Experimental Methods Workshop

Simon Poulding

University of York

7 September 2007

Part I

Introduction

Purpose

- To share knowledge and experience of reliable, efficient and principled experimentation.
- Focus: experiments on stochastic search algorithms in general, SBSE in particular.
- Emphasis on concepts, terminology and methods rather than theory and tools,
- . . . with suggestions for further reading.

Format

- Two sessions:
  - morning – 2 hours
  - afternoon – 3 hours
- Plenty of (short) breaks.
- Exercises - as a group.
- Ask questions at any time.
- Will distribute slides after workshop.

Exercise

The average acceleration experienced by astronauts in our new rocket to Mars (codename 'YORKIE') must not exceed 10g as it takes off.

$$\frac{\text{Acceleration (g)}}{12}$$

Exercise

The average acceleration experienced by astronauts in our new rocket to Mars (codename 'YORKIE') must not exceed 10g as it takes off.

Acceleration (g)
12
17
14

Exercise

The average acceleration experienced by astronauts in our new rocket to Mars (codename 'YORKIE') must not exceed 10g as it takes off.

Acceleration (g)
12
17
14
9

Exercise

The average acceleration experienced by astronauts in our new rocket to Mars (codename 'YORKIE') must not exceed 10g as it takes off.

Acceleration (g)
12
17
14
9
15
14
12

Exercise

The average acceleration experienced by astronauts in our new rocket to Mars (codename 'YORKIE') must not exceed 10g as it takes off.

Acceleration (g)
12
17
14
9
15
14
12

Exercise

The average acceleration experienced by astronauts in our new rocket to Mars (codename 'YORKIE') must not exceed 10g as it takes off.

Acceleration (g)
9

Exercise

The existing best algorithm takes 10 seconds to find a solution. Is my new algorithm (called 'YORKIE'), on average, any faster?

Time (s)
12
17
14
9
15
14
12

Exercise

The existing best algorithm takes 10 seconds to find a solution. Is my new algorithm (called 'YORKIE'), on average, any faster?

Time (s)
12
17
14
9
15
14
12

Exercise

The existing best algorithm takes 10 seconds to find a solution. Is my new algorithm (called 'YORKIE'), on average, any faster?

Time (s)
9

Computer Science Experimentation

Computer scientists often do not perform reliable experiments (or fail to communicate that they do).

Other sciences are more rigorous in their experimentation.

Objectives

1. Understand experimental techniques and statistical methods that enable experiments to be:

   reliable - significant experimental results are obtained using proven techniques;

   efficient - with both experimenter's time and computing resources.

2. Discuss how to perform experiments and communicate results.

Course Structure

1. Introduction
2. Concepts
3. Comparison Experiments
4. Algorithm Tuning and Factor Investigations
5. Experimental Method
6. Resources

Part II

Concepts

Why Experiment?

Why Experiment?

Experiments provide the information that enables science to move forward.

- Experiments enable insight and understanding.
- Hypotheses might be made based on the results
- . . . and experiments can verify or refute an hypothesis.

Why Use Statistics?

Experimental data and experimental hypotheses are usually expressed quantitatively.

Statistics provide a set of mathematical tools for formulating and testing experimental hypotheses.

- Statistical Concepts
- Experimental Model
- Sampling
- Confidence Intervals

- Statistical Concepts
  - Random Variables
  - Probability Mass and Density Functions
  - Expectation
  - Population Mean and Variance
  - Probability Distributions
- Experimental Model
- Sampling
- Confidence Intervals

Random Variables

Definition

Random variables are quantities where each value has an associated probability.

