population is normally distributed



- Know (or assume) population is normally distributed
- Our sample statistic is the mean
- Then the distribution of the sample statistic is known

The *t* distribution



- Mean of sample of *n* Normal items (scaled by standard error) . . .
- Follows the *t* distribution with $n - 1$ *degrees of freedom*
- This observation forms the basis for the *t* test

The *t* test

```
> sc.a <− c(0.15, 0.06, 0.40, 0.08, 0.71, 0.17, ...)
> sc.b <− c(0.08, 0.03, 0.58, 0.14, 0.50, 0.23, ...)
> t.test(sc.a, sc.b, type="paired")

        Paired t−test

data:  sc.1 and sc.2
t = 2.4407, df = 24, p−value = 0.02242
alternative hypothesis: true difference in means \
   is not equal to 0
95 percent confidence interval:
 0.01687137 0.20171263
sample estimates:
mean of the differences
               0.109292
```

Limits of *t* test

However *t* test is usually only applicable if:

- The statistic of interest is the mean
- The underlying population is normally distributed

Assumption of normal distribution particularly troublesome

Central limit theorem

Central limit theorem:

> *The mean of a large enough i.i.d. sample from any distribution with finite variance is approximately normally distributed*

- 'i.i.d' shorthand for: independent, and identically distributed
- So if our sample size is "large enough" ...
- We can treat the sample mean's distribution as normal!

Using the *t* test under the CLT

- Formally, should use statistical test based on normal distribution
- In practice, generally just use *t* distribution
- Because *t* distribution is close to normal . . .
- and is more conservative . . .
- and for all we know the population may be normal . . .
- and it's convenient and conventional!

Assumptions of central limit theorem

To use CLT plus *t* test:

- Sample must be “large enough”
- “Large enough” depends on true distribution, but 30 is a typical threshold
- Also, only works for the mean

What can we do about sample statistics other than the sample mean...

The plug-in principle

The plug-in principle is where we use a function (as a statistic) of the sample to estimate the value of the same function (as a parameter) of the population.

- For instance, we estimate the mean of the population based on the mean of the sample.
- We can also estimate the standard deviation of the population based on the standard deviation of the sample.

Treating the sample as our population

We can push this even further:

- Estimate the distribution of the population from the distribution of the sample
- Resample with replacement from the sample to simulate sampling from the population
- Use these re-samples to estimate the distribution of the statistic

This is the *bootstrap principle*: ‘easy’ to implement, incredibly powerful

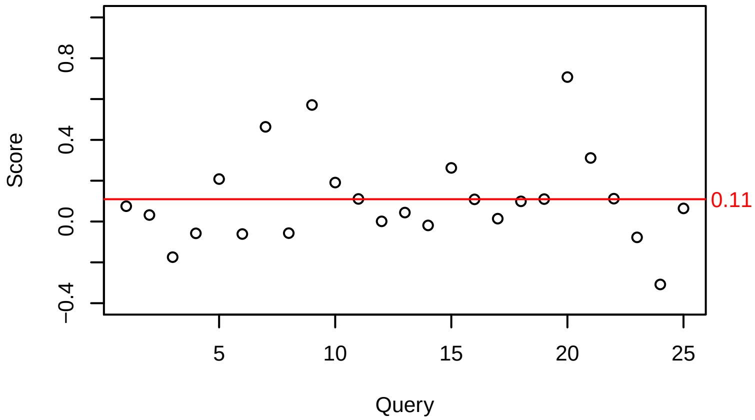
Bootstrapping

Bootstrapping:

- Resample, with replacement, from our sample
- For each resample, calculate statistic
- Collect thousands of resamplings, calculate distribution

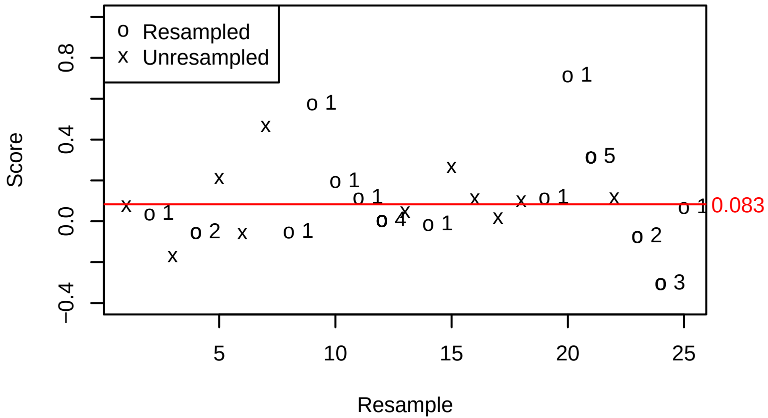
Distribution of resamplings is our estimate of sampling distribution from original population.

