The 38th CREST Open Workshop Working Tutorial on Statistical Methods in Experimental Software Engineering

Simon Poulding

Department of Software Engineering
Blekinge Institute of Technology

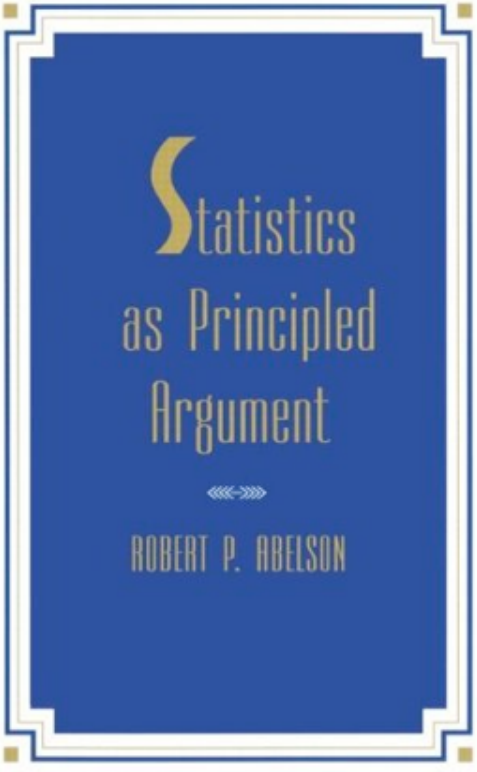
January 2015

Part I

Introduction

- Statistics as Principled Argument
- Tutorial Topics

Statistics as Principled Argument



Robert P. Abelson, "Statistics as Principled Argument", 1995

statistics as evidence supporting an argument

. . . and this argument should be part of an engaging narrative about the quantitative research

Abelson's MAGIC Criteria

Magnitude	the size of the quantitative support for the claim
Articulation	amount of comprehensible detail in the conclusions
Generality	are the conclusions broadly applicable?
Interestingness	how important is the result; does it change belief?
Credibility	methodological soundness and coherence with theory

Effect of Reward Experiment

1. each participant performs a boring task
2. they are asked to tell the next subject that the task is interesting, and paid $1 or $20 for doing so
3. later asked to rate the task from -5 (very boring) to +5 (very interesting)

amount paid	mean rating
$1	1.35
$20	-0.05

Effect of Reward Experiment

Magnitude	difference in mean score of t-test gave a p-value < 0.03
Articulation	"smaller reward causes greater tendency to change belief to conform to behaviour"
Generality	does it only occur in this specific situation? only under lab conditions?
Interestingness	counter to expectations
Credibility	coherency with cognitive dissonance theory some minor questions as to statistical analysis open to other interpretations

Magnitude Example

Table II
FOR EACH PROJECT, AVERAGE COVERAGE ON ALL OF ITS CLASSES WHEN NO BYTECODE CONSTANT IS USED ($P_{\text{BYTECODE}} = 0$) AND WHEN THEY ARE USED WITH PROBABILITY $P_{\text{BYTECODE}} = 0.2$. THE $\hat{A}_{12}$ OF THESE COMPARISONS ARE CALCULATED BY AGGREGATING ALL RUNS OF ALL CLASSES PER PROJECT (IN BOLD IF STATISTICALLY SIGNIFICANT). ON HIGHER GRANULARITY, IT IS REPORTED THE PERCENTAGE % OF CLASSES FOR WHICH WE HAVE A SIGNIFICANT $\hat{A}_{12} > 0.5$ AND $\hat{A}_{12} < 0.5$.

Project	$P = 0$	$P = 0.2$	$\hat{A}_{12}$	$\% > 0.5$	$\% < 0.5$
COL	0.74	0.73	**0.48**	0.04	0.27
CCL	0.87	0.90	**0.56**	0.21	0.07
CCD	0.87	0.88	0.52	0.29	0.00
CCO	0.91	0.91	0.50	0.02	0.01
CMA	0.75	0.75	0.51	0.12	0.02
CPR	0.93	0.94	**0.52**	0.15	0.005
GCO	0.74	0.74	0.50	0.04	0.01
ICS	0.85	0.86	0.50	0.05	0.00
JCO	0.82	0.82	0.50	0.07	0.00
JDO	0.73	0.73	0.50	0.12	0.00
JGR	0.75	0.75	0.50	0.03	0.01
JTI	0.84	0.85	**0.52**	0.18	0.00
NXM	0.59	0.59	0.51	0.00	0.00
NCS	0.97	0.97	0.51	0.09	0.00
REG	0.75	0.75	0.50	0.00	0.00
SCS	0.63	0.85	**0.77**	0.75	0.00
TRO	0.88	0.87	**0.46**	0.005	0.32
XEN	0.65	0.72	**0.57**	0.29	0.00
XOM	0.76	0.77	0.51	0.17	0.00
ZIP	0.80	0.83	**0.69**	1.00	0.00
Average	0.79	0.81	0.53	0.18	0.04

Fraser and Arcuri, "The Seed is Strong: Seeding Strategies in Search-Based Software Testing", ICST 2012

Articulation Example

"Our experiments show with strong statistical confidence that, even for a testing tool that is already able to achieve high coverage, the use of appropriate seeding strategies can further improve performance."

Fraser and Arcuri, "The Seed is Strong: Seeding Strategies in Search-Based Software Testing", ICST 2012

Generality Example

"To avoid a bias in our case study class selection, we therefore randomly chose 20 classes out of the SF100 corpus of Java projects randomly selected from Sourceforge . . ."

Pavlov and Fraser, "Semi-Automatic Search-based Test Generation", SBST 2012

Interestingness Example

"This paper presents an approach in which examples of inputs are sought from the Internet by reformulating program identifiers into web queries."

McMinn, Shahbaz, and Stevenson, "Search-Based Test Input Generation for String Data Types Using the Results of Web Queries", ICST 2012

Credibility Example

"One of the critical aspects of the application of search based techniques is finding the right configuration of parameters. We sampled the set of all parameter values combinations . . . by selecting all pair-wise interactions between values . . ."

Gómez, Baudry, Sahraoui, "Searching the boundaries of a modeling space to test metamodels", ICST 2012

Abelson's Rule of Two Criticism

MAGI Criteria

Two or more criticisms in Magnitude, Articulation, Generality, and Interestingness should lead to rejection.

C Criterion

Lack of Credibility should lead to debate.

- Statistics as Principled Argument
- Tutorial Topics

Tutorial Topics - Motivation

1. techniques I use myself in my research
2. topics similarly suggested by Mark
3. topics covered in Arcuri and Briand, "A Hitchhiker's guide to statistical tests for assessing randomized algorithms in software engineering", Software Testing, Verification and Reliability 2014; **24**:219–250

Tutorial Topics - MAGIC Criteria

Hypothesis Testing Credibility
Effect Size Magnitude
Confidence Intervals Credibility
Sample Size Credibility

Linear Models Magnitude
Relationships between Variables Magnitude
Controlling Nuisance Factors Credibility
Visualisation Articulation

Part II

Introduction to R

Objects and Expressions

R

```
> foo <- 42
> bar = 19
> baz = log(7)
> foo + bar
[1] 61
> baz
[1] 1.94591
```

Vectors

R

```
> x <- c(1, 1, 2, 3, 5, 8, 13, 21)
> x[7]
[1] 13
> 2*x
[1]  2  2  4  6 10 16 26 42
> x+10
[1] 11 11 12 13 15 18 23 31
> length(x)
[1] 8
> z = 1:10
> z
[1]  1  2  3  4  5  6  7  8  9 10
```

Some Statistics

R

```
> y <- c(6.1, 4.2, 4.6, 3.9, 5.8, 6.2, 5.4)
> mean(y)
[1] 5.171429
> median(y)
[1] 5.3
> sd(y)
[1] 0.9357961
> IQR(y)
[1] 1.55
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.900   4.400   5.400   5.171   5.950   6.200
```

Data Frames

R

```
> y1 <- c(10.3, 11.2, 5.3, 7.7, 8.0)
> y2 <- c(230, 251, 173, 174, 209)
> df <- data.frame(time=y1, memory=y2)
> df
  time memory
1 10.3    230
2 11.2    251
3  5.3    173
4  7.7    174
5  8.0    209
> df$memory
[1] 230 251 173 174 209
> names(df)
[1] "time"   "memory"
```

Importing Data

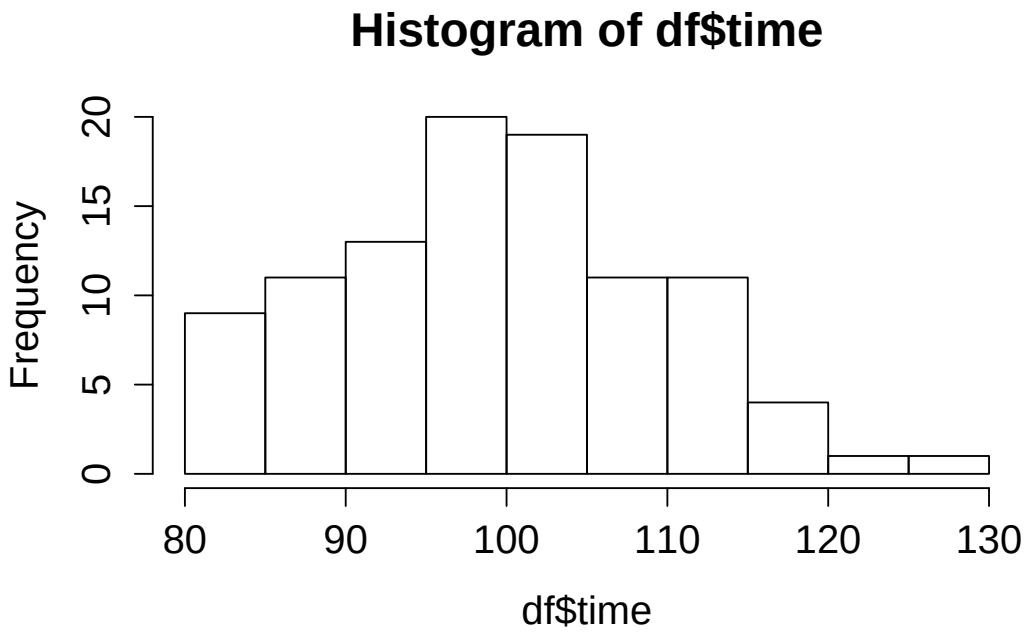
R

```
> df <- read.table("cow38A.dat")
> df
     time memory
1   111.1 225.34
2    96.7 189.54
3   115.6 241.86
4    86.4 158.39
5    84.0 179.93
6   106.4 219.67
```

Plotting Data

R

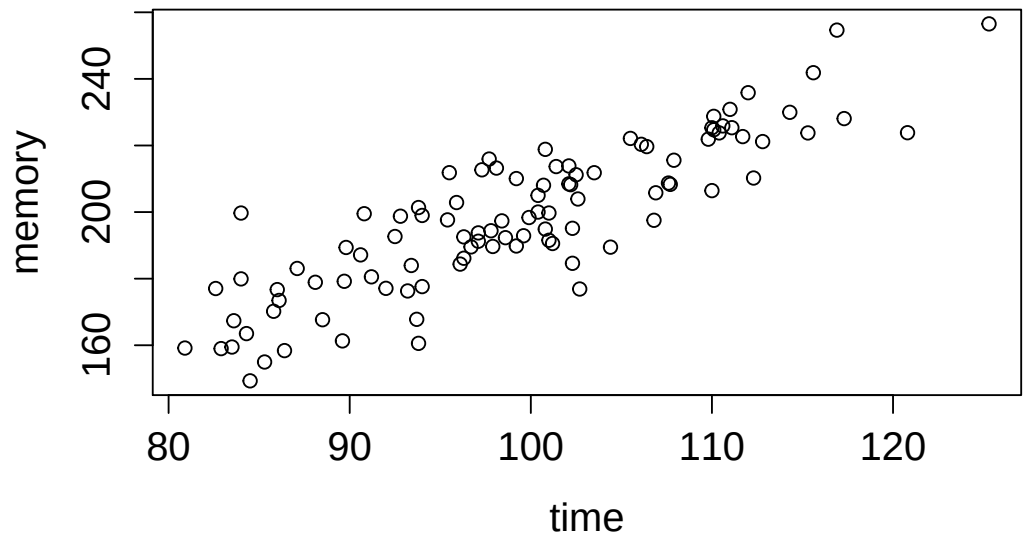
```
> hist(df$time)
```



Plotting Data

R

```
> plot(df)
```



Executing Scripts

cow38Aplot.R

```
df <- read.table("cow38A.dat")
pdf("cow38_time_v_memory.pdf", height=4, width=5)
plot(df)
dev.off()
```

R

```
> source("cow38Aplot.R")
```

Getting Help

R

```
> help(plot)
> ?plot
> help.start()
```

Workspace and Exiting

R

```
> q()
Save workspace image? [y/n/c]:
```

Practical

Background

Experiment cow38B measured the coverage obtained by a testing tool.
Experiment cow38C measured the time taken to perform automated refactorings

For each datasets cow38B.dat and cow38C.dat:

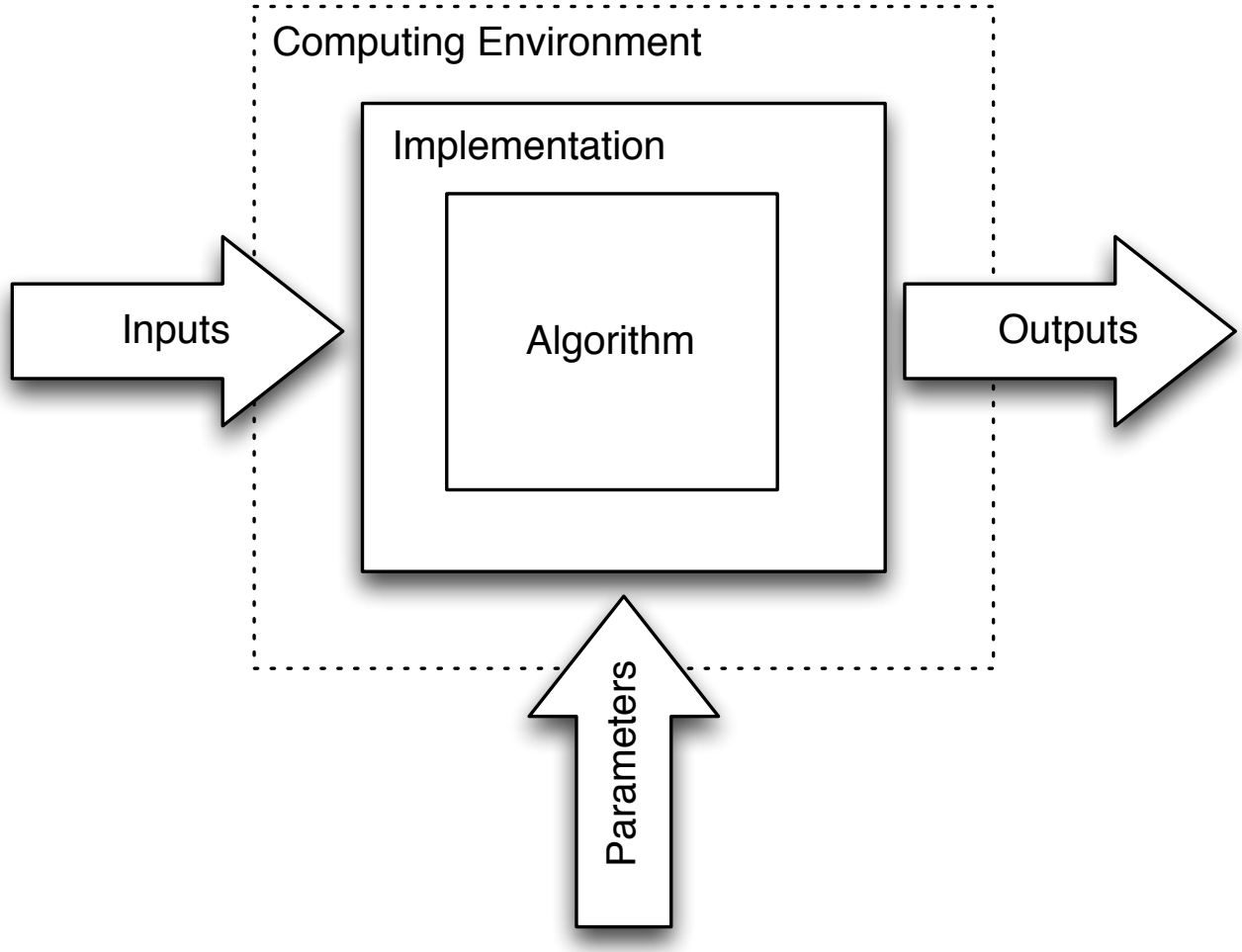
1. Import the data into R. [Note: cow38C.dat has no row nor column names.]
2. Calculate the mean, median, standard deviation, and interquartile range.
3. Create a histogram.
4. Create a boxplot.

Part III

Statistical Inference

- Experiments, Factors and Responses
- Example Experiments
- Parametric and Non-Parametric Techniques

Experiments on Algorithms



Response

Definition

The response is the quantity that we are interested in.

Examples

- the quality of the solution that algorithm finds
- whether or not the algorithm finds a solution
- how long the algorithm takes to find a solution
- the power consumed by the processor while running the algorithm

Factors

Definition

A factor is anything that could affect the response.

Examples

- the input to the algorithm
- the setting of the algorithm parameters
- the language used to implement the algorithm
- CPU and memory specs of the server
- other users on the server

Stochastic Algorithms

Stochastic algorithms use a Pseudo-Random Number Generator (PRNG) to make random choices. Since the choices made affect the output of the algorithm, the initial state of the PRNG is a factor.

The initial state is typically set by:

- seeding from a source of entropy from the computing environment (e.g. the system clock)
- explicitly setting the seed each time the algorithm is run from a source of entropy such as random.org

Controlled and Uncontrolled Factors

Controlled

Factors such as the inputs, parameter settings, language of implementation can be controlled.
However only some of these may be of interest for the purpose of the experiment; the others are nuisance factors.

Uncontrolled

Some factors, such other users on the server, may not be in our control.

Choosing an appropriate response can limit controlled and uncontrolled factors that are not of interest.

Research Questions

Research questions are typically about the probability of distribution of the response.

Examples

- What is the mean quality of the solutions found by the algorithm?
- For what proportion of runs is the algorithm successful?
- How long do I need to run the algorithm to get a solution 90% of the time?
- What is the variance in the power consumed by the processor when running the algorithm?

Population

The research question is typically posed over all values of one or more factors.

For example, "What is the mean coverage achieved by the test data generation algorithm?" is implicitly over a set of possible software-under-test, e.g. all Java programs.

If the algorithm is stochastic, it is implicitly over all set of initial states of the PRNG.

Using the terminology of statistics, the response over all the values of these factors is the population distribution.

Sample

Typically we cannot consider the entire population: the set of Java programs is infinite; the set of PRNG states is extremely large.

Therefore we measure the response for a small sample taken from the population.

We then infer properties of the population distribution from sample distribution.

- Experiments, Factors and Responses
- Example Experiments
- Parametric and Non-Parametric Techniques

Example 1: ScandiTest Quality

Background

ScandiTest is a search-based algorithm for creating tests for structural coverage.

Research Question

What is the average coverage achieved by ScandiTest?

Example 2: ScandiTest Robustness

Background

Although ScandiTest typically achieves high coverage, it only finished successfully on 70% of the runs, even on the same SUT. ScandiTest2 is an improved algorithm that is designed to finish more often.

Research Question

Have the improvements in ScandiTest2 had an effect on robustness?

Example 3: ScandiTest versus BritTest Performance

Background

BritTest is the current state-of-the-art algorithm that uses a genetic algorithm.

Research Question

For a given coverage, does ScandiTest achieve this coverage more quickly than BritTest?

Example 4: ScandiTest Parameter Settings

Background

ScandiTest uses a novel bio-inspired optimisation technique called Reindeer Herd Search. There are three categorical parameters that can be on or off: rednose, heavysnow, pullingsleigh.

Research Question

The coverage of ScandiTest does not depend on any of these parameter settings.