Examples

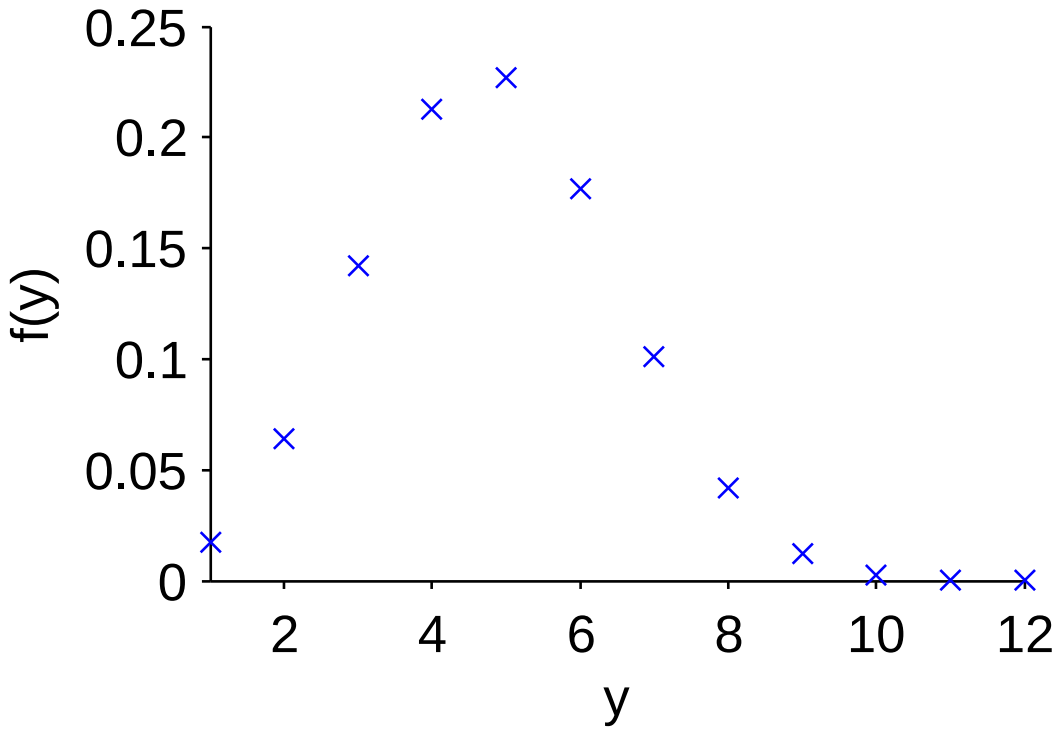
X = result of a single throw of a die
Y = a person's age
Z = performance of my YORKIE algorithm

Y denotes a random variable; y is a realised value of Y.

Probability Mass Function

- Y takes discrete values - a discrete distribution.
- Probability of random variable Y having the value y is denoted $f(y)$:

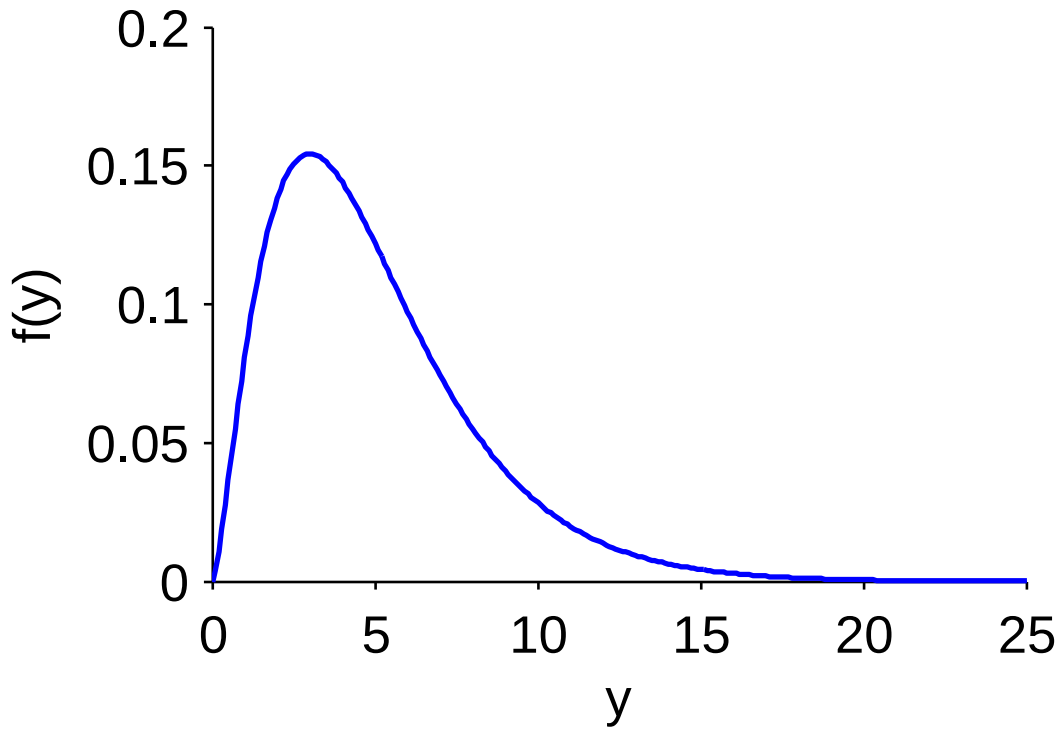
$$\mathbb{P}(Y = y) = f(y)$$



Probability Density Function

- Y is continuous - a continuous distribution.
- Probability of random variable Y having a value in a small interval of length δy around y is $f(y)\,\delta y$:

$$\mathbb{P}(y \leq Y < y + \delta y) = f(y)\,\delta y$$



Expectation

Definition

Expectation is the 'average' value of a quantity over a very large sample.

If Y has a probability distribution $f(y)$, then:

$$\mathbb{E}(Y) = \sum_y y\, f(y)$$

$$\mathbb{E}(Y) = \int_y y\, f(y)\, \mathrm{d}y$$

Exercise

Y is a discrete random variable with the following distribution:

$$f(1) = \mathbb{P}(Y = 1) = 3/11$$
$$f(2) = \mathbb{P}(Y = 2) = 4/11$$
$$f(3) = \mathbb{P}(Y = 3) = 2/11$$
$$f(4) = \mathbb{P}(Y = 4) = 2/11$$

What is the expected value of Y?

Exercise

Y is a discrete random variable with the following distribution:

$$
\begin{aligned}
f(1) &= \mathbb{P}(Y = 1) = 3/11 \\
f(2) &= \mathbb{P}(Y = 2) = 4/11 \\
f(3) &= \mathbb{P}(Y = 3) = 2/11 \\
f(4) &= \mathbb{P}(Y = 4) = 2/11
\end{aligned}
$$

What is the expected value of Y?

$$
\begin{aligned}
\mathbb{E}(Y) &= \sum_y y\, f(y) \\
&= 1 \times \tfrac{3}{11} + 2 \times \tfrac{4}{11} + 3 \times \tfrac{2}{11} + 4 \times \tfrac{2}{11} \\
&= 2\tfrac{3}{11}
\end{aligned}
$$