The Original Sample



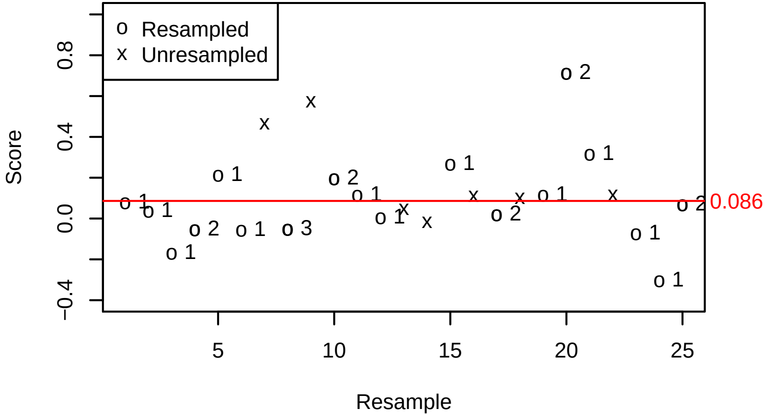
We start with our original sample

Re-Sampling



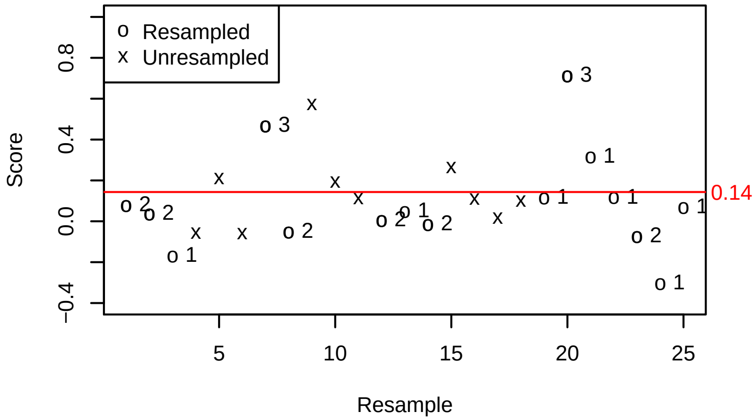
- Draw resamples, with replacement, from original sample
- Sample size same as original sample
- Record summary statistic (estimator)

Re-Sampling



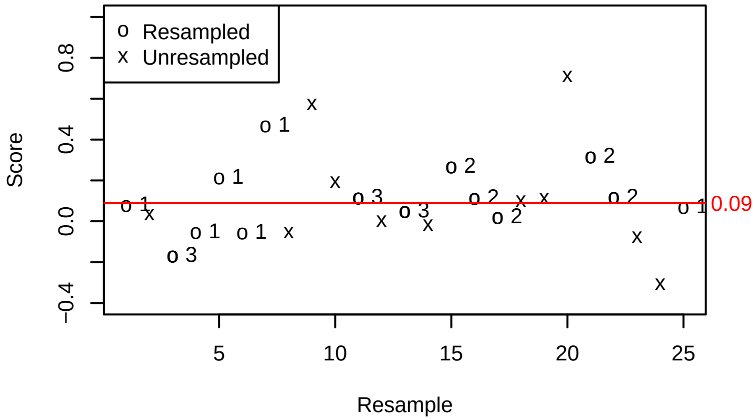
- Draw resamples, with replacement, from original sample
- Sample size same as original sample
- Record summary statistic (estimator)

Re-Sampling



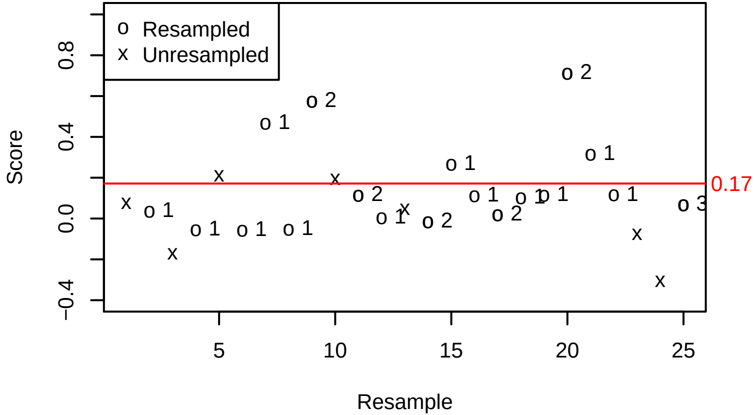
- Draw resamples, with replacement, from original sample
- Sample size same as original sample
- Record summary statistic (estimator)

Re-Sampling



- Draw resamples, with replacement, from original sample
- Sample size same as original sample
- Record summary statistic (estimator)

Re-Sampling



- Draw resamples, with replacement, from original sample
- Sample size same as original sample
- Record summary statistic (estimator)

Re-sampled statistics

0.13, 0.09, 0.15, 0.09, 0.07, 0.07, 0.08, 0.13, 0.00, 0.13,

- Repeat the resampling multiple times
- Record value of summary statistic on each resample
- ...until stable distribution of resampled statistic appears

Re-sampled statistics

0.13, 0.09, 0.15, 0.09, 0.07, 0.07, 0.08, 0.13, 0.00, 0.13, 0.07, 0.07, 0.11, 0.19, 0.08, 0.09, 0.18, 0.04, 0.18, 0.13, 0.07, 0.08, 0.04, 0.14, 0.13, 0.13, 0.16, 0.15, 0.11, 0.08,