Example 5: ScandiTest (Memory) Scalability

Background

A major practical constraint on using ScandiTest is the amount of memory it appears to consume.

Research Question

How does the memory used scale with the number of structural elements in the SUT?

Example 6: ScandiTest (Time) Performance

Background

Anecdotal evidence is that the slowest runs of ScandiTest are also the ones that consume most memory.

Research Question

Is the the time performance of ScandiTest correlated with its memory performance?

Example 7: ScandiTest Parameters

Background

ScandiTest uses a novel bio-inspired optimisation technique called Reindeer Herd Search. It has 5 numeric parameters.

Research Question

What are the best setting of the parameters?

- Experiments, Factors and Responses
- Example Experiments
- Parametric and Non-Parametric Techniques

Parametric and Non-Parametric Techniques

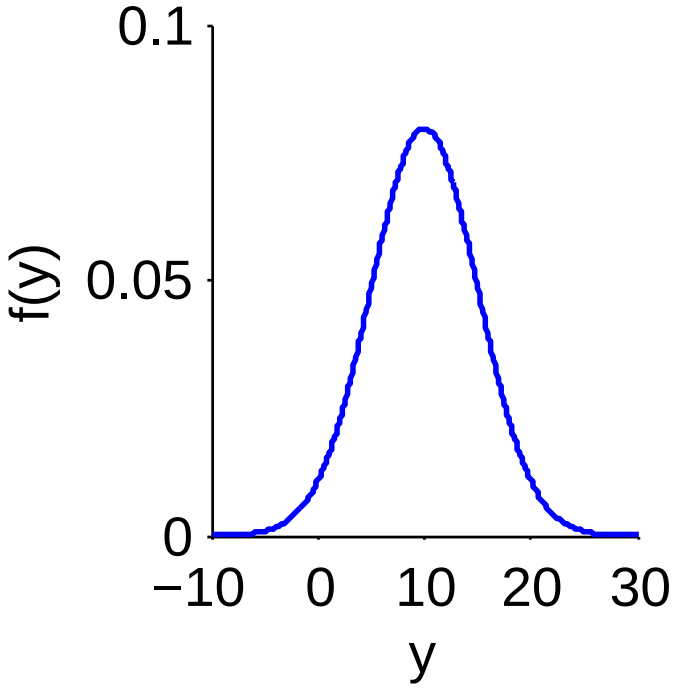
Parametric

Parametric techniques assume that the population distribution has a particular form.

Non-Parametric

Non-parametric techniques make few assumptions about the population distribution.

Normal Distribution

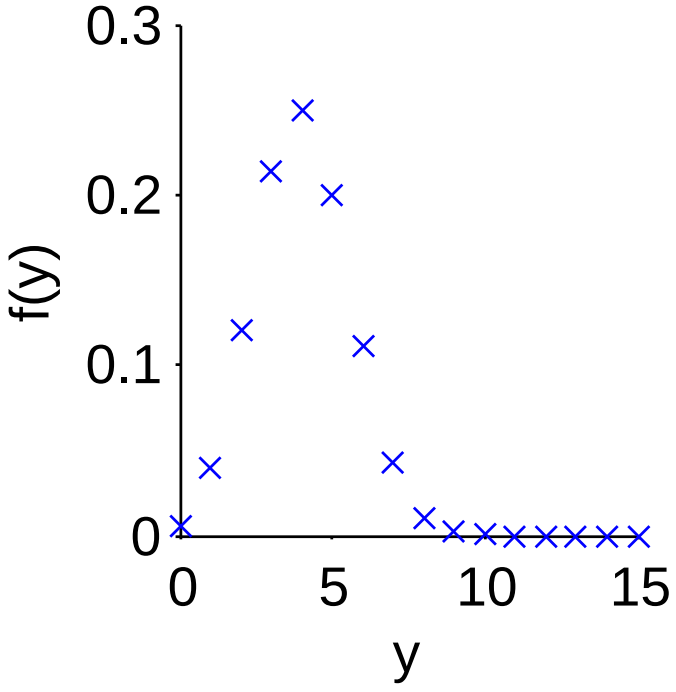


distribution parameters

population mean: μ
population variance: σ^2

often denoted $\mathcal{N}(\mu, \sigma^2)$

Binomial Distribution



distribution parameters

number of trials: N

probability of success: p

properties (moments)

population mean: Np

population variance: $Np(1-p)$

Arguments in Favour of Parametric Techniques

- Parametric techniques can make 'stronger' inferences than non-parametric techniques for a given sample size.
- Some types of analysis are only possible when parametric assumptions are made.
- Parametric techniques are more widely understood that non-parametric.

Arguments in Favour of Non-Parametric Techniques

- There is often no theoretical justification for expecting a response to have a particular distribution.
- Instead you must demonstrate empirically the assumptions of parametric techniques are met.
- Small deviations from parametric assumptions may make any conclusions unreliable.

Key Points

- The response is the output of the experiment and it depends on input factors.
- Research questions are often concerned with properties of the population distribution.
- Experiments take samples from which properties of the population distribution can be inferred.
- Parametric statistical techniques make assumptions as to the form of the population distribution.

Part IV

Hypothesis Testing

Example 2: ScandiTest Robustness

Background

Although ScandiTest typically achieves high coverage, it only finished successfully on 70% of the runs, even on the same SUT. ScandiTest2 is an improved algorithm that is designed to finish more often.

Research Question

Have the improvements in ScandiTest2 had an effect on robustness?

Example 3: ScandiTest versus BritTest Performance

Background

BritTest is the current state-of-the-art algorithm that uses a genetic algorithm.

Research Question

For a given coverage, does ScandiTest achieve this coverage more quickly than BritTest?

- Hypothesis Testing
  - General Principle
  - Hypothesis Testing Example
  - Formal Process
- Significance and Power
- Variants of Hypothesis Test
- Standard Tests
- Multiple Comparisons

Informal Process

1. Sample responses from the algorithm(s).
2. Calculate an appropriate statistic on the observed data.
3. Determine the probability of the observed value of the statistic or a more extreme value, assuming no difference between the algorithms.
4. If the probability is small, assume that there is a difference in algorithms.

Example 2: ScandiTest Robustness

Background

Although ScandiTest typically achieves high coverage, it only finished successfully on 70% of the runs, even on the same SUT. ScandiTest2 is an improved algorithm that is designed to finish more often.

Research Question

Have the improvements in ScandiTest2 had an effect on robustness?

Step 1 - Sample responses

Method

Run the algorithm 10 times. For each trial, record whether algorithm finished successfully or not.

Step 2 - Calculate a statistic

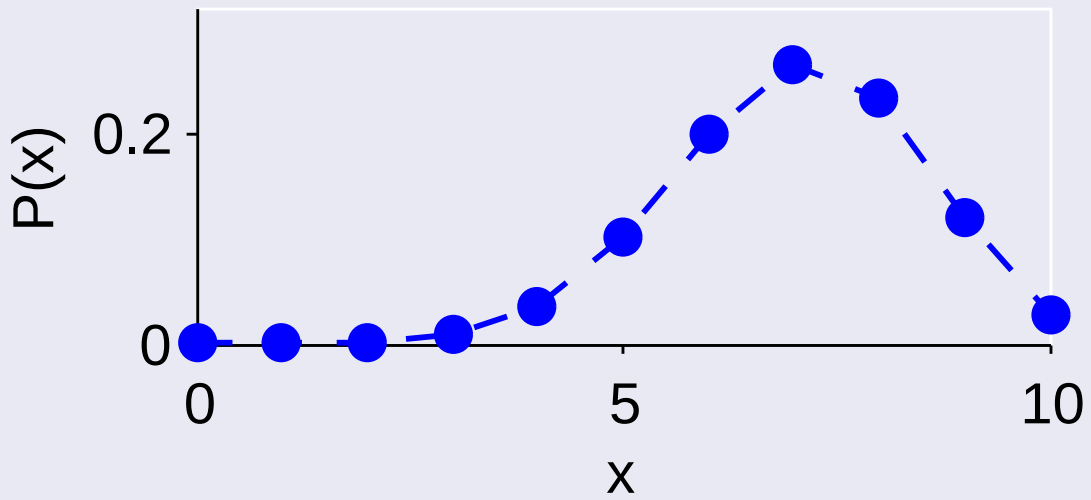
Statistic

X = the number of algorithm trial for which the algorithm completed successfully. We observed the algorithm finishing successfully 4 times.

Step 3 - Determine the probability of observed (or more extreme) value

Distribution

Distribution of X is Binomial with $N = 10$ and $p = 0.7$.



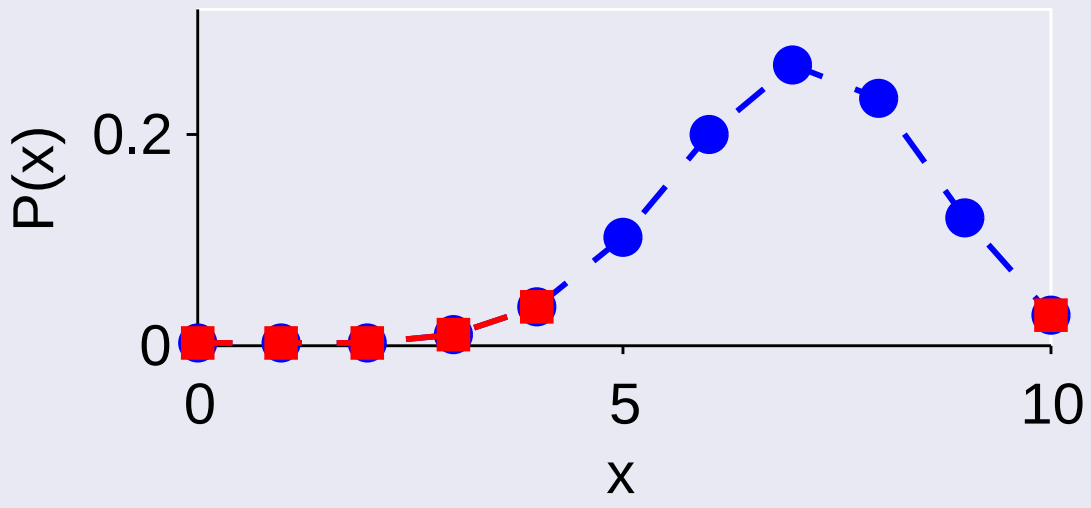
Step 3 - Determine the probability of observed (or more extreme) value

Observed Probability

$$\begin{aligned}\mathbb{P}(X = 4) &= \binom{10}{4}(0.7)^4(0.3)^6 \\ &= \frac{10!}{6!4!}(0.7)^4(0.3)^6 \\ &\approx 0.0368\end{aligned}$$

Step 3 - Determine the probability of observed (or more extreme) value

. . . Or More Extreme Probability



$$\mathbb{P}\left(X = 4 \text{ or more extreme}\right) \approx 0.0756$$

Step 4 - Is the probability small?

Decision

Is 0.0756 a small probability?
It depends, but traditionally this isn't quite small enough to show that there is a difference in the robustness of ScandiTest2.

Slightly More Formal Process

1. Define null and alternative hypotheses
2. Decide an appropriate statistic (to be calculated from the observed data)
3. Assuming null hypothesis to be true, determine the statistic's probability distribution
4. Identify values for the statistic that are unlikely ('extreme') if the hypothesis is true
5. If the observed statistic falls in this critical region, reject the hypothesis; otherwise, accept it

Step 1 - Hypotheses

Definition (Null Hypothesis - H_0)

Often the 'default' or current state of knowledge which is retained unless there is good evidence to the contrary.
It is normally convenient to express this as an equality.

Definition (Alternative Hypothesis - H_1)

A competing hypothesis - usually a hypothesis put forward as new knowledge.

Example

H_0: $p = 0.7$
H_1: $p \neq 0.7$

Step 2 - Test Statistic

Choice of test statistic depends on:

- What data is being observed
- What assumptions can be made about the data (underlying probability distributions etc.)
- Form of the hypotheses

We often use standard hypothesis test in which this decision has already been made.

Example

X = the number of algorithm trial for which the algorithm completed successfully.

Step 3 - Test Statistic Distribution

In reality, we consider this in conjunction with steps 1 and 2:

- In step 1, choose a null hypothesis which enables the distribution to be determined completely.
- In step 2, choose a test statistic whose distribution is relatively easy to calculate.

Again if we use a standard test, the distribution of test statistic will be calculated for us.

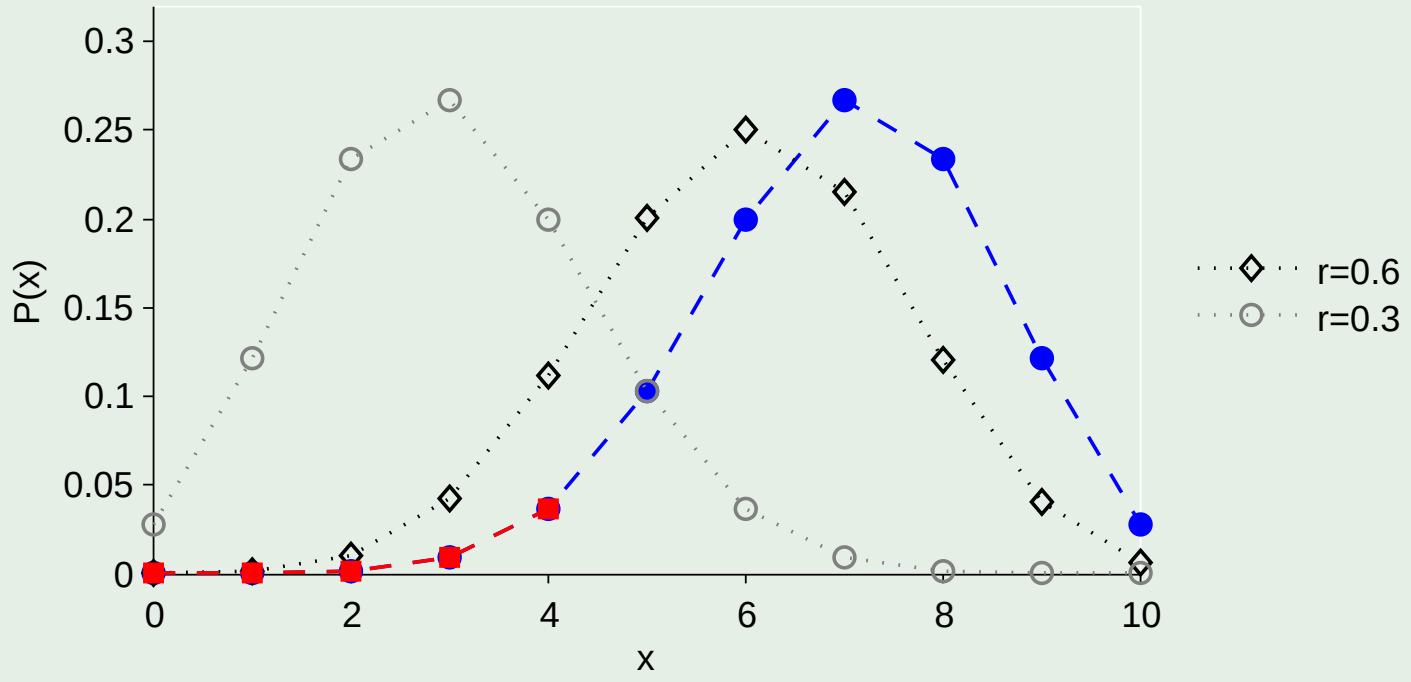
Example

X has a Binomial Distribution with $N = 10$, $p = 0.7$

Step 4 - Unlikely Test Statistic Values

More precisely: identify values where the alternative hypothesis is much more likely to be true than the null hypothesis.

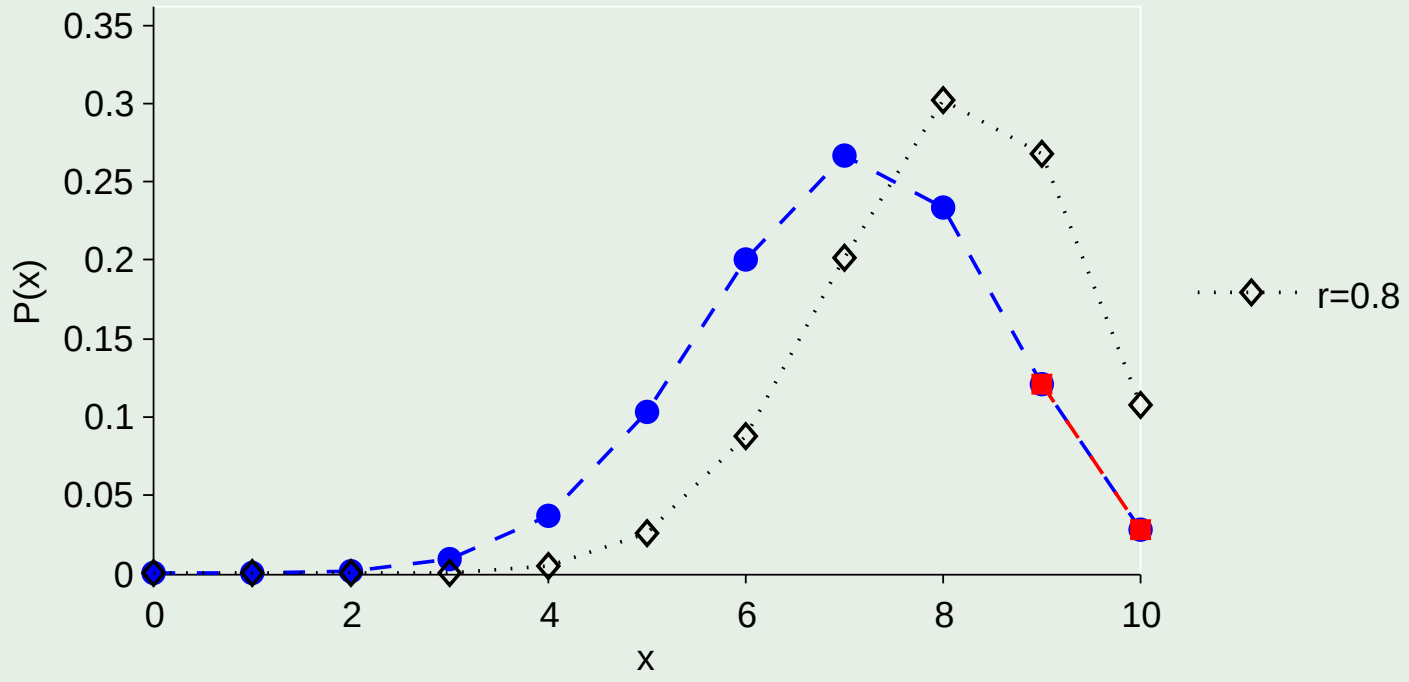
Example



Step 4 - Unlikely Test Statistic Values

More precisely: identify values where the alternative hypothesis is much more likely to be true than the null hypothesis.

Example



Step 5 - Decision Rule

Decision Rule

If test statistic is in critical region, reject H_0 (and accept H_1). Otherwise accept H_0 (and reject H_1).

- Hypothesis Testing
- Significance and Power
- Variants of Hypothesis Test
- Standard Tests
- Multiple Comparisons

Errors

Definition (Type I Error)

H_0 is really true, but it is rejected by the test.
. . . a type of false positive

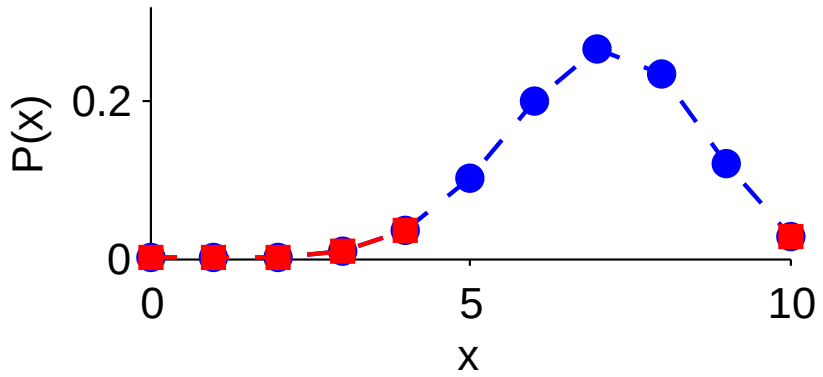
Definition (Type II Error)

H_0 is really false, but it is accepted by the test.
. . . a type of false negative

Well-designed tests attempt to minimise both type I and type II errors.