Population Mean

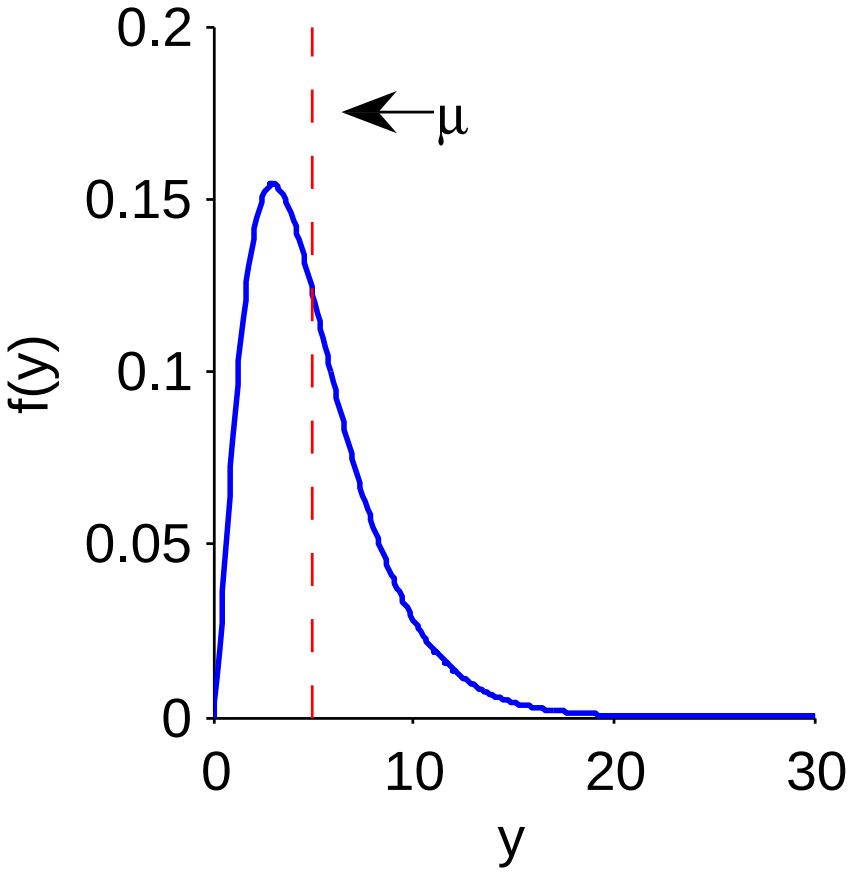
Definition

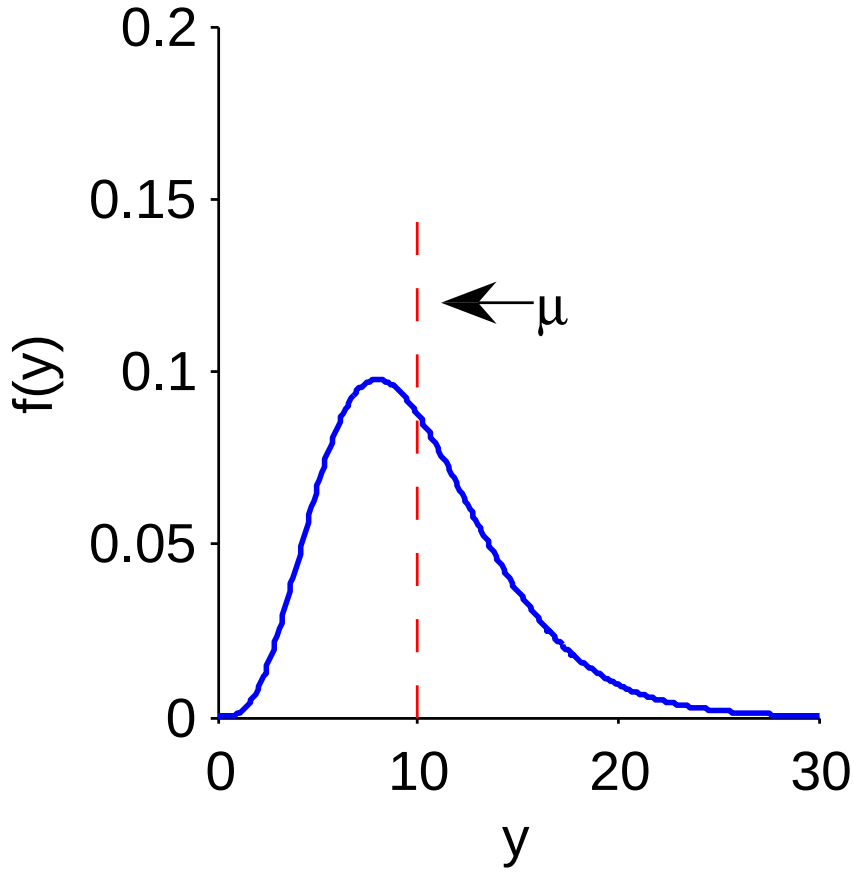
Population mean is the expected value of the random variable with a given probability distribution:

$$\text{mean}(Y) = \mu = \mathbb{E}(Y)$$

A measure of the 'central tendency' of a distribution.