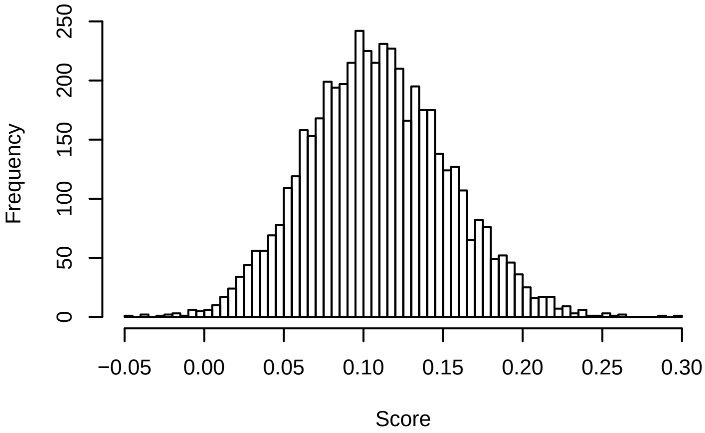
- Repeat the resampling multiple times
- Record value of summary statistic on each resample
- ...until stable distribution of resampled statistic appears

Re-sampled statistics

0.13, 0.09, 0.15, 0.09, 0.07, 0.07, 0.08, 0.13, 0.00, 0.13, 0.07, 0.07, 0.11, 0.19, 0.08, 0.09, 0.18, 0.04, 0.18, 0.13, 0.07, 0.08, 0.04, 0.14, 0.13, 0.13, 0.16, 0.15, 0.11, 0.08, 0.11, 0.15, 0.10, 0.06, 0.15, 0.03, 0.13, 0.04, 0.08, 0.12, 0.05, 0.16, 0.06, 0.04, 0.11, 0.18, 0.13, 0.02, 0.12, 0.18,

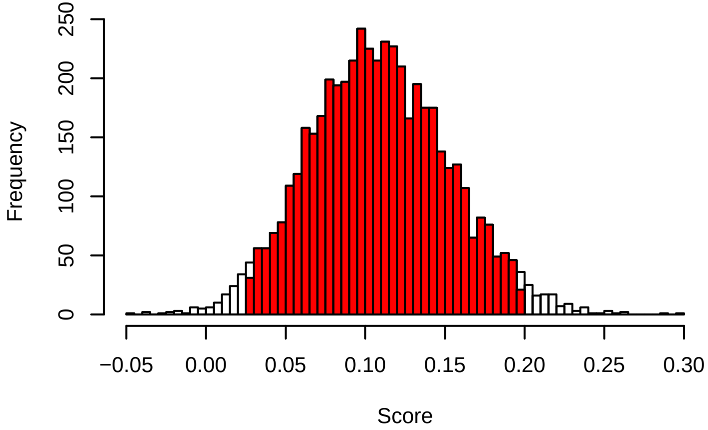
- Repeat the resampling multiple times
- Record value of summary statistic on each resample
- ...until stable distribution of resampled statistic appears

Re-sampling (bootstrap) distribution



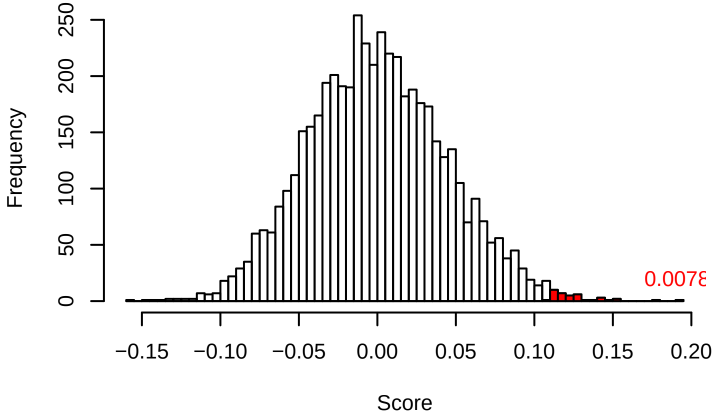
- Distribution of resampled statistic estimates true static distribution
- Apply as usual to confidence intervals, significance tests

Confidence interval



The 95% confidence interval on the mean score delta between the `pirc` and `uog` is 0.025 – 0.195.

Significance



This result is significant at $\alpha = 0.01$, in a one-tailed, paired bootstrap test.

Summary

Experimental uncertainty:

- Results on experimental data are not enough: we want to generalize them
- Generalisation involves uncertainty
- It is not enough simply to quote the mean value of our experiments: we need to indicate how certain we are of this mean

Summary (continued)

Statistical modelling

- Assume that our experimental sample has been randomly sampled from the full population
- The statistic (mean) on the sample becomes our estimate for the parameter (true mean) on the population
- Quantify uncertainty of this estimate

Summary (continued)

Statistical inference: Need handle on sampling distribution of statistic

Alternatives for the mean (a particularly common example)

- Assume population is normally distribution (*t* test)
- Have large sample and invoke Central Limit Theorem (*t* test)
- Estimate population distribution from sample, resample to generate sampling distribution of statistic (Boostrap)

The Bootstrap can be used for many other statistics too