Type I Errors

Occur when test statistic (by unfortunate chance) happens to fall in the critical region.



Definition (Significance)

If total probability for values in the critical region is α, then this is the chance of a type I error. It is the significance of the test.

p-value

Analysis tools will return the p-value of a hypothesis test.

Definition (p-value)

The p-value is the probability of the observed data – or data more 'extreme' – occurring by chance given that null hypothesis is true.

If the p-value is smaller than the significance of the test, then the observed data is inside the critical region, and the null hypothesis can be rejected.

Example

$$\mathbb{P}(X = 4 \text{ or more extreme}) \approx 0.0756$$

p-value Thresholds

Set an *a priori* threshold

Choose the significance level of the test before calculating the *p*-value. Traditionally a significance level of 5% is used.

. . . OR . . .

Set no threshold

Report the *p*-value and interpret as part of your overall argument.

A Common Misinterpretation

p-value

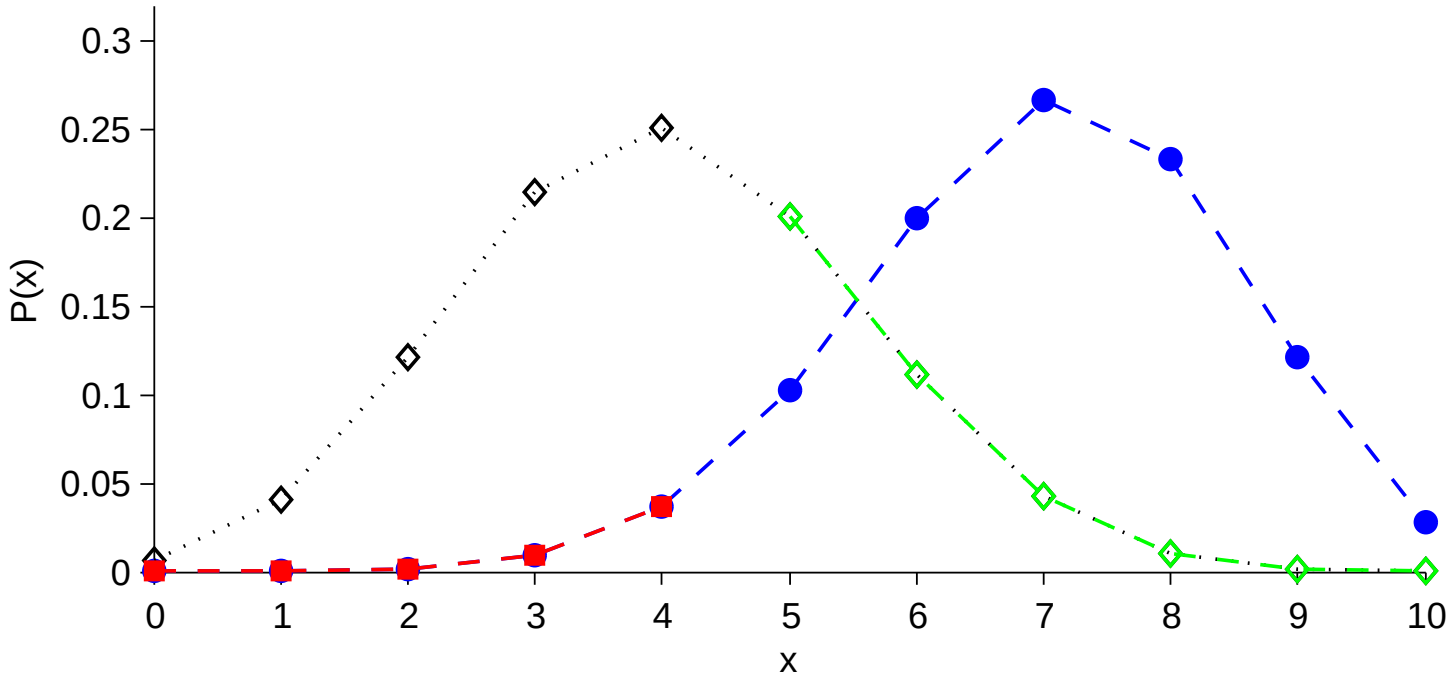
The p-value is not the likelihood that the hypothesis test returns an erroneous result.

Type II Errors

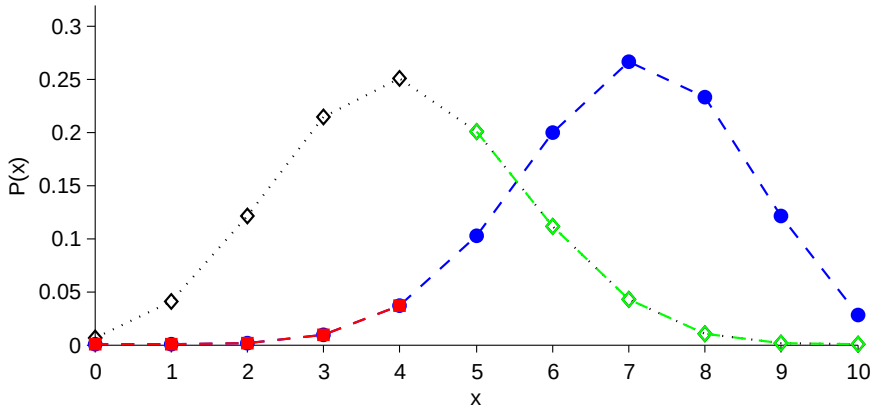
Easiest to consider when H_1 is a simple hypothesis:

Example

H_0: $p = 0.7$
H_1: $p = 0.4$



Type II Errors and Power



The chance of a type II error is denoted β. (In example above, $\beta \approx 0.367$.)

Definition (Power)

The power of a test is the chance that a type II will not occur, i.e. $1 - \beta$.

(The power is equivalent to the probability of values being in the critical region if H_1 is true.)

Best Tests

Typically, the significance α is chosen, and a critical region of this size found that minimises the size of β (or, equivalently, maximises the power). This is called the 'best' test.

The value of β is typically larger than the significance (α). For example, β may be chosen as 20%; or equivalently, a power of 80%.

Standard hypothesis tests have already determined the optimal critical region and therefore the 'best' test.

- Hypothesis Testing
- Significance and Power
- Variants of Hypothesis Test
  - One- and Two-Sample Tests
  - Paired Tests
  - One-Tailed Tests
- Standard Tests
- Multiple Comparisons

One- and Two-Sample Tests

One-Sample Tests

The hypotheses compare a property of one sample against a known value.
Example: H_0: mean run time is 121 seconds.

Two-Sample Tests

The hypotheses compare the same property of two different samples - e.g. the responses of two different algorithms.
Example: H_0: mean run time the two algorithms are the same.

Example 3: ScandiTest versus BritTest Performance

Background

BritTest is the current state-of-the-art algorithm that uses a genetic algorithm.

Research Question

For a given coverage, does ScandiTest achieve this coverage more quickly than BritTest?

Practical

My experiment showed that ScandiTest achieves coverage more quickly than BritTest. H_0 — that the two algorithms have the same performance — was rejected at the 5% significance level. β, the type II error probability, was 30%.

If I now take a new sample of timings, what is the probability that I'll get the same result, i.e. that hypothesis test will again reject H_0?

Practical

Depends on whether H_0 is really correct or not:

If H_0 is really true (the algorithms have the same performance): probability is 5% that the new sample will also reject H_0.

If H_0 is really false (the algorithms have difference performance): probability is 70% that the new sample will also reject H_0

Paired Samples

When comparing two algorithms, the samples may be paired: for corresponding members in each sample, one or more of the controlled factors are the same.

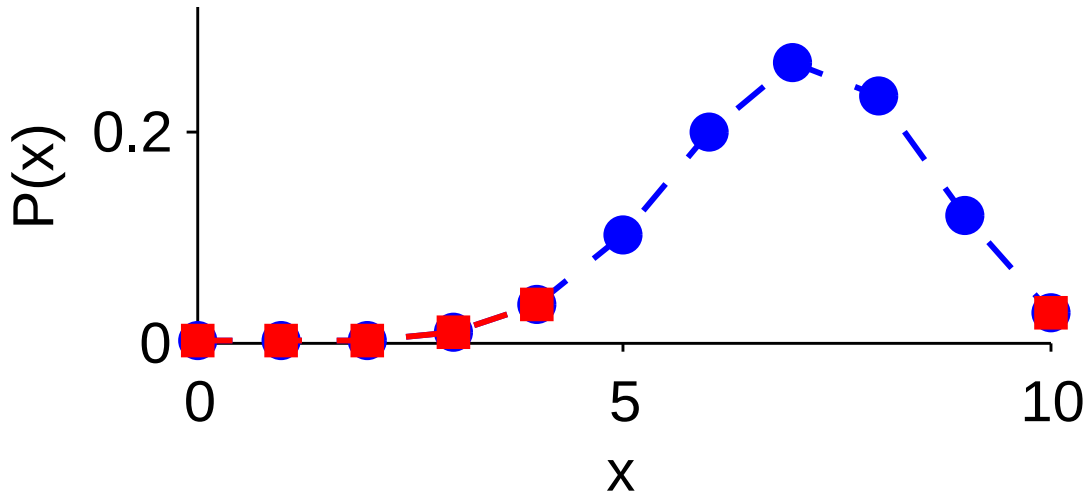
Some standard hypothesis tests can be more sensitive if the samples are known to be paired.

Example

When comparing the performance of ScandiTest and BritTest, the same set of software-under-test is used.

Software	ScandiTest	BritTest
triangle	10.4	9.3
replace	12.3	11.8
tcas	8.7	8.9
⋮	⋮	⋮

One-Tailed Tests



If the alternative hypothesis is not inequality, but instead has a good reason to be either less than or more than, then the critical region should only on the corresponding tail.

Example 2: ScandiTest Robustness

Background

ScandiTest2 is an improved algorithm that is designed to finish more often. Owing to the nature of the changes, we know that the improved algorithm must have a robustness at least as good the original ScandiTest.

Research Question

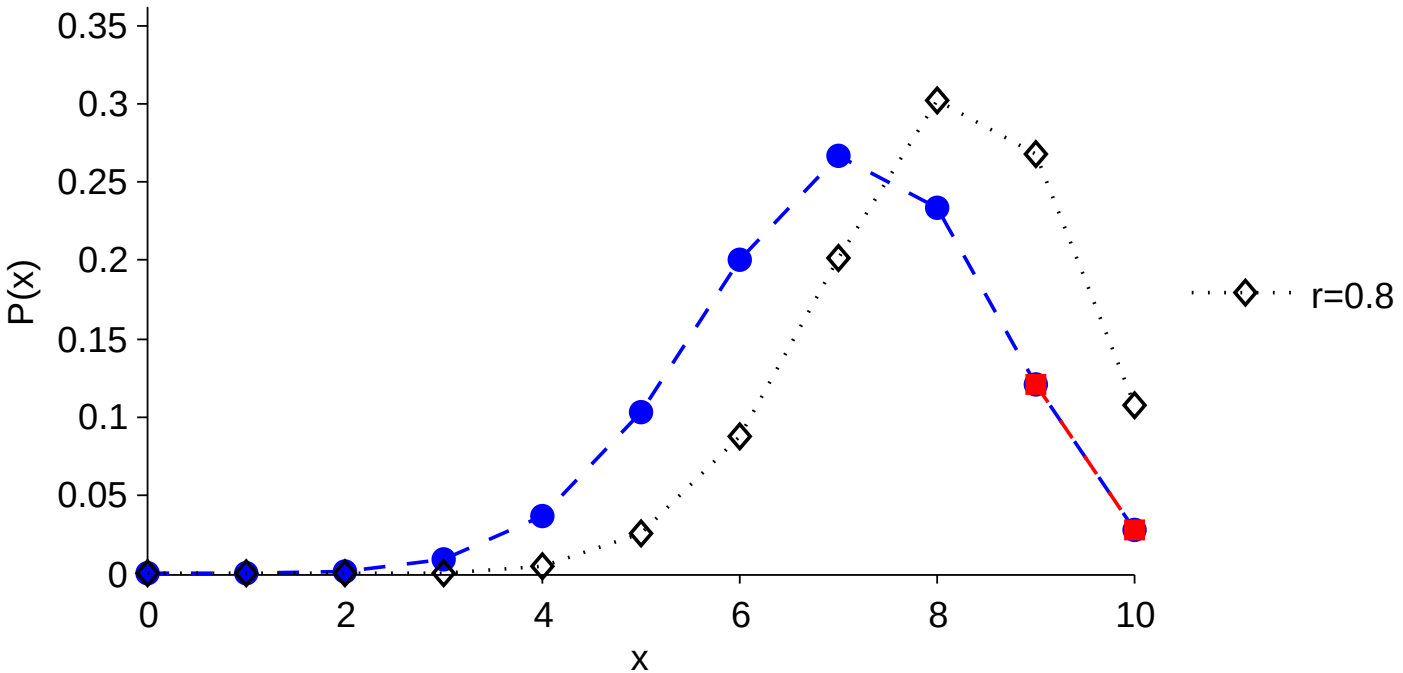
Have the improvements in ScandiTest2 had an effect on robustness?

Example

H_0: $p = 0.7$
H_1: $p > 0.7$

Example 2: ScandiTest Robustness

H_0: $p = 0.7$ H_1: $p > 0.7$



A Word of Caution

For a given significance level (e.g. 5%), a one-tailed test will have a larger critical region on that tail than it would have using a two-tailed test.

two-tailed but should have been one-tailed

The effective significance of the two-tailed test is smaller than the 5% expected.

one-tailed but should have been two-tailed

If the observed data falls on tail containing the critical region, the effective significance is too large. If it falls in the other tail, the effective significance is too small.

- Hypothesis Testing
- Significance and Power
- Variants of Hypothesis Test
- Standard Tests
- Multiple Comparisons

Rank Sum Test (Mann-Whitney-Wilcoxon, also U-test

Characteristics

- non-parametric
- two-sample unpaired

Hypotheses

H_0: Y_A and Y_B are 'stochastically the same'
H_1: Y_A or Y_B is 'stochastically larger'

Example (R)

```
> wilcox.test(y1, y2)
```

What hypotheses does Rank Sum test?

"The null hypothesis is that A and B have the same distribution . . . The alternative hypothesis, H_1, . . . is that A is stochastically larger than B, a directional hypothesis . . . For a two-tailed test, i.e., for a prediction of differences which does not state direction, H_1 would be that $p(a > b) \neq 1/2$" – Siegel, Nonparametric Statistics for the Behavioral Sciences, 1956

"the hypotheses may be stated as: H_0: $F(x) = G(x)$ for all x, H_1: $F(x) \neq G(x)$ for some x. The test is sensitive for H_1: $E(X) \neq E(Y)$ and can be used as a test for means. In many real situations any difference between distributions implies $P(X > Y)$ is no longer equal to $1/2$." – Conover, Practical Nonparametric Statistics, 1999

"the null hypothesis is that the distributions of x and y differ by a location shift of mu and the alternative is that they differ by some other location shift" – R help

"a two-sided rank sum test of the null hypothesis that data in the vectors x and y are independent samples from identical continuous distributions with equal median" – MATLAB help

"a non-parametric test for assessing whether two independent samples come from the same distribution" – Wikipedia

Signed Rank Test (Wilcoxon)

Characteristics

- non-parametric
- one-sample, or two-sample paired

Hypotheses (One Sample)

H_0: median$(Y) = \mu$
H_1: median$(Y) \neq \mu$

Example (R)

```
> wilcox.test(y1, mu=...)
> wilcox.test(y1, y2, paired=TRUE)
```

t-Test (Student) – One Sample

Characteristics

- parametric: assumes a Normal distribution

Hypotheses

H_0: mean of Y is μ
H_1: mean of Y is not μ

Example (R)

```
> t.test(y, mu=...)
```

t-Test (Student) – Two Sample

Characteristics

- parametric: assumes a Normal distribution

Hypotheses

H_0: mean of Y_1 and Y_2 are the same
H_1: mean of Y_1 and Y_2 differ

Example (R)

```
> t.test(y1, y2)
```

Set `var.equal=TRUE` if the variances of the two populations are known to be the same; set `var.equal=FALSE` (the default) for the Welch correction is applied.
Set `paired=TRUE` for paired samples.

Shapiro-Wilk Test of Normality

Characteristics

- one-sample

Hypotheses

H_0: sample is from a Normal distribution

H_1: sample is not form a Normal distribution

Example (R)

```
> shapiro.test(y)
```

Practical (part 1)

Experiment cow38D

Research question: does ScandiTest attain coverage more quickly that BritTest?
Method: 80 SUTs (software-under-test) were randomly split into two sets, one set of 40 for each algorithm. The time taken by the algorithm to attain 80% coverage of each SUT in its set was measured.

1. Download the datasets cow38D.dat and import the data into R.
2. Perform a hypothesis test (or tests) to analyse whether the data supports our research question.
3. Create side-by-side boxplots using the parameter `notch=TRUE`.

Practical (part 2)

Experiment cow38E

Research question: does ScandiTest use more memory that BritTest?

Method: 40 SUTs were chosen at random and each algorithm was run against each of the 40 SUT. The total memory required by the algorithm was measured.

1. Download the datasets cow38E.dat and import the data into R.
2. Perform a hypothesis test (or tests) to analyse whether the data supports our research question.

- Hypothesis Testing
- Significance and Power
- Variants of Hypothesis Test
- Standard Tests
- Multiple Comparisons

Example 4: ScandiTest Parameter Settings

Background

ScandiTest uses a novel bio-inspired optimisation technique called Reindeer Herd Search. There are three categorical parameters that can be on or off: rednose, heavysnow, pullingsleigh.

Research Question

The coverage of ScandiTest does not depend on any of these parameter settings.

Possible Experiment Method

1. Perform three individual experiments – one for each parameter – that compares the paremeter on and off.
2. Compare the samples using three hypothesis tests, each one at the 5% significance level.
   - H_0: no difference between rednose on and off; H_1 ...
   - H_0: no difference between heavysnow on and off; H_1 ...
   - H_0: no difference between pullingsleigh on and off; H_1 ...
3. If any one of the three tests rejects its H_0, then reject research question hypothesis.

A Problem

Assume that in reality none of the parameters have an effect, i.e. each of the H_0 is true.

Then each hypothesis tests, considered individually, has a probability of 5% of rejecting H_0

The probability that one or more of the three tests rejects its null hypothesis is:

$$\begin{aligned}\mathbb{P}\left(\geq 1 \text{ test rejects } H_0\right) &= 1 - (1 - 0.05)^3 \\ &\approx 0.143\end{aligned}$$

So taken as a whole, our family of tests has a significance level of only 14.3%.

Bonferroni Correction

Definition (Bonferroni Correction)

The significance level of each individual test is reduced by a factor of $\frac{1}{N}$ where N is the number of tests.

Example

For the three tests of the nasal_colour parameter, the significance level would be reduced to $\frac{5\%}{3} \approx 1.667\%$.
Take as a whole, the overall family of tests now has a significance level of:

$$\begin{aligned}\mathbb{P}(\geq 1 \text{ test rejects } H_0) &= 1 - (1 - 0.01667)^3 \\ &\approx 0.049\end{aligned}$$

Conservative

Especially when there are many tests the Bonferroni correction is conservative.

Benjamini-Hochberg Procedure

False Discovery

A false discovery is an individual hypothesis test that incorrectly reject the null hypothesis (i.e. a false positive).

Benjamini-Hochberg Procedure

The Benjamini-Hochberg procedure determines which individual tests should be regarded as significant based on an acceptable false discovery rate (i.e. the proportion of significant tests that are false positive) that the experimenter chooses.

False Discovery Rate

The false discovery rate is not the significance of the overall family of tests. It is a measure of how many false positive are acceptable to the experimenter - for example, depending on how easy it is to follow up on significant tests using further experiments.