Population Mean





Population Variance

Definition

Population variance is defined as:

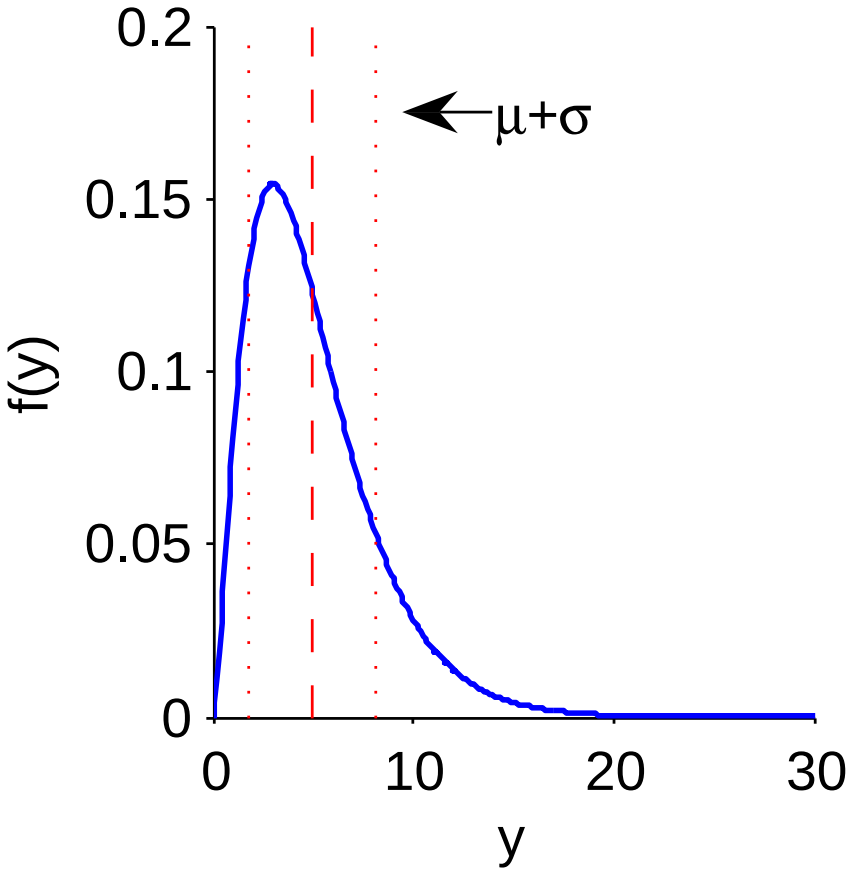
$$\text{var}(Y) = \sigma^2 = \mathbb{E}\left(\{Y - \mu\}^2\right)$$

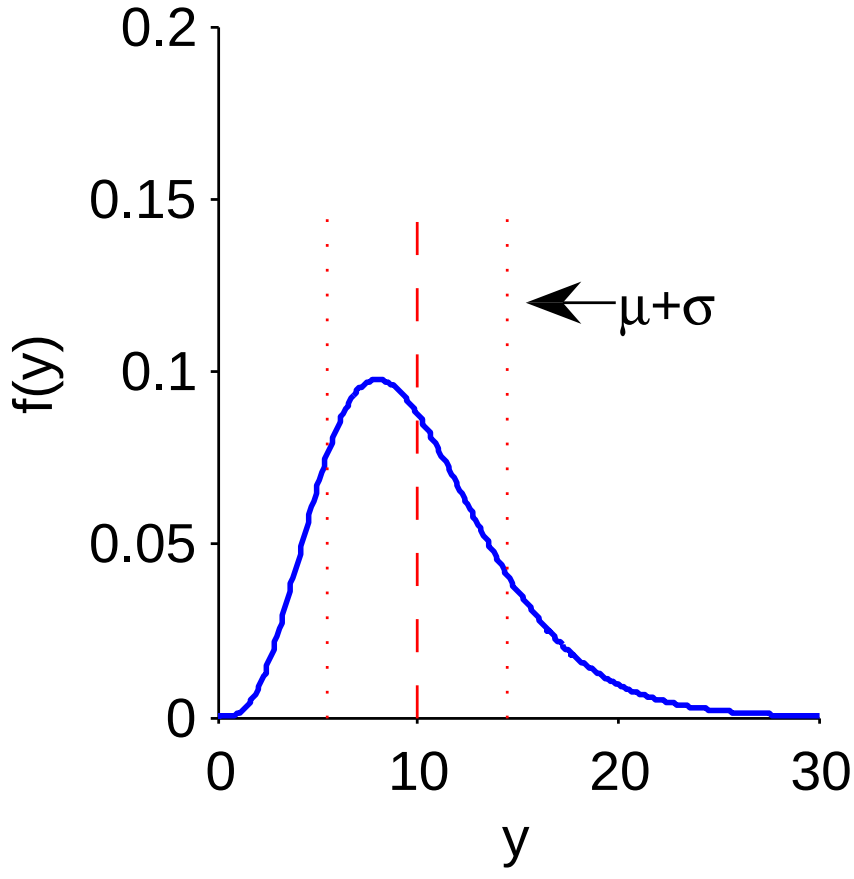
A measure of how much a distribution disperses (spreads) from its mean value.

Definition

Standard deviation, σ, is the square root of the variance.

Population Variance / Standard Deviation





Named Probability Distributions

- Large number of 'named' probability distributions (e.g. Binomial, Poisson, Uniform, Normal, Exponential, Gamma, Weibull, Chi-squared, ...).
- Useful because they are mathematically tractable or have specific desirable properties.

But ...

- A random variable can have any probability distribution, not necessary a named one.
- Often a named distribution is assumed because:
  - there are theoretical reasons which justify the assumption;
  - it is a simplifying approximation.

Distribution Parameters

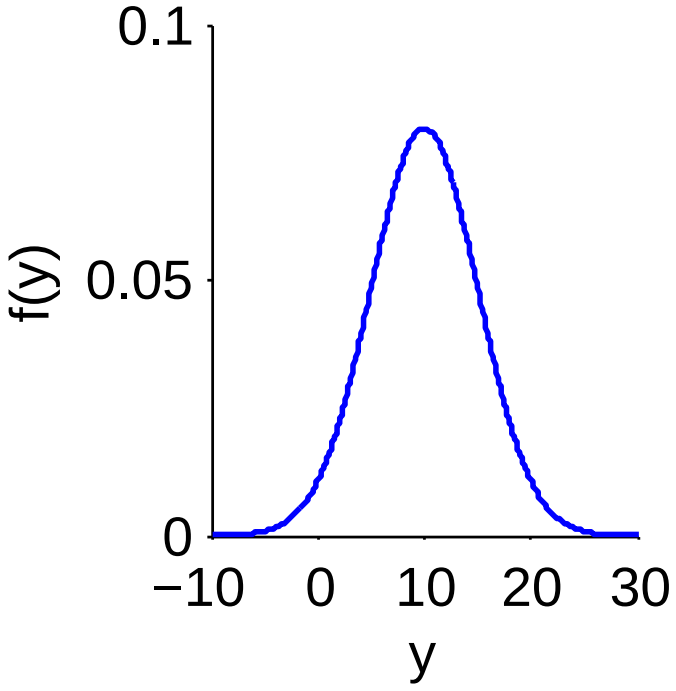
Definition

Many named probability distribution have distribution parameters which control the shape of the distribution.

In this case, the named probability distribution is really a family of related distributions.

Properties of the distribution, such as the population mean and variance, depend on the distribution parameters.