Benjamini-Hochberg Procedure - Example

1. Decide on an acceptable false discovery rate, Q (e.g. 20%)
2. Sort the p-value of the individual tests in ascending order.
3. For each, calculate a critical value of $\frac{iQ}{N}$ where i is the index of the sorted p-value (starting at 1), and N the number of tests.
4. Find the largest p-value that is lower than its corresponding critical value.
5. That test and all tests with smaller p-values are considered significant (even if the p-value is not less than its corresponding critical value).

Benjamini-Hochberg Procedure - Example

$Q = 20\%,\ N = 10$

i	p-value	$\frac{iQ}{N}$
1	0.62%	2%
2	2.9%	4%
3	6.3%	6%
4	7.7%	8%
5	13.2%	10%
6	27.8%	12%
7	40.5%	14%
8	66.9%	16%
9	72.2%	18%
10	76.2%	20%

Multiple Comparisons in R

Example (Bonferroni Correction)

```
> p.adjust(c(0.09,0.21,0.01),"bonferroni")
[1] 0.27 0.63 0.03
```

Example (Benjamini-Hochberg Procedure)

```
> p.adjust(c(0.0062,0.029,0.063,0.077,0.132,0.278,0.405,0.669,0.722,0.762),"BH")
  [1] 0.0620000 0.1450000 0.1925000 0.1925000 0.2640000 0.4633333 0.5785714
  [8] 0.7620000 0.7620000 0.7620000
```

Key Points

- Hypothesis testing is a principled method of comparison.
- Small p-values indicate that the observed effect are unlikely to be a result of chance.
- There are a number of standard non-parametric and parametric tets.
- When multiple comparisons are made the significance levels must be adjusted.

Part V

Effect Size

- Effect Size
- Standardised Effect Size
- Relationship to Statistical Significance
- Calculating Standardised Effect Sizes

Effect Size

Definition (Effect Size)

The magnitude of the change in the response as a result of different ‘treatments’.

Example 3: ScandiTest versus BritTest Performance

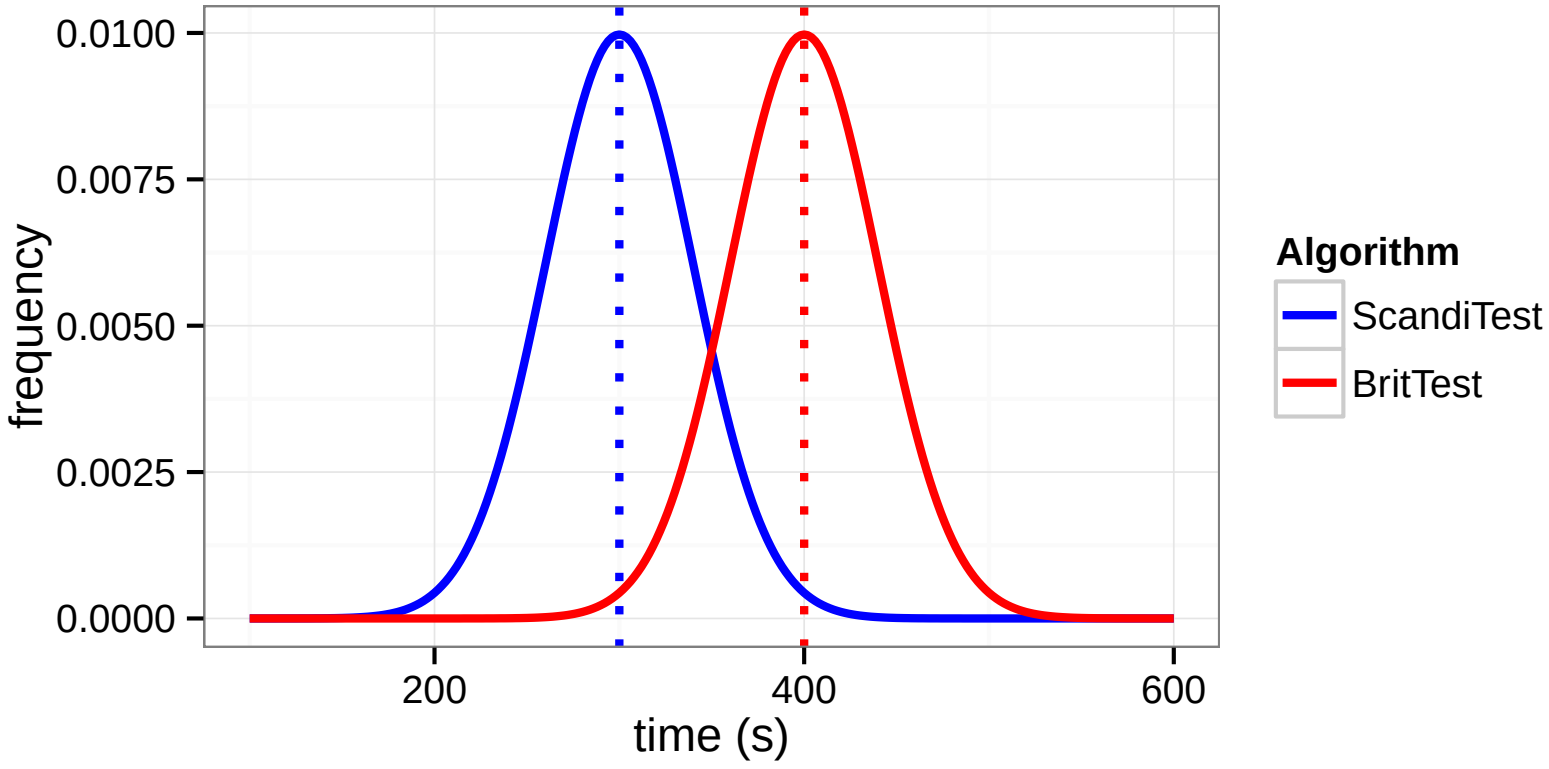
Background

BritTest is the current state-of-the-art algorithm that uses a genetic algorithm.

Research Question

For a given coverage, does ScandiTest achieve this coverage more quickly than BritTest?

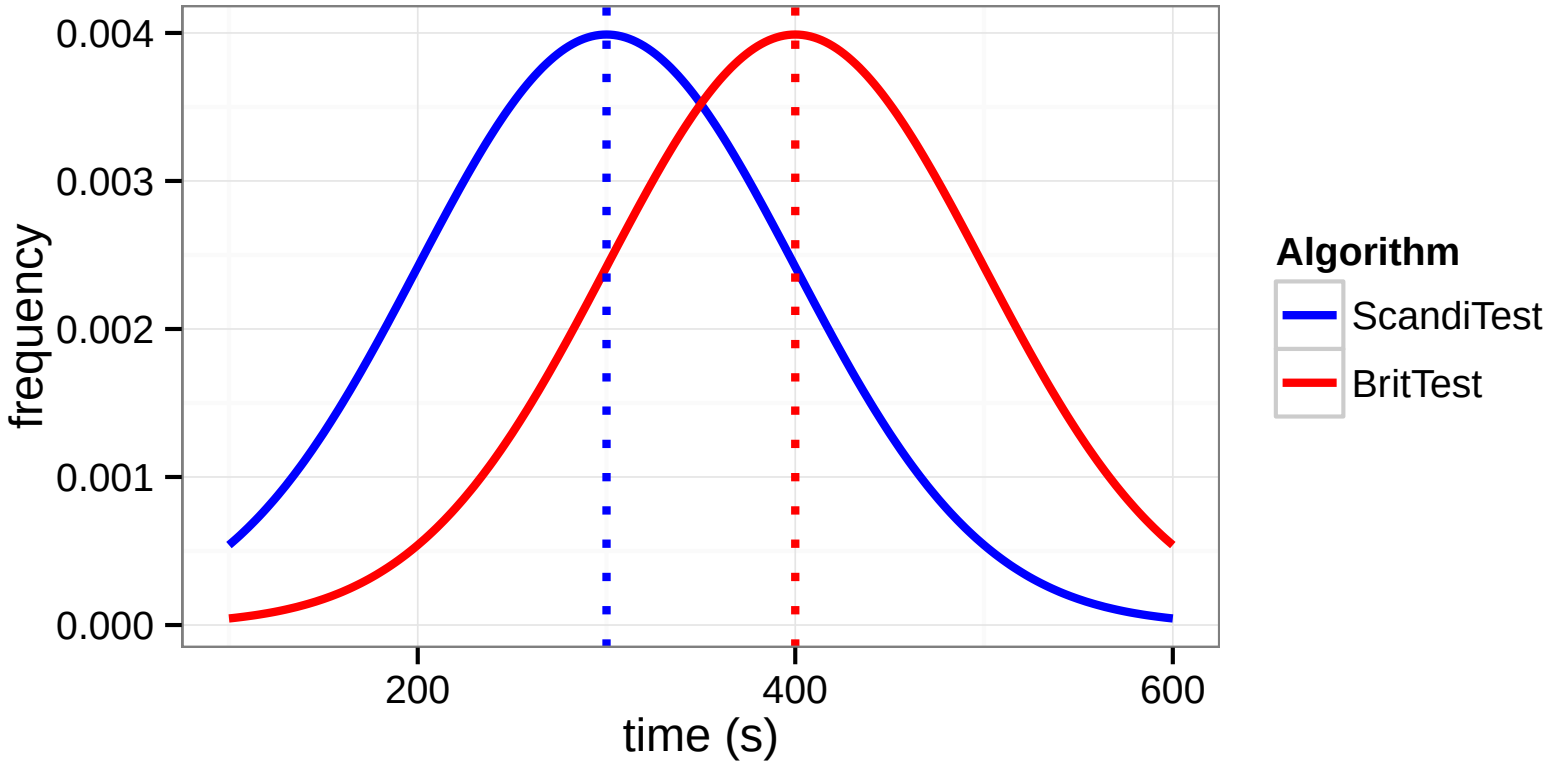
Experiment Results



Difference in mean time to achieve coverage is 100 seconds.

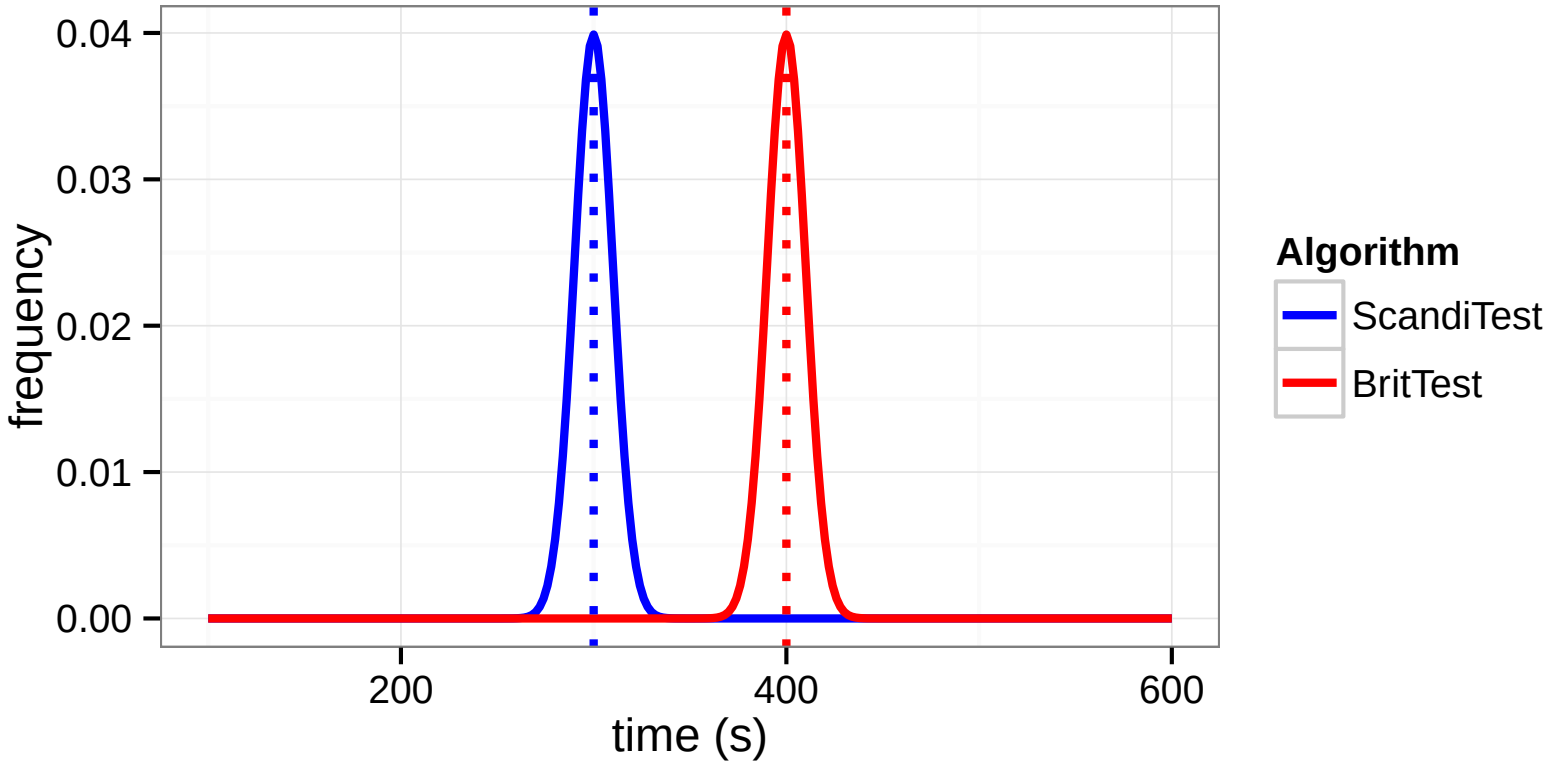
- Effect Size
- Standardised Effect Size
- Relationship to Statistical Significance
- Calculating Standardised Effect Sizes

Experiment Results



Difference in mean time to achieve coverage is 100 seconds.

Experiment Results



Difference in mean time to achieve coverage is 100 seconds.

Standardised Effect Size

Definition (Standardised Effect Size)

Standardised effect size measures normalise the effect in relation to the variability in the data, and calculate a scale-free measure that may be used to compare effect sizes across multiple different experiments.

Cohen's d

$$d = \frac{\overline{y_1} - \overline{y_1}}{s}$$

where $\overline{y_1}$, $\overline{y_2}$ are the sample means, and s the (pooled) sample variance

0.2	'small'
0.5	'medium'
0.8	'large'

Vargha-Delaney A

A has a nice real-world interpretation: it is the probability that if a single response is sampled from algorithms 1 and 2, then the value of algorithm 1's response will be the larger one.

0	algorithm 2 is always the larger
0.29	'large' (if favour of algorithm 2)
0.36	'medium'
0.44	'small'
0.5	no difference between algorithms
0.56	'small'
0.64	'medium'
0.71	'large' (in favour of algorithm 1)
1	algorithm 1 is always the larger

- Effect Size
- Standardised Effect Size
- Relationship to Statistical Significance
- Calculating Standardised Effect Sizes

Effect Size and Statistical Significance

Best Practice

Statistical analysis should report both results of hypothesis tests and effect size.

Relationship

Effect size and statistical significance are different metrics: an experiment may be statistically significant but still have a small effect size.

- Effect Size
- Standardised Effect Size
- Relationship to Statistical Significance
- Calculating Standardised Effect Sizes

'Long Form' Data in R

Example

```
> y1 <- c(10.3, 11.2, 5.3)
> y2 <- c(23.0, 25.1, 17.3)
> df <- data.frame(Y1=y1, Y1=y2)
> df
    Y1       Y2
1 10.3     23.0
2 11.2     25.1
3  5.3     17.3
> library(reshape2)
> melt(df, measure.vars=1:2)
  variable value
1       Y1  10.3
2       Y1  11.2
3       Y1   5.3
4       Y2  23.0
5       Y2  25.1
6       Y2  17.3
```

Effect Size in R

Example (Cohen's D)

```
> mdf <- melt(df, measure.vars=1:2)
> library(effsize)
> cohen.d(mdf$value, mdf$variable)
```

Example (Vargha-Delaney's A)

```
> mdf <- melt(df, measure.vars=1:2)
> library(effsize)
> VD.A(mdf$value, mdf$variable)
```

Practical (part 1)

Experiment cow38D

Research question: does ScandiTest attain coverage more quickly that BritTest?
Method: 80 SUTs (software-under-test) were randomly split into two sets, one set of 40 for each algorithm. The time taken by the algorithm to attain 80% coverage of each SUT in its set was measured.

1. Download the datasets cow38D.dat and import the data into R.
2. Calculate standardised effect size(s) for this dataset.

Practical (part 2)

Experiment cow38E

Research question: does ScandiTest use more memory that BritTest?

Method: 40 SUTs were chosen at random and each algorithm was run against each of the 40 SUT. The total memory required by the algorithm was measured.

1. How would you calculate the effect size for this experiment?

Key Points

- The effect size is a measure of the magnitude.
- Standardised effect sizes measures compare the effect size to the variability in the data.
- Effect sizes should be reported in addition to p-values.

Part VI

Confidence Intervals

- Statistics as Random Variables
- Confidence Intervals
- Bootstrapping

Statistics are Random Variables Too

Example

Measured 100 samples, each of 10 responses, and calculated the sample mean, $\overline{Y}$ of each sample:

	Sample	$\overline{Y}$
1	7.4 10.0 13.5 15.3 7.5 16.5 15.1 10.1 8.5 17.8	12.17
2	4.2 7.0 7.1 9.1 18.1 26.7 8.2 5.7 21.2 1.1	10.84
⋮	⋮	⋮
100	14.9 10.2 6.5 18.6 11.9 13.9 11.4 23.3 11.2 11.7	13.35

Example 1: ScandiTest Quality

Background

ScandiTest is a search-based algorithm for creating tests for structural coverage.

Research Question

What is the average coverage achieved by ScandiTest?

- Statistics as Random Variables
- Confidence Intervals
- Bootstrapping

Confidence Intervals

Often we want to use the sample mean, $\overline{Y}$, to estimate the population mean, μ_Y, of the response Y. We take a set of observations, and calculate a single value of the sample mean: $\overline{y}$.

But since $\overline{Y}$ is a random variable, the particular value $\overline{y}$ is unlikely to be equal to the population mean of Y.

Confidence intervals are a way of expressing how 'close' we expect the true population mean to be to our estimate, $\overline{y}$.

Sample Mean Probability Distribution

Following results hold for any distribution of Y with population mean μ and population variance σ^2.