Normal Distribution

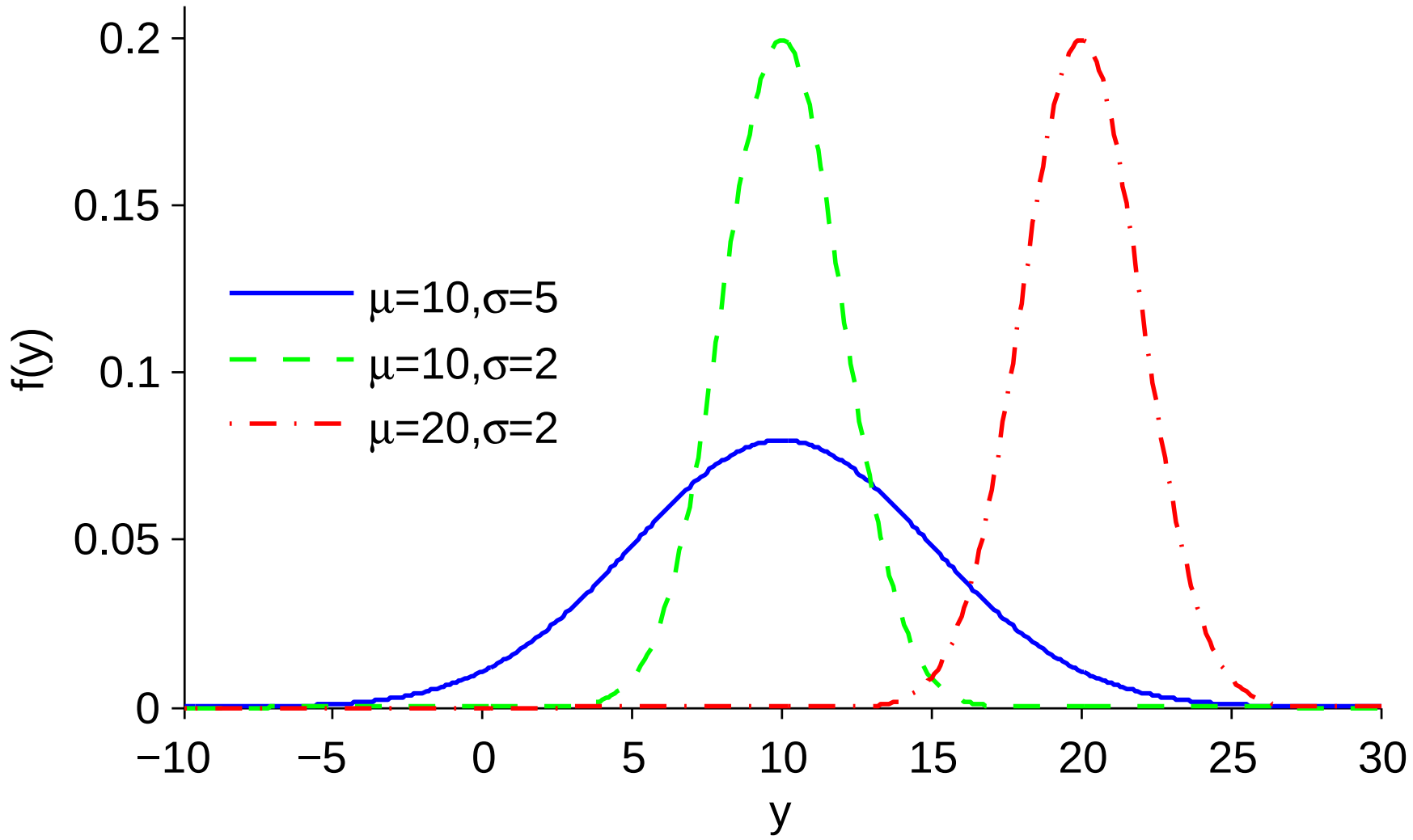


a continuous distribution

parameters: mean μ, variance σ^2

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

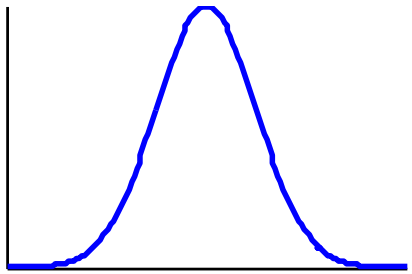
Normal Distribution - Parameters

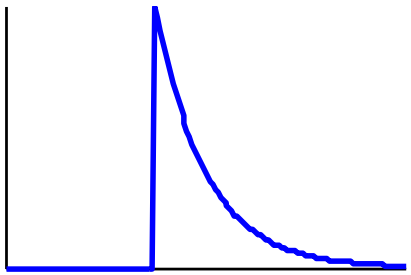