Theorem

For the sample mean $\overline{Y}$ calculated from samples of size n:

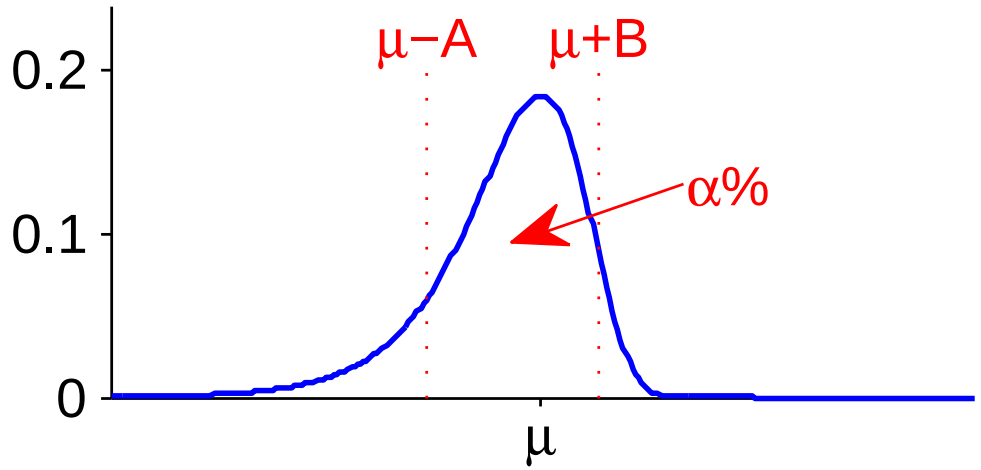
$$\mu_{\overline{Y}} = \mu_Y$$

$$\sigma^2_{\overline{Y}} = \frac{1}{n}\sigma^2_Y$$

Confidence Intervals - Principle I

1. Want a confidence interval for the mean of the response, μ_Y. Since $\mu_Y = \mu_{\overline{Y}}$, this is equivalent to finding a confidence interval for $\mu_{\overline{Y}}$. For simplicity, call the value μ.
2. Derive a range (expressed in terms of μ) that contains a certain percentage, α, of the distribution of $\overline{Y}$, i.e.:

$$\mathbb{P}(\mu - A \leq \overline{Y} \leq \mu + B) = \alpha$$



Confidence Intervals - Principle II

3. Calculate a single sample mean value, $\overline{y}$.
4. If we can show that α percent of the time:

$$\mu - A \leq \overline{y} \leq \mu + B$$

   Equivalently, α percent of the time:

$$\overline{y} - B \leq \mu \leq \overline{y} + A$$

5. This gives a confidence interval for μ.

Confidence Intervals - Example I

Sample									
3.7	12.6	13.4	6.3	18.0	17.9	11.8	13.6	12.9	11.1

1. Find the 95% confidence interval for the mean of the response Y.

   We may assume:

   - response has a Normal distribution
   - variance of response, σ^2_Y, is 25

Confidence Intervals - Example II

2. We make use of the following mathematical result:

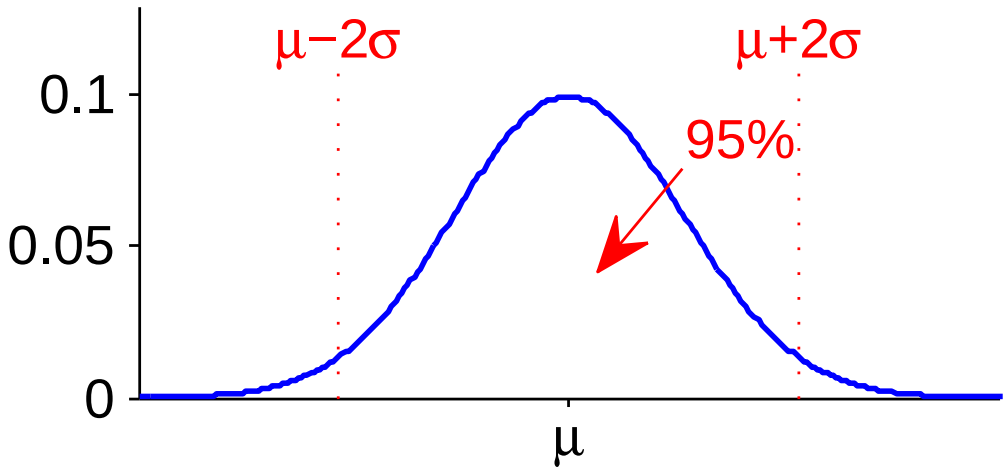
Distribution of the Sample Mean

In general, the distribution of the sample mean, $\overline{Y}$, is not the same as the distribution of Y. But when Y has a Normal distribution, so does $\overline{Y}$.

$\overline{Y}$ is therefore Normally distributed with variance $25/10 = 2.5$.

Confidence Intervals - Example III

For a Normal distribution, (approximately) 95% of the distribution is contained within 2 standard deviations of the mean.



Confidence Intervals - Example IV

Now,

$$\sigma_{\overline{Y}} = \sqrt{2.5} \approx 1.5811$$

Giving,

$$\mathbb{P}(\mu - 3.16 \leq \overline{Y} \leq \mu + 3.16) = 0.95$$

3. From sample, calculate $\overline{y} = 12.13$.
4. So 95% percent of the time:

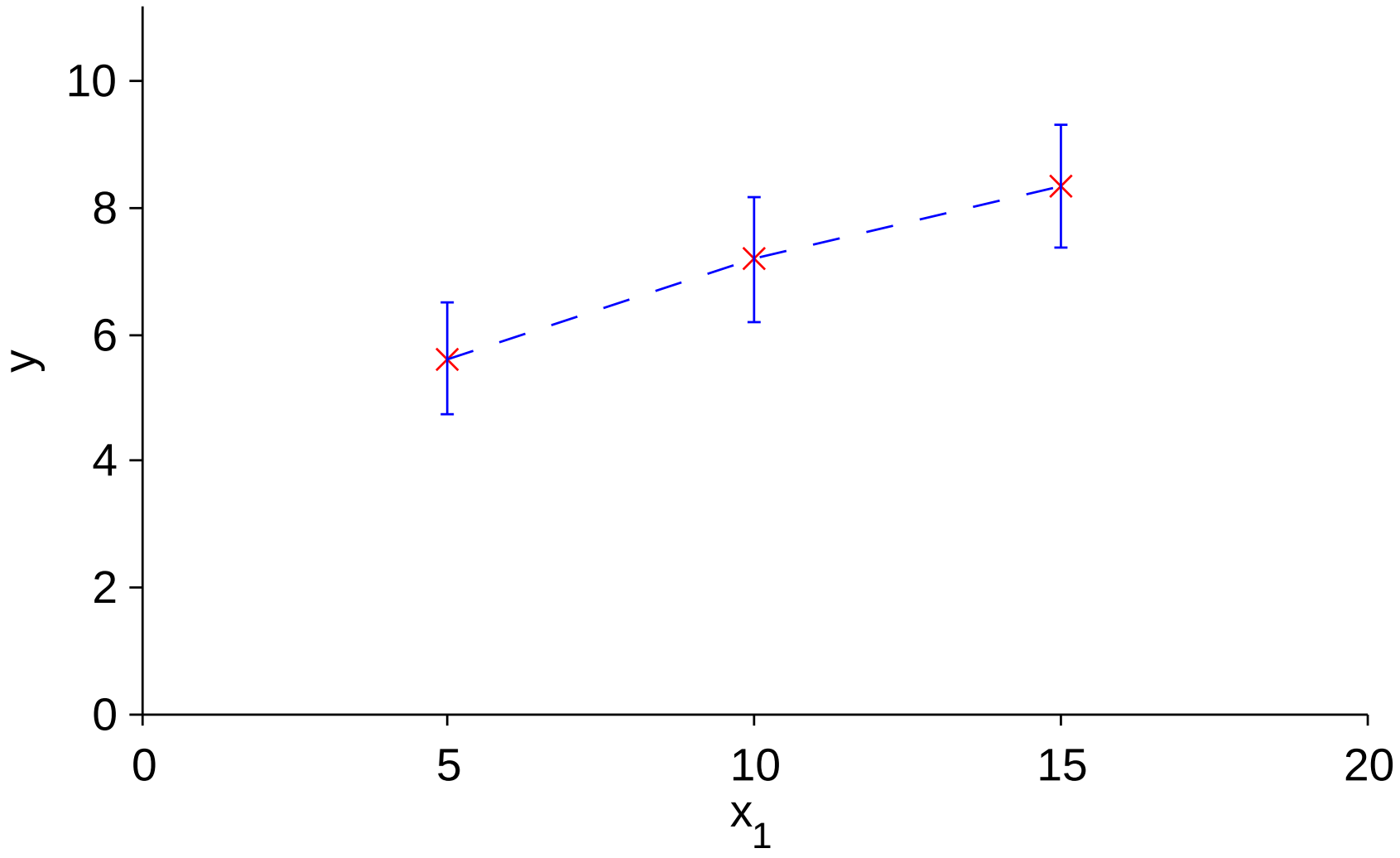
$$\mu - 3.16 \leq 12.13 \leq \mu + 3.16$$

Equivalently, 95% percent of the time:

$$12.13 - 3.16 \leq \mu \leq 12.13 + 3.16$$

5. This gives a 95% confidence interval for μ of 12.13 ± 3.16

Error Bars



- Statistics as Random Variables
- Confidence Intervals
- Bootstrapping

General Principle

We want to estimate a population parameter θ (e.g. the mean μ) and use an appropriate estimator, t, calculated from the sample (e.g. the sample mean, $\overline{y}$). Assume the sample is $y_1, y_2, \ldots, y_n$ (size n).

1. From the original sample, draw a new sample of size n, $y_1^*, y_2^*, \ldots, y_n^*$ by picking randomly with replacement
2. Calculate t^*, the value of the estimator for this new sample
3. Repeat m times, to give $t_1^*, t_2^*, \ldots, t_m^*$
4. The set $t_1^*, t_2^*, \ldots, t_m^*$ gives information about the distribution of the estimator, t, e.g.:

$$\overline{t^*} = \frac{1}{m} \sum_{j=1}^{m} t_j^*$$

$$s_{t^*}^2 = \frac{1}{m-1} \sum_{j=1}^{m} (t_j^* - \overline{t^*})^2$$

Simple Example

Original Sample

0.8 0.5 5.3 7.8 9.3 1.3 5.7 4.7 0.1 3.4

$\overline{y} = 3.89$

Bootstrap Samples

j	bootstrap sample	$\overline{y_j^*}$
1	0.5 4.7 7.8 1.3 0.5 5.7 5.3 5.7 5.7 4.7	4.19
2	9.3 0.8 5.3 3.4 0.5 0.1 1.3 3.4 0.8 9.3	3.42
3	0.5 3.4 0.8 4.7 0.1 0.1 0.8 7.8 5.3 0.1	2.36
4	9.3 3.4 0.5 5.3 0.5 0.5 0.1 1.3 1.3 0.5	2.27
5	0.1 5.7 7.8 1.3 9.3 0.8 5.3 0.5 0.5 5.3	3.66

sample mean of $\overline{y^*}$: $\frac{1}{5}\sum_{i=1}^{5}\overline{y_j^*} = 3.18$

sample variance of $\overline{y^*}$: $\frac{1}{4}\sum_{i=1}^{5}(\overline{y_j^*} - 3.18)^2 = 0.702$

Parametric BootStrapping

So far, when picking bootstrap samples, each point in the original sample has had the same chance of being selected. This is non-parametric bootstrapping.

But if distribution for the original sample points is already known (e.g. from theory) and can estimate distribution parameters from original sample, then can assign chances of a sample point being picked for a bootstrap sample according to its probability in this distribution. This is parametric bootstrapping.

Balanced Bootstrapping

Balanced Bootstrapping ensures that the number of times that each original sample point is picked across *all* bootstrap samples is the same for every sample point.

Example

Original Sample: A B C D

Bootstrap Sample 1: B C C A
Bootstrap Sample 2: C D A D
Bootstrap Sample 3: A D B A
Bootstrap Sample 4: C B B D

Bootstrapping Confidence Intervals in R

Example

```
> library(boot)
> y <- c(10.3, 11.2, 5.3, 7.7, 8.0)
> medianfn <- function(d,i) { median(d[i]) }
> y.median.boot <- boot(y, medianfn, 1000)
> boot.ci(y.median.boot, type="perc")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = y.median.boot, type = "perc")

Intervals :
Level     Percentile
95%   ( 5.3, 11.2 )
Calculations and Intervals on Original Scale
```

Practical

Background

Experiment cow38B measured the coverage obtained by a testing tool.

1. Download the dataset cow38B.dat and import into R
2. Calculate a 95% confidence interval for the mean coverage.

Key Points

- Confidence intervals indicate the accuracy of a statistic.
- Bootstrapping is one (non-parametric) method of deriving a confidence interval.

Part VII

Sample Size

- Confidence Intervals
- Effect Size
- Hypothesis Tests
- Central Limit Theorem

Confidence Intervals

The larger the sample size, the tighter the bounds of the confidence interval.

Example

For the sample mean $\overline{Y}$ calculated from samples of size n:

$$\sigma^2_{\overline{Y}} = \frac{1}{n}\sigma^2_Y$$

- Confidence Intervals
- Effect Size
- Hypothesis Tests
- Central Limit Theorem

Effect Size

How does sample size affect the effect size?

- Confidence Intervals
- Effect Size
- Hypothesis Tests
- Central Limit Theorem

Example: ScandiTest Performance

Research Question

Does ScandiTest take 100 seconds or 110 seconds to achieve 80% coverage?

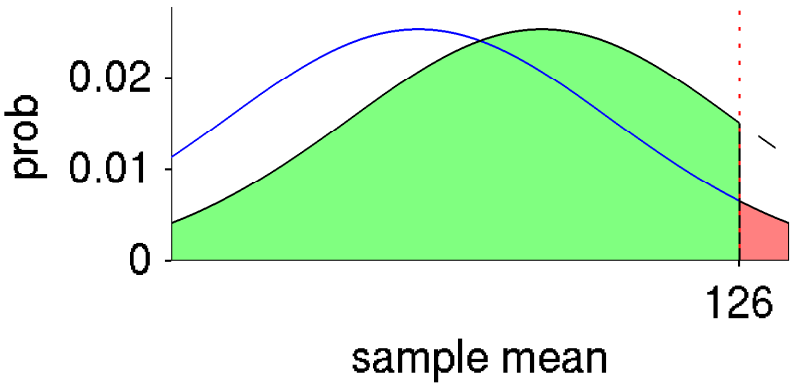
Example

Y = time taken to achieve 80% coverage

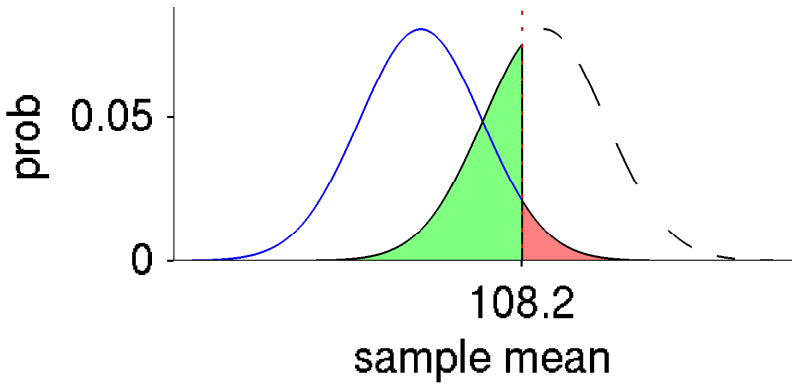
H_0: $\mu_Y = 100$

H_1: $\mu_Y = 110$

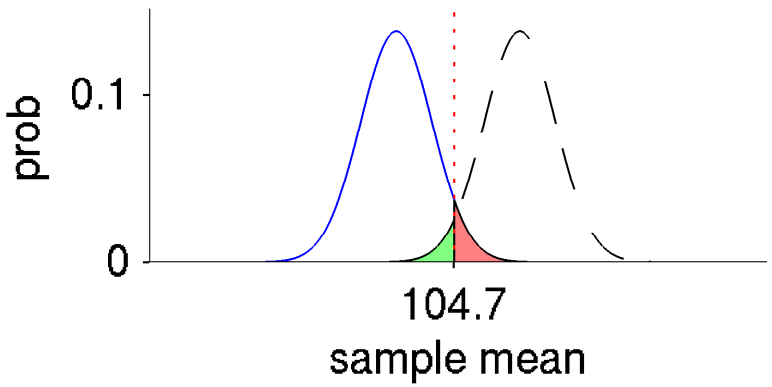
Different Sample Sizes



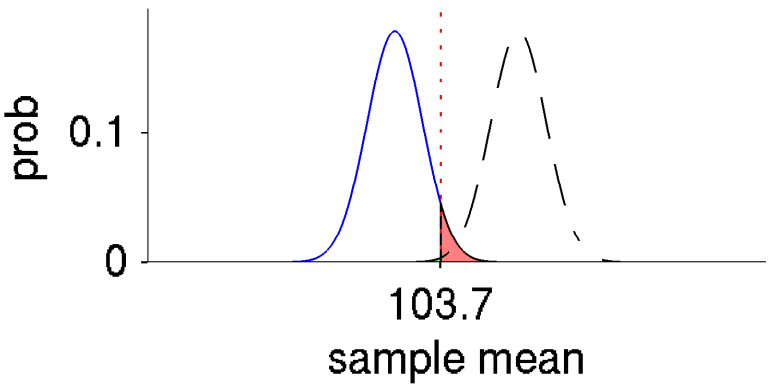
$N = 1 \quad \beta = 0.844$



$N = 10 \quad \beta = 0.361$

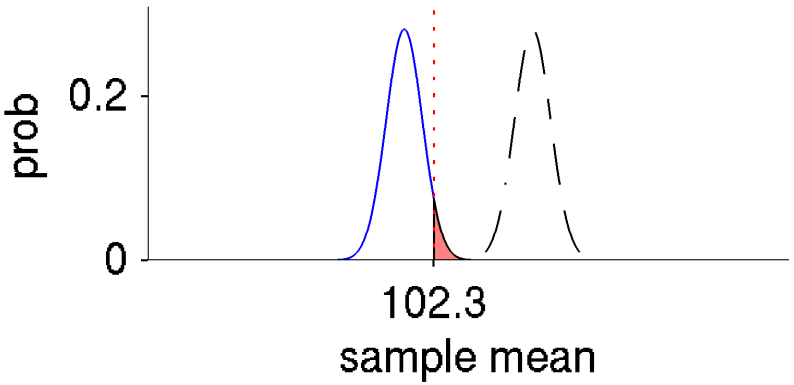


$N = 30 \quad \beta = 0.0344$

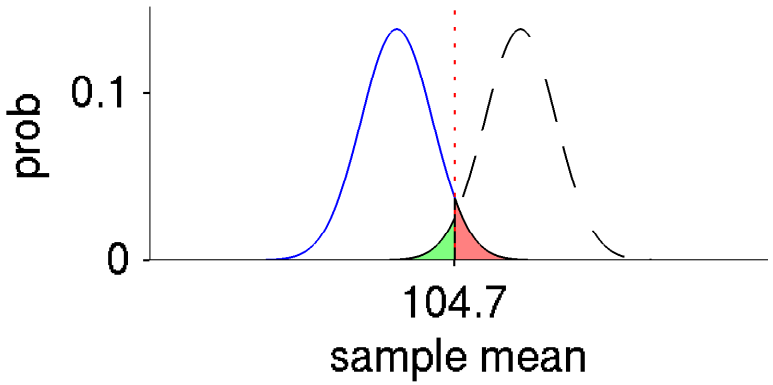


$N = 50 \quad \beta = 0.00235$

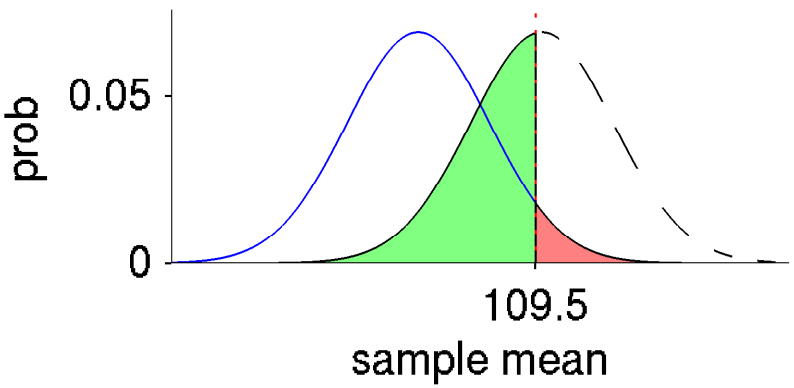
Different Variances ($N = 30$)



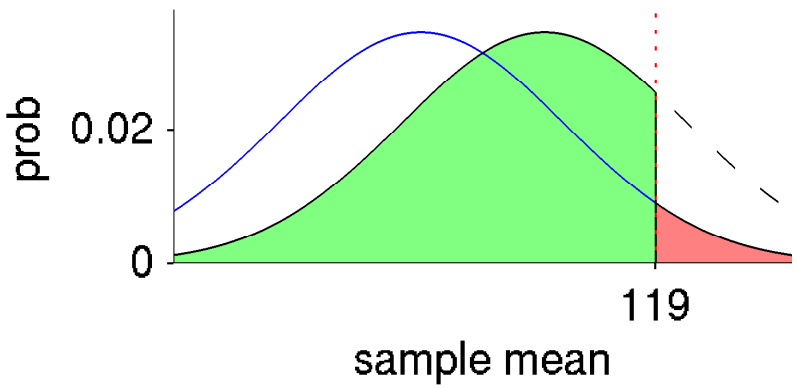
$\sigma_Y^2 = 60 \quad \beta = 2.88 \times 10^{-8}$



$\sigma_Y^2 = 250 \quad \beta = 0.0344$

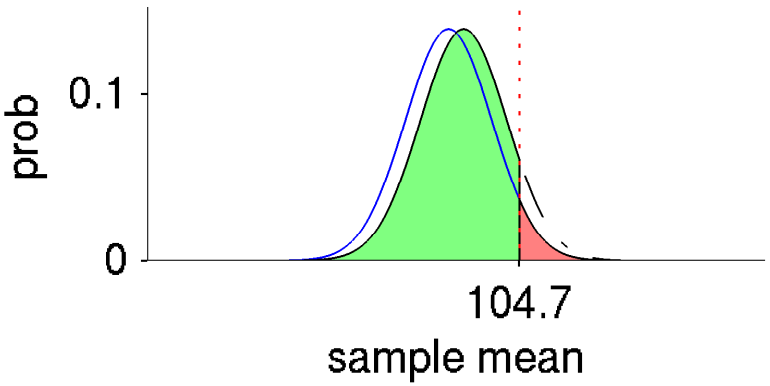


$\sigma_Y^2 = 1000 \quad \beta = 0.467$

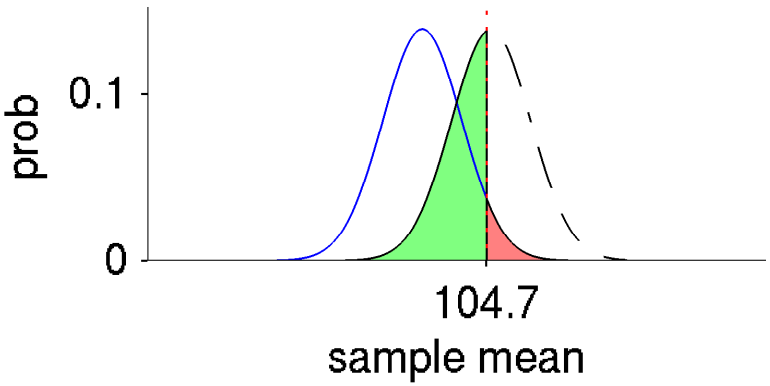


$\sigma_Y^2 = 4000 \quad \beta = 0.782$

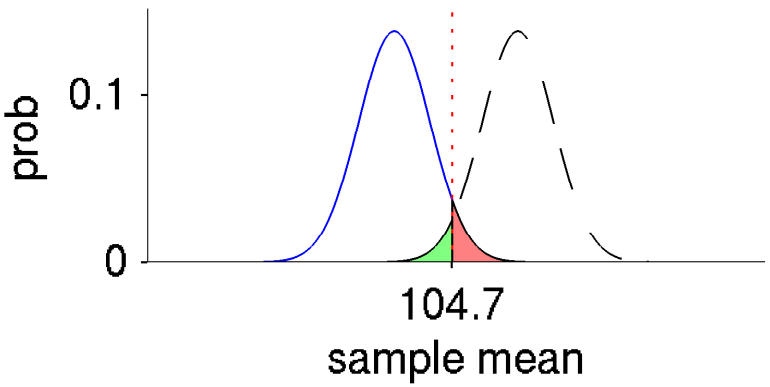
Different Effect Sizes ($N = 30$, $\sigma_Y^2 = 250$)



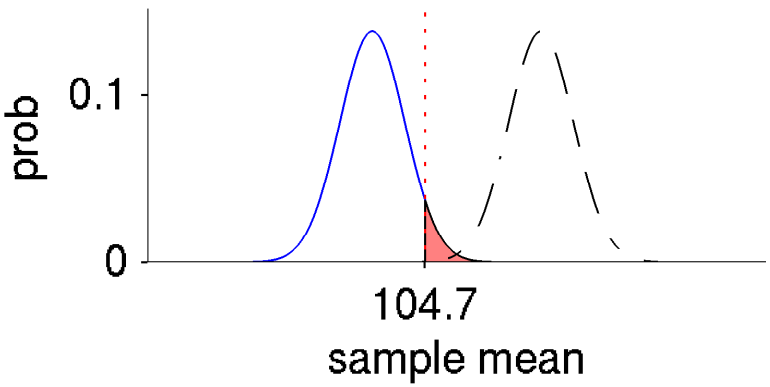
H_1: $\mu_Y = 101$ $\beta = 0.903$



H_1: $\mu_Y = 105$ $\beta = 0.465$

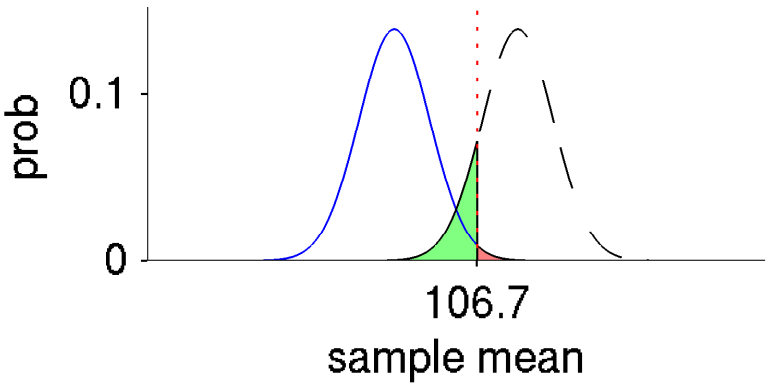


H_1: $\mu_Y = 110$ $\beta = 0.0344$

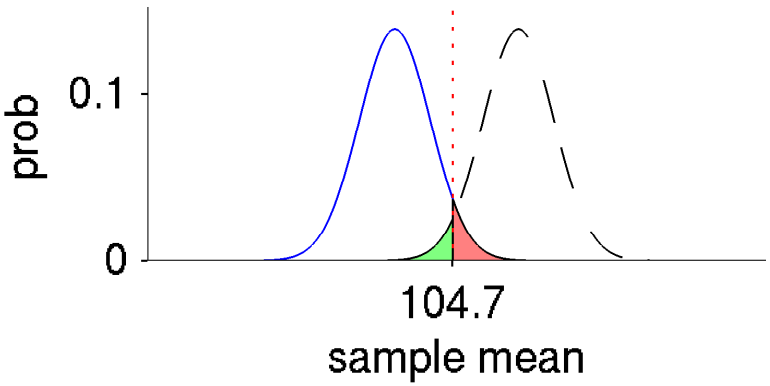


H_1: $\mu_Y = 115$ $\beta = 0.000192$

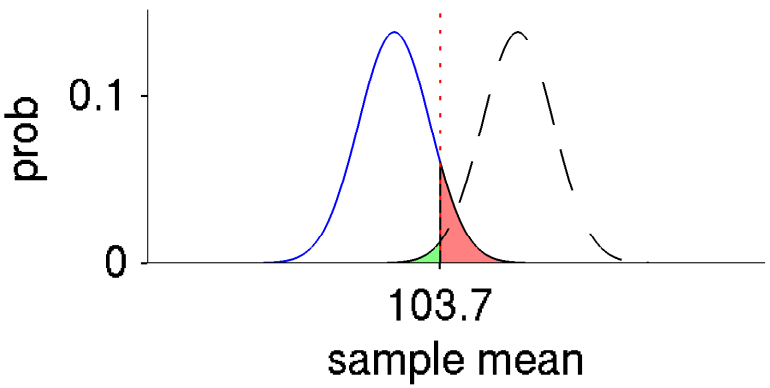
Different Significances ($N = 30$, $\sigma_Y^2 = 250$, H_1: $\mu_Y = 110$)



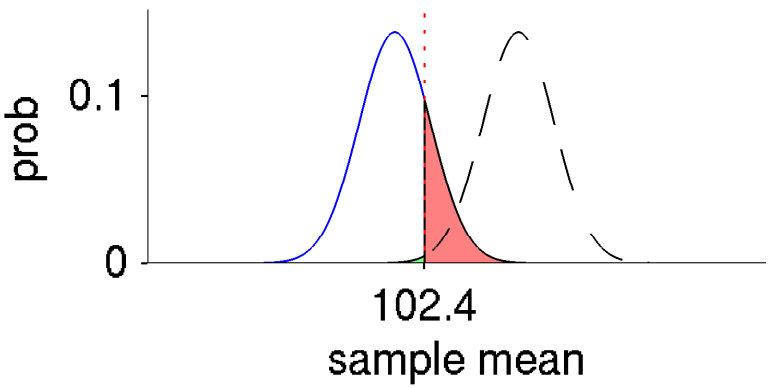
$\alpha = 0.01$ $\beta = 0.128$



$\alpha = 0.05$ $\beta = 0.0344$



$\alpha = 0.10$ $\beta = 0.0145$



$\alpha = 0.20$ $\beta = 0.00436$

Summary

The sample size required to obtain a given power depends on:

- the variance in the data
- the effect size
- the significance level

Estimating Sample Size

a priori

If these factors are known ahead of time, might be able to estimate the sample size required.

post hoc

Otherwise, if (estimates of) the factors are obtained by the test itself (e.g. the variance in the data, or the actual effect size), then can estimate the power after performing the test.

If sample size is inadequate, could take further samples until a sufficient test power is obtained.

Calculating Sample Size and Power in R

R

```
> power.t.test(n=30,power=NULL,delta=10,sd=16,sig.level=0.05)

     Two-sample t test power calculation

              n = 30
          delta = 10
             sd = 16
      sig.level = 0.05
          power = 0.6629097
    alternative = two.sided

NOTE: n is number in *each* group
```

One of `n`, `delta`, `power`, `sd`, `sig.level` must be NULL, and is the value that will be estimated.

Practical

Experiment

Research question: does ScandiTest use more memory that BritTest?
Method: n SUTs will be separately chosen at random for each algorithm, and the total memory required by the algorithm will be measured.

1. Assume the difference in means is 100, and the standard deviation is 230.
2. Estimate the sample size (i.e., the value of n) for a paired t-test to have a power of 80% at a 5% significance level.

- Confidence Intervals
- Effect Size
- Hypothesis Tests
- Central Limit Theorem

Central Limit Theorem

Theorem (Central Limit Theorem)

For independent random variables $Y_1, Y_2, Y_3, \ldots, Y_N$ with any distribution, the distribution of the sum, $\sum_{i=1}^{N} Y_i$, gets closer to a Normal distribution as N increases.

Theorem (Corollary)

For large samples, the sample mean $\overline{Y}$ is approximately Normally distributed for any distribution of Y.

Key Points

- A larger sample size improves the accuracy of confidence intervals.
- A larger sample size does not change the magnitude of the effect size.
- For a given significance, a larger sample size will improve the power of a hypothesis test.
- At larger sample size, some statistics will be approximately Normally distributed.

Part VIII

Correlation

- Multivariate Probability Distributions
- Correlation (Pearson)
- Rank Correlation (Spearman)

Multivariate Probability Distributions

So far have considered population distribution of a single variable - univariate probability distributions.

In this section, we consider distributions of two or more variables - multivariate probability distributions.

Example

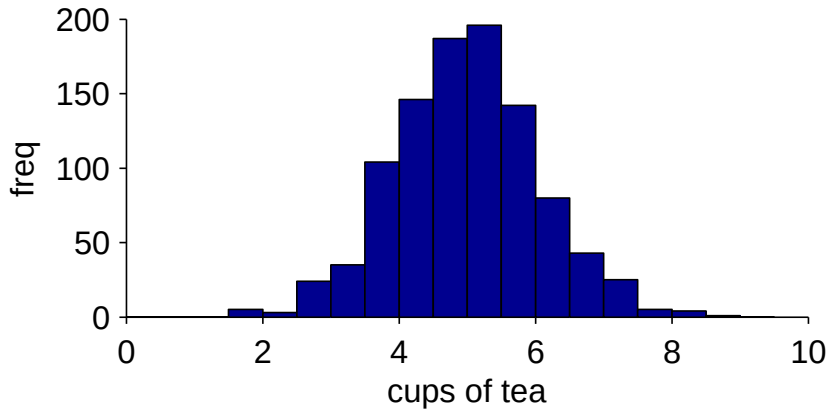
I'm interested in the relationship between IQ and the amount of tea a person drinks.

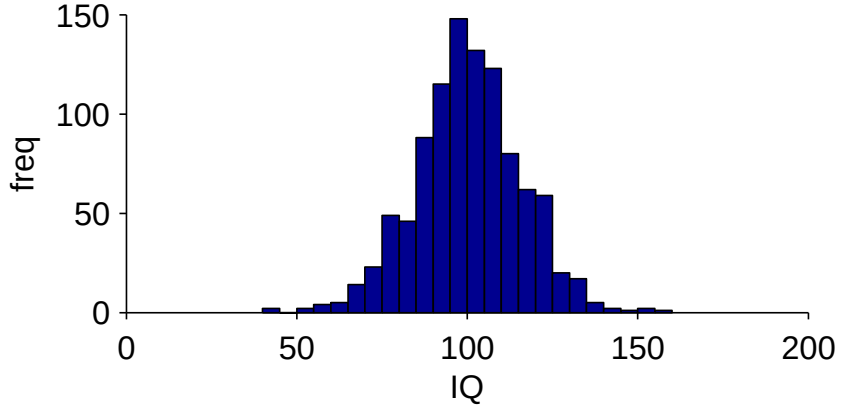
For a sample of 1000 regular tea drinkers, I record the average number of cups they drink per day and also measure their IQ.

Sample Data

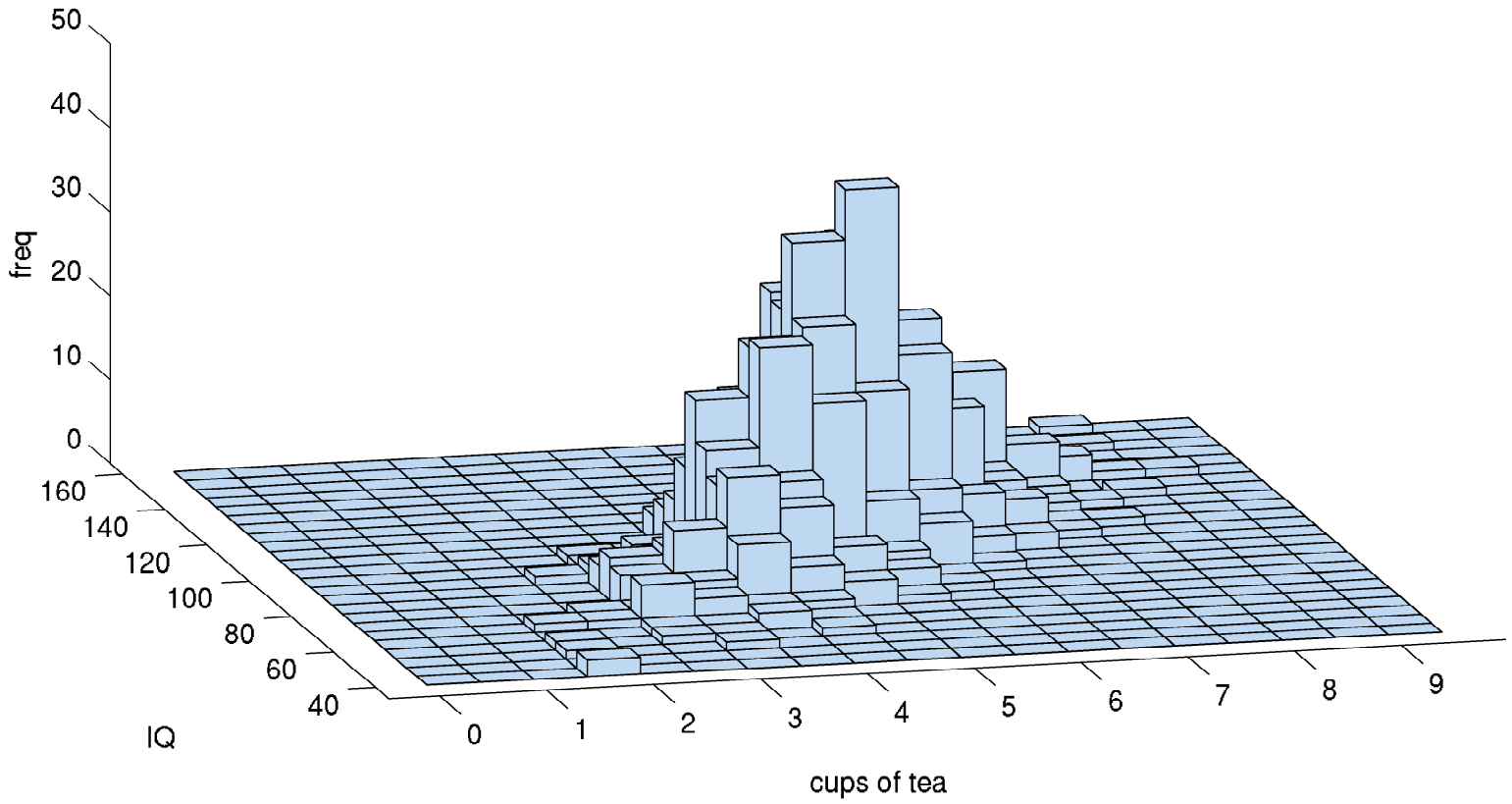
cups of tea	IQ
6.3	109
3.2	73
4.7	95
4.9	114
5.0	90
⋮	⋮

Marginal Distributions





Joint Distribution



Example 6: ScandiTest (Time) Performance

Background

Anecdotal evidence is that the slowest runs of ScandiTest are also the ones that consume most memory.

Research Question

Is the the time performance of ScandiTest correlated with its memory performance?

- Multivariate Probability Distributions
- Correlation (Pearson)
- Rank Correlation (Spearman)

Sample Estimators

Mean

Sample Estimator for μ_X:

$$\overline{X} = \frac{1}{n}\sum x_i$$

Variation

Sample Estimator for σ_X^2:

$$s_X^2 = \frac{1}{n-1}\sum_i (x_i - \overline{X})^2$$

Sample Estimators for Covariance and Correlation

Covariance

$$s_{XY} = \frac{\sum_i (x_i - \overline{X})(y_i - \overline{Y})}{n-1}$$

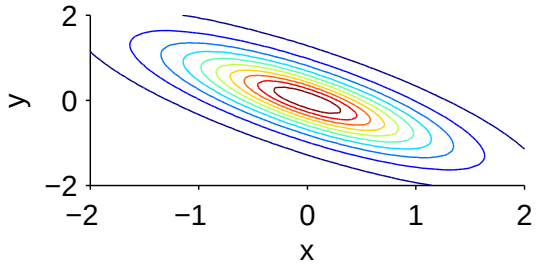
(Pearson) Correlation Coefficient ρ

$$r = \frac{s_{XY}}{s_X s_Y}$$

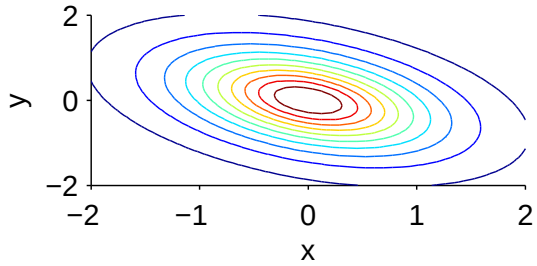
Note: $-1 \leq r \leq 1$

Examples of ρ

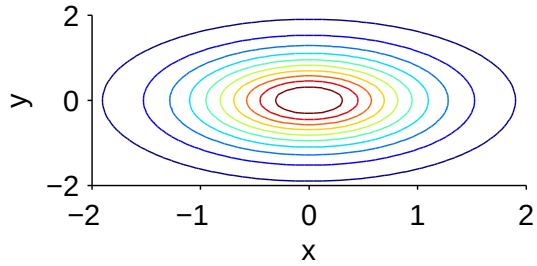
As $|\rho|$ gets closer to 1, X and Y get closer to a linear dependence:



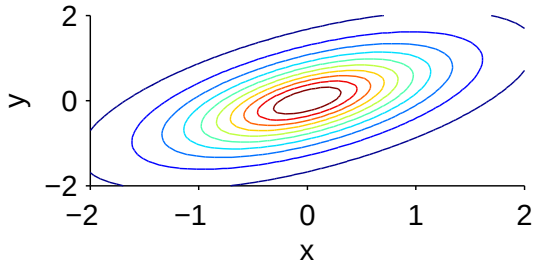
$\rho = -0.8$



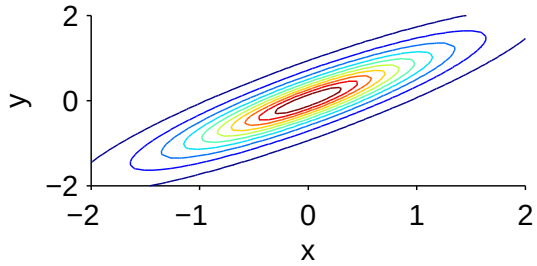
$\rho = -0.4$



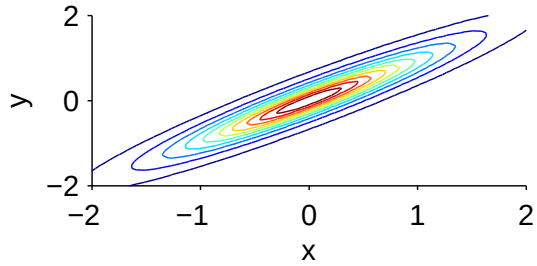
$\rho = 0$



$\rho = 0.6$



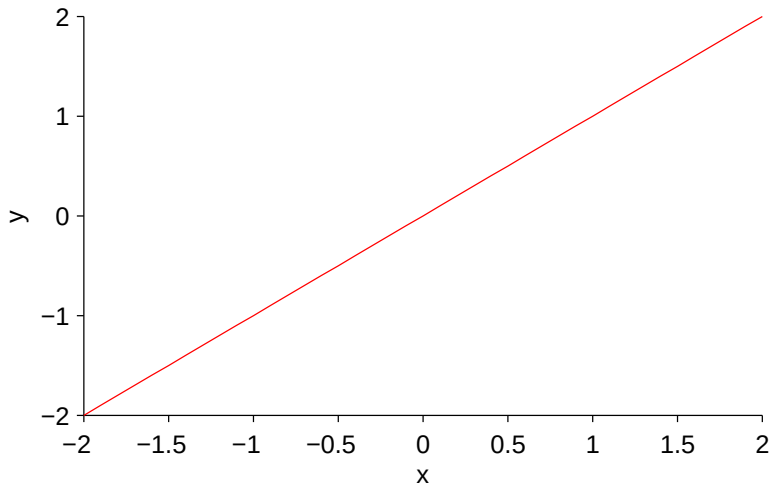
$\rho = 0.9$



$\rho = 0.95$

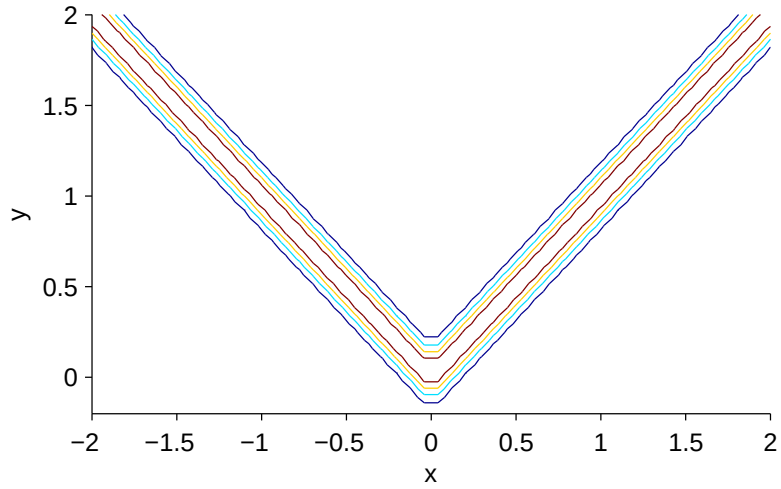
$\rho = \pm 1$

If $\rho = \pm 1$ then X and Y are exactly linearly related: all the probability lies along a line in the x-y plane:

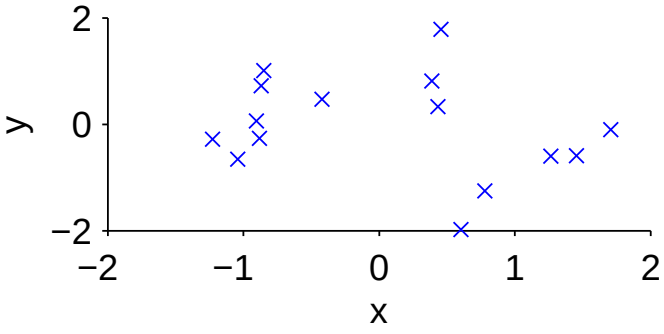


$\rho = 0$

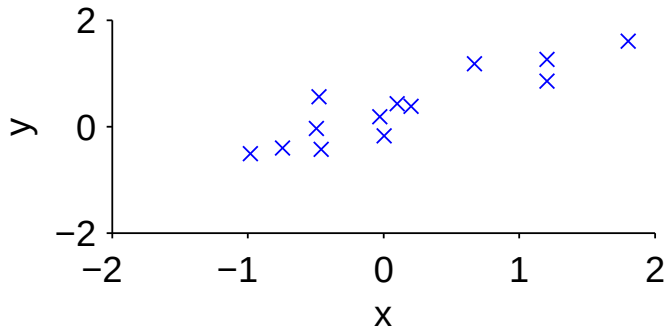
- if X and Y are independent, then $\rho = 0$
- converse does not hold—X,Y can be dependent but nevertheless have $\rho = 0$:



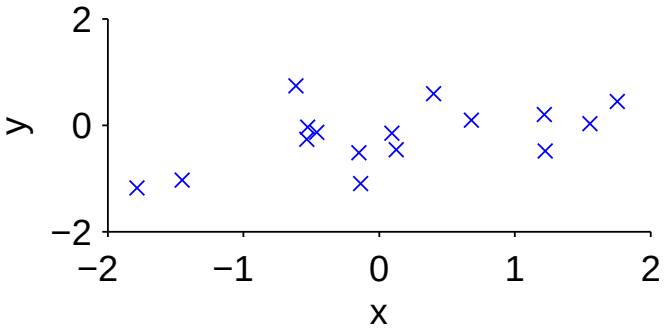
Scatter Plots



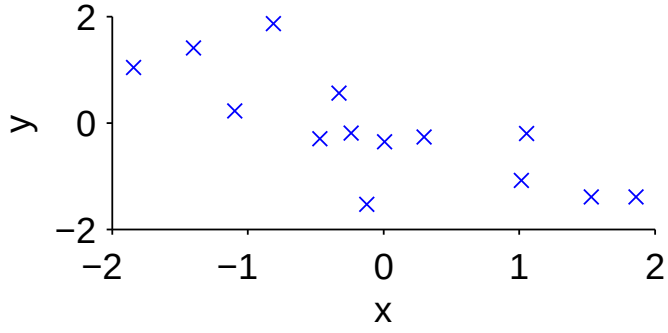
$r = -0.11$



$r = 0.94$



$r = 0.54$



$r = -0.89$

Interpretation of r

- $|r| > 0$ is evidence that there is a linear dependence between the two variables
- the larger the magnitude of r, the closer the dependence is to exactly linear
- $r = 0$ is not necessarily evidence that the variables are independent

But . . .

1. Evidence of a statistical dependence between the variables is *not*, by itself, evidence of a causal relationship
2. Since it is a sample estimator, r is a random variable

- Multivariate Probability Distributions
- Correlation (Pearson)
- Rank Correlation (Spearman)

Spearman's Rank Correlation

1. Rank the observations, x_i of X in order; denote the rank $(1, 2, \ldots, n)$ of observation i as x_i'
2. Do the same for observations, y_i of Y
3. Calculate the (Pearson) correlation between x_i' and y_i': this gives the Spearman Rank Correlation, r' for the sample

Calculating Pearson Correlation Using R

Example (R)

```
x <- c(10.2, 12.4, 8.4, 9.7)
y <- c(3.5, 3.6, 3.1, 3.0)
cor(x,y,method="pearson")
[1] 0.7951368
> cor(x,y,method="spearman")
[1] 0.8
```

Practical

Experiment cow38F

Research Question: Is the the time performance of ScandiTest correlated with its memory performance?
Method: Ran scandiTest against a set of SUTs, and for each measured the time taken, and the memory consumed by the algorithm.

1. Download the dataset cow38F.dat and import into R.
2. Create a scatter plot of the data.
3. Calculate the Pearson and Spearman correlation for this dataset.
4. If you have time, use bootstrapping to calculate a confidence interval for the correlation.

Key Points

- Correlation measures one form of dependency between random variables.
- Correlation does not imply causation.
- Considered two measures: Pearson correlation and Spearman (rank) correlation

Part IX

Linear Models

- Motivation
- Linear Models
- Experimental Designs
- Model Fitting (Analysis)
- ANOVA
- Model Interpretation

Example 5: ScandiTest (Memory) Scalability

Background

A major practical constraint on using ScandiTest is the amount of memory it appears to consume.

Research Question

How does the memory used scale with the number of structural elements in the SUT?

Example 7: ScandiTest Parameters

Background

ScandiTest uses a novel bio-inspired optimisation technique called Reindeer Herd Search. It has 5 numeric parameters.

Research Question

What are the best setting of the parameters?

- Motivation
- Linear Models
  - Linear Models
  - Higher Order Linear Models
- Experimental Designs
- Model Fitting (Analysis)
- ANOVA
- Model Interpretation

Experimental Model

If we are interested in how the mean response changes, we can express this mathematically as:

$$\mu_Y = f(\boldsymbol{x})$$

But $f(\boldsymbol{x})$ could be any function of $\boldsymbol{x}$.

Such a generic function is difficult to analyse, so often assume a simpler model.

Linear Model

Linear Model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

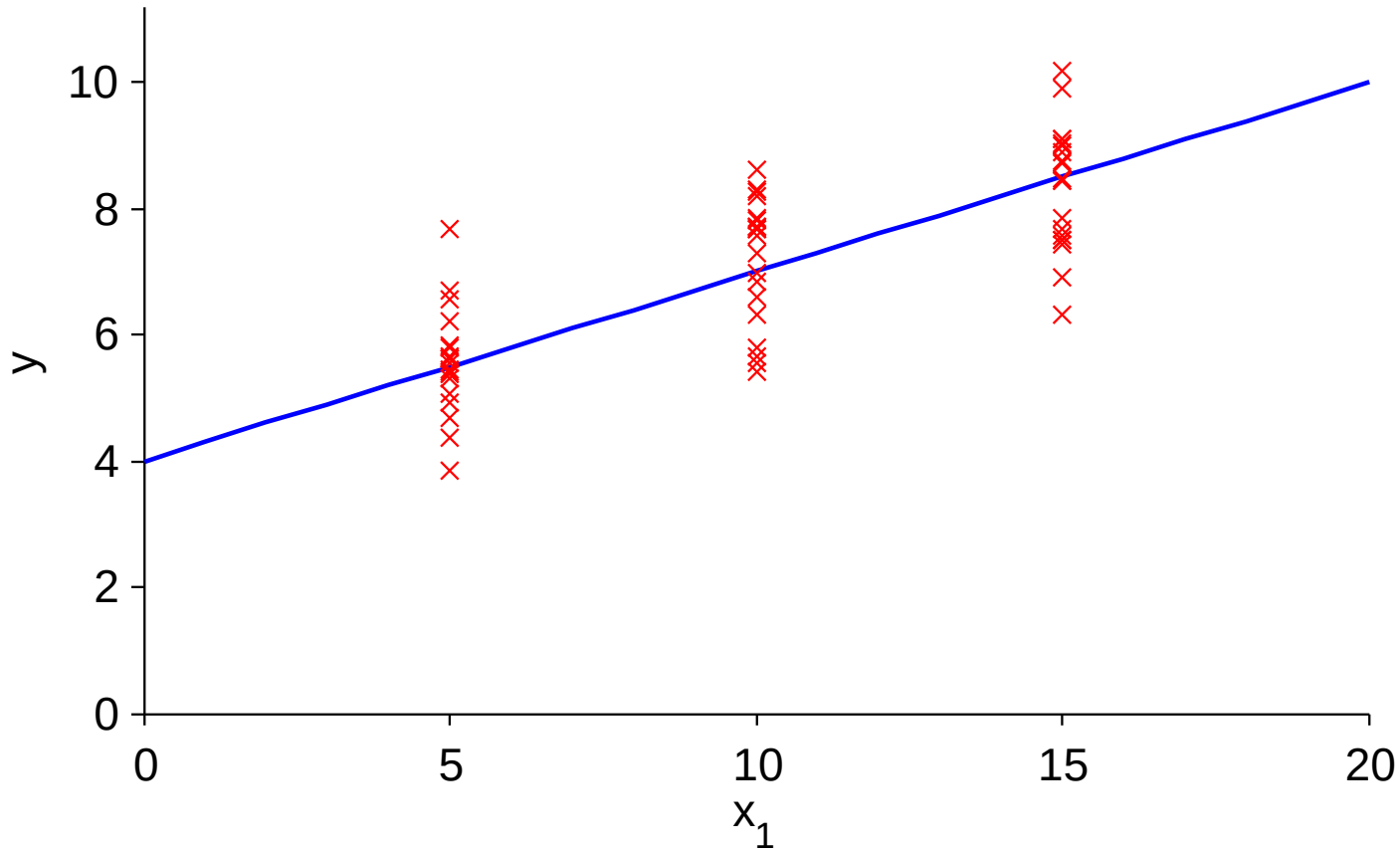
- β_i are model parameters
- β_0 is the intercept
- ε is the noise term

Noise Term Assumptions

Assumes noise term, ε, is a random variable that it is:

independent	Each time a response is measured, the value of the error term is independent of the values it took for previous responses.
identically distributed	The probability distribution is the same regardless of the factor values.
Normally distributed	The distribution is Normal with zero mean (and constant variance σ^2)

Noise Term Assumptions



Linear Model - Mean Response

Linear Model

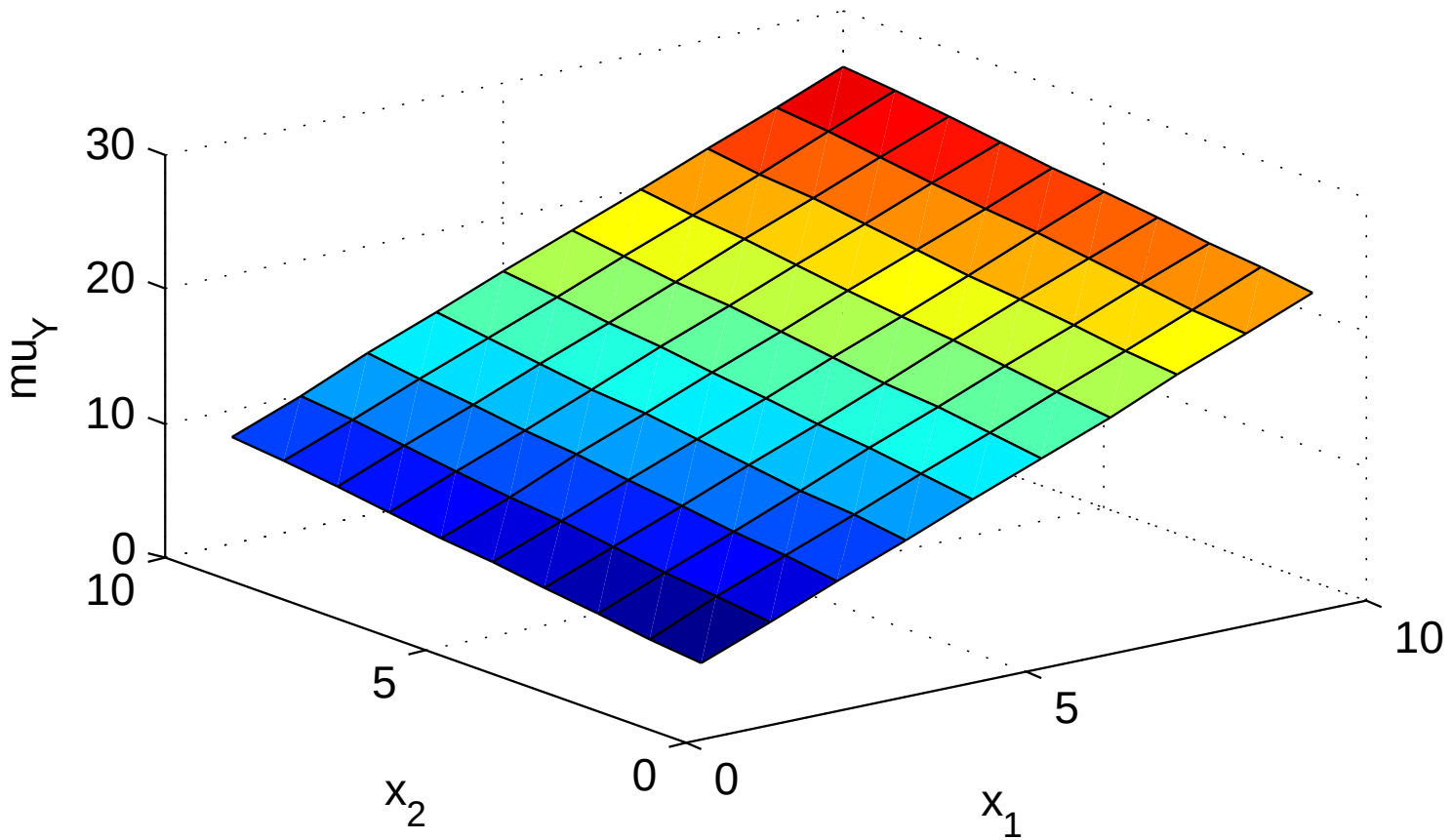
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

The linear terms ($\beta_0 + \sum_i \beta_i x_i$) explain the mean response.
The noise term (ε) explains the variance in the response.

Linear Model

$$\mu_Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Linear Model - 2 Factors



Higher Order Linear Models

Model is linear in terms of model parameters.

Other forms include:

Interaction

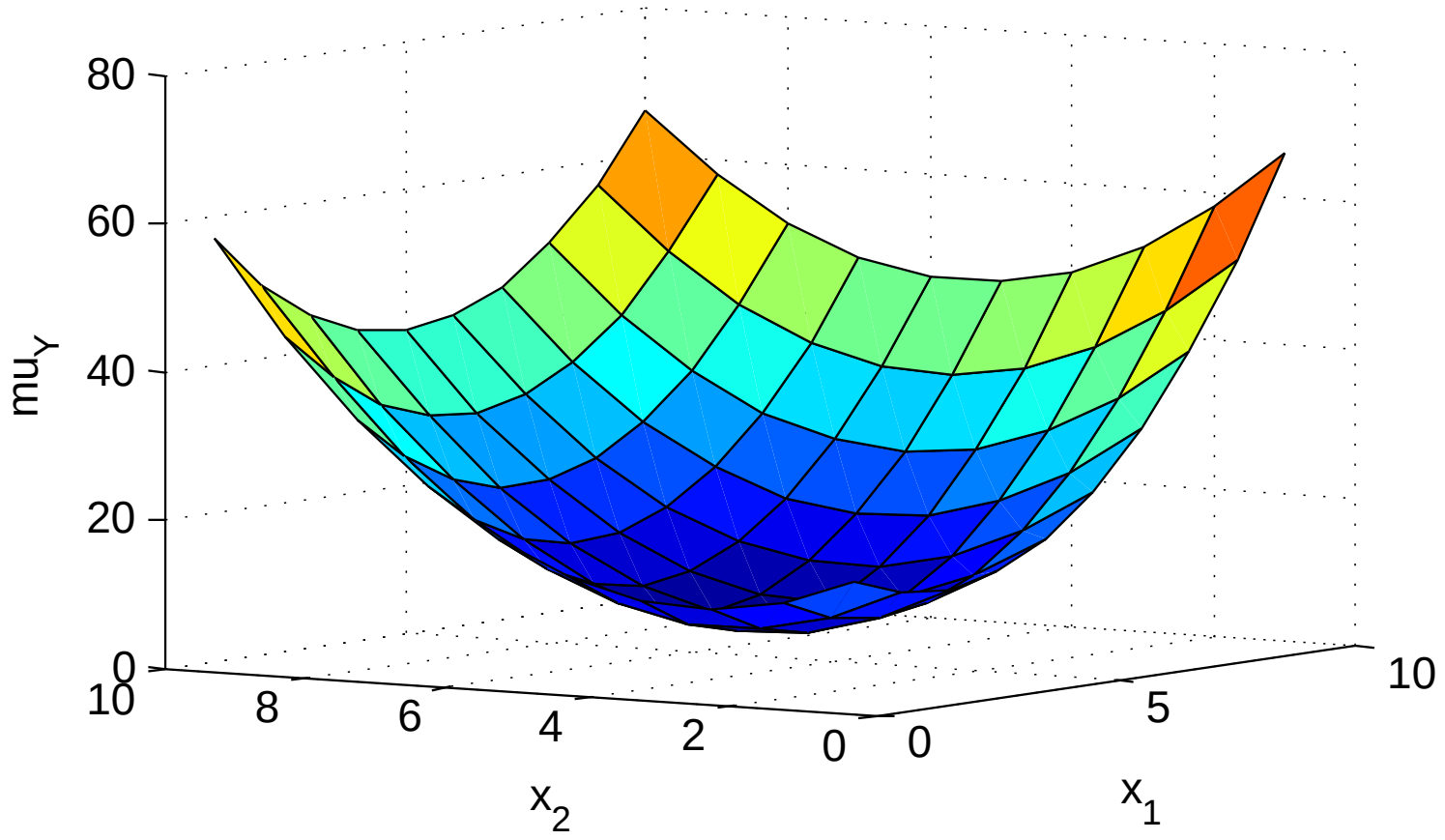
$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \beta_{ij} x_i x_j + \varepsilon$$

Quadratic

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \beta_{ij} x_i x_j + \sum_{k=1}^{n} \beta_k x_k^2 + \varepsilon$$

These forms are useful as they can express different forms of 'curvature' in the response.

Quadratic Model - 2 Factors



- Motivation
- Linear Models
- Experimental Designs
  - Experimental Designs
  - Factorial Design
  - Fractional Factorial Designs
- Model Fitting (Analysis)
- ANOVA
- Model Interpretation

Experimental Designs

Definition (Experimental Design)

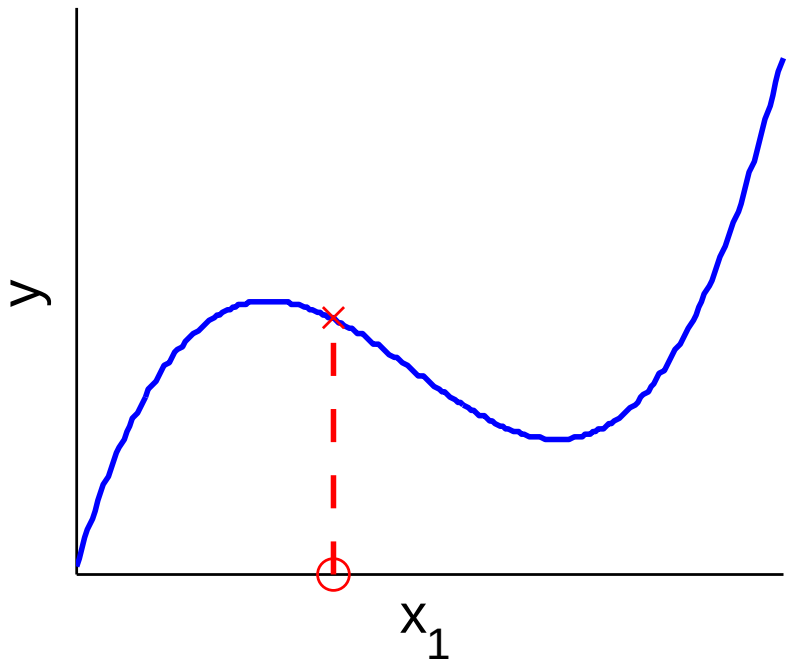
An experimental design is a set of factor settings (design points) for experimental trials.

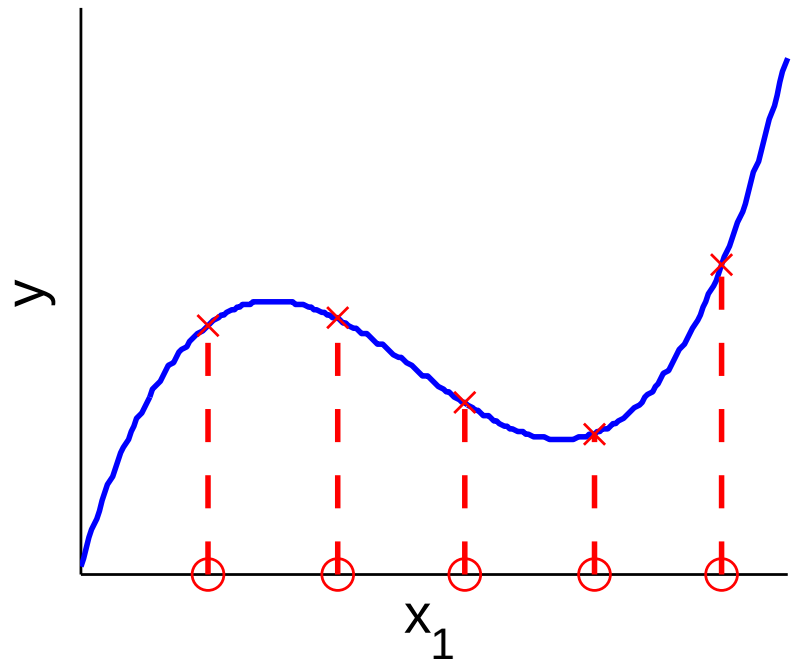
Generally, experimental designs are chosen that:

- enable the effect of each factor to be identified;
- require as few experimental trials as possible to achieve a desired level of accuracy.

Many designs are possible - the choice depends on the objective of the experiment.

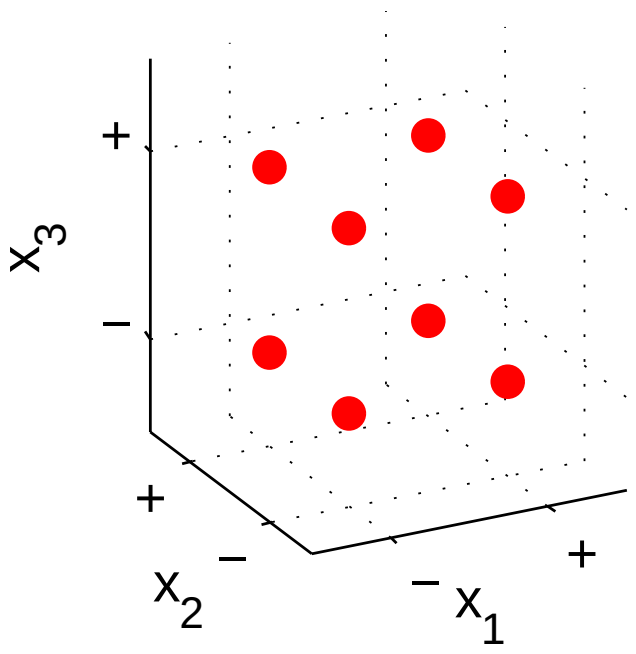
Experimental Designs





Factorial Design

Pick high (+) and low (-) values for each factor.



x_1	x_2	x_3
-	-	-
-	-	+
-	+	-
-	+	+
+	-	-
+	-	+
+	+	-
+	+	+

Factorial Design - Example

mutation rate values in range 0.05 to 0.2 appear to be good;

crossover rate values 0.4 to 0.8 give a good response;

population size found populations 100 to 260 give the best response.

mutation rate	crossover rate	population rate
0.07	0.45	120
0.07	0.45	230
0.07	0.75	120
0.07	0.75	230
0.18	0.45	120
0.18	0.45	230
0.18	0.75	120
0.18	0.75	230

Factorial Design in R

R

```
> library(AlgDesign)
> des <- gen.factorial(levels=c(2,2,3))
> des
   X1 X2 X3
1  -1 -1 -1
2   1 -1 -1
3  -1  1 -1
4   1  1 -1
5  -1 -1  0
6   1 -1  0
7  -1  1  0
8   1  1  0
9  -1 -1  1
10  1 -1  1
11 -1  1  1
12  1  1  1
```

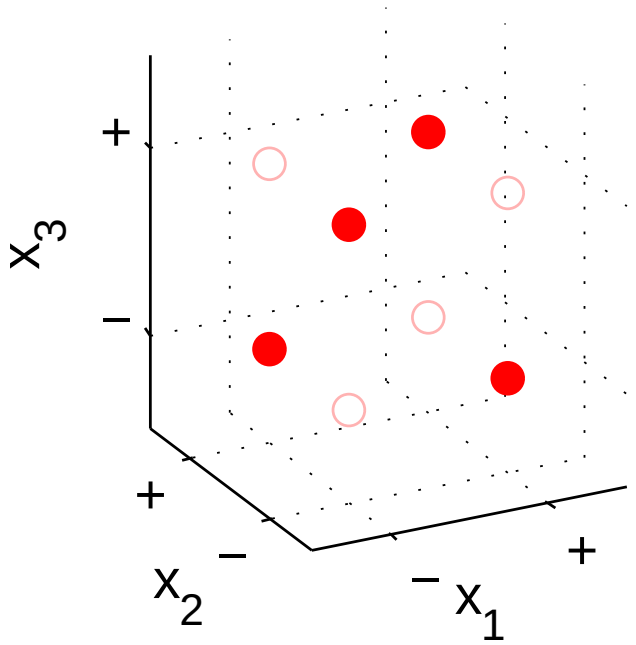
Fractional Factorial Designs

Factorial design of n factors: 2^n trials

Fractional factorial designs use special subsets of a full fractional design to reduce number of trials.

Advantage: fewer experiments
Disadvantage: some β parameters cannot be distinguished from one another in higher order models

Fractional Factorial Designs



x_1	x_2	x_3
-	-	+
-	+	-
+	-	-
+	+	+

Fractional Factorial Design in R

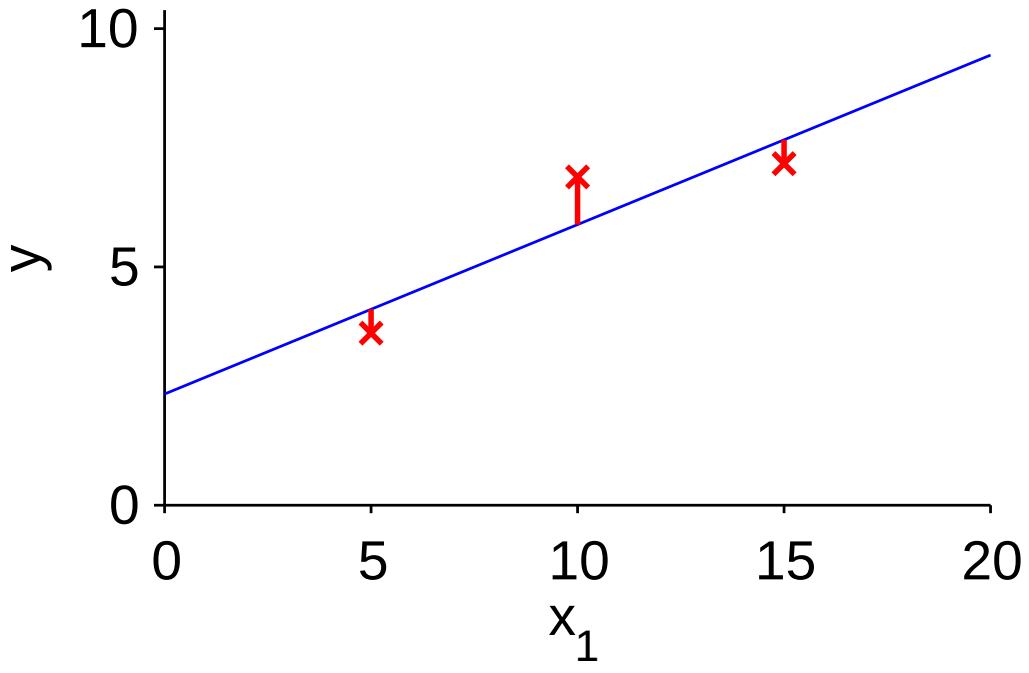
R

```
> library(FrF2)
> FrF2(nfactors=10, nruns=16)
    A  B  C  D  E  F  G  H  J  K
1   1  1  1  1  1  1  1  1  1  1
2  -1  1  1 -1 -1 -1  1  1 -1  1
3   1  1 -1  1  1 -1 -1  1 -1 -1
4   1 -1  1 -1 -1  1 -1 -1  1  1
5  -1 -1  1  1  1 -1 -1 -1 -1  1
6   1  1  1 -1  1  1  1 -1 -1 -1
7   1  1 -1 -1  1 -1 -1 -1  1  1
8  -1 -1 -1 -1  1  1  1  1 -1  1
9  -1 -1 -1  1  1  1  1 -1  1 -1
10  1 -1 -1 -1 -1 -1  1 -1 -1 -1
11  1 -1 -1  1 -1 -1  1  1  1  1
12 -1 -1  1 -1  1 -1 -1  1  1 -1
13 -1  1 -1  1 -1  1 -1 -1 -1  1
14 -1  1  1  1 -1 -1  1 -1  1 -1
15 -1  1 -1 -1 -1  1 -1  1  1 -1
16  1 -1  1  1 -1  1 -1  1 -1 -1
```

- Motivation
- Linear Models
- Experimental Designs
- Model Fitting (Analysis)
  - Least Squares Linear Regression
  - Residuals
- ANOVA
- Model Interpretation

Least Squares Linear Regression

Minimises square of distance from predicted and observed responses.



Returns estimate of parameters, $\hat{\beta}$, and variance, $\hat{\sigma}^2$, of noise term.

Least Squares Linear Regression - Example

Linear Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

x_1	x_2	x_3	y
60	12	3	107
60	12	6	114
60	18	3	86
60	18	6	72
90	12	3	163
90	12	6	173
90	18	3	138
90	18	6	143

Least Squares Linear Regression in R

R

```
> x1 = c(60,60,60,60,90,90,90,90)
> x2 = c(12,12,18,18,12,12,18,18)
> x3 = c(3,6,3,6,3,6,3,6)
> y = c(107,114,86,72,163,173,138,143)
> lm.ex = lm(y~x1+x2+x3)
> lm.ex

Call:
lm.default(formula = y ~ x1 + x2 + x3)

Coefficients:
(Intercept)           x1           x2           x3
    46.5000       1.9833      -4.9167       0.6667
```

Fitted Linear Model

$$y = 46.5 + 1.98x_1 - 4.92x_2 + 0.67x_3$$

Residuals

$\hat{y}$ is the predicted response for particular setting of the factors.

The residual is difference between observed and predicted response:

$$\hat{\varepsilon} = y - \hat{y}$$

Residuals are instances of the random variable ε constituting the noise term.

Residuals

Fitted Linear Model

$$y = 46.5 + 1.98x_1 - 4.92x_2 + 0.67x_3$$

Example

x_1	x_2	x_3	y	$\hat{y}$	$\hat{\varepsilon}$
60	12	3	107	108.5	-1.5
60	12	6	114	110.5	3.5
60	18	3	86	79.0	7.0
60	18	6	72	81.0	-9.0
90	12	3	163	168.0	-5.0
90	12	6	173	170.0	3.0
90	18	3	138	138.5	-0.5
90	18	6	143	140.5	2.5

- Motivation
- Linear Models
- Experimental Designs
- Model Fitting (Analysis)
- ANOVA
- Model Interpretation

ANOVA (Analysis of Variance)

Used to determine which factors influence the response.

For each factor in linear model, gives p-value for the hypothesis test:

Hypotheses

H_0: different levels of factor x_i have no effect on distribution of response
H_1: different levels of factor x_i do have an effect

In terms of the linear model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Hypotheses are similar to: H_0: $\beta_i = 0$ $\quad$ H_1: $\beta_i \neq 0$

ANOVA - Example

Linear Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

x_1	x_2	x_3	y
60	12	3	107
60	12	6	114
60	18	3	86
60	18	6	72
90	12	3	163
90	12	6	173
90	18	3	138
90	18	6	143

ANOVA in R

R

```
> x1 = c(60,60,60,60,90,90,90,90)
> x2 = c(12,12,18,18,12,12,18,18)
> x3 = c(3,6,3,6,3,6,3,6)
> y = c(107,114,86,72,163,173,138,143)
> aov.ex = aov(y~x1+x2+x3)
> summary(aov.ex)
            Df Sum Sq Mean Sq F value   Pr(>F)
x1           1   7080    7080 153.092 0.000245 ***
x2           1   1740    1740  37.632 0.003579 **
x3           1      8       8   0.173 0.698828
Residuals    4    185      46
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Kruskal-Wallis in R

R

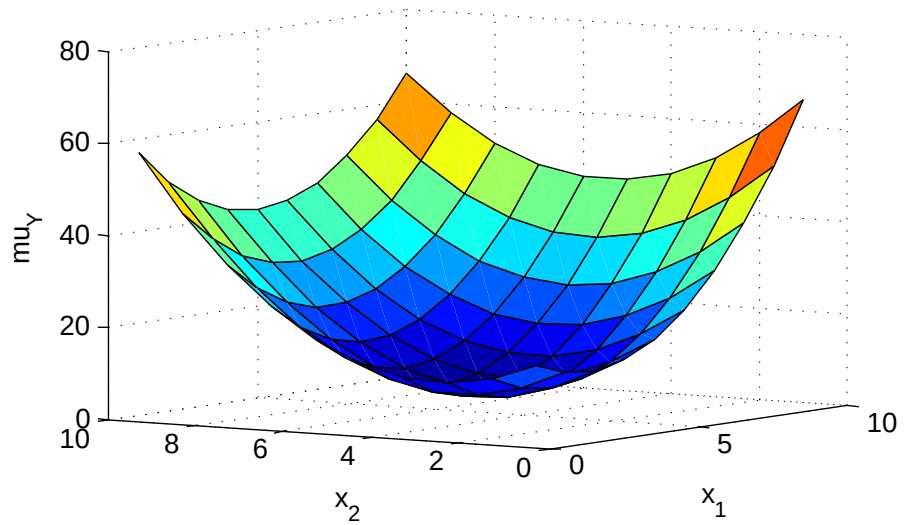
```
> x = c(1,1,2,2,3,3,4,4)
> y = c(107,114,86,72,163,173,138,143)
> kruskal.test(y,x)

Kruskal-Wallis rank sum test

data:  y and x
Kruskal-Wallis chi-squared = 6.6667, df = 3, p-value = 0.08332
```

- Motivation
- Linear Models
- Experimental Designs
- Model Fitting (Analysis)
- ANOVA
- Model Interpretation

Model Interpretation



Tuning Optimise resulting model (using calculus or deterministic optimisation methods) to find factors that give best response.

Scalability Express response in terms of scale factor, e.g.:

$$y = \beta_0 + \beta_1 x_1^2$$

where x_1 is the scale (problem characteristic factor).

Practical

Background

Research Question: How does the performance (the time required to achieve a given coverage) of ScandiTest depend on the 5 numeric parameters of Reindeer Herd Search?
Method: Created a two-level factorial design, and measured the performance at each design point.

1. Download the dataset cow38G.dat and import into R.
2. Use linear regression to fit a first-order linear model to the data.
3. Use ANOVA to identify which of the parameters has an effect of the performance.

Key Points

- Linear models - simple (but very widely used) models.
- Experimental designs - factorial, fractional factorial.
- Model fitting - parameter estimation using linear regression.
- ANOVA and Kruskal-Wallis test - to identify significant factors.
- Model interpretation - scalability, algorithm tuning.

Part X

Selected Resources

Resources I

website NIST/SEMATECH e-Handbook of Engineering Statistics
`http://www.itl.nist.gov/div898/handbook/`

book R Abelson
Statistics as Principled Argument, 1995

paper D Johnson
A Theoretician's Guide to the Experimental Analysis of Algorithms
Proc. 5th and 6th DIMACS Implementation Challenges, 59:215–250, 2002

example paper D White and S Poulding
A Rigorous Evaluation of Crossover and Mutation in Genetic Programming,
Proc. 12th European Conference on Genetic Programming (EuroGP), 220-231, 2009

Resources II

example paper Simon Poulding, John A Clark, and Hélène Waeselynck
A Principled Evaluation of the Effect of Directed Mutation on Search-Based Statistical Testing,
Proc. 4th International Workshop on Search-Based Software Testing (SBST), 184–193, 2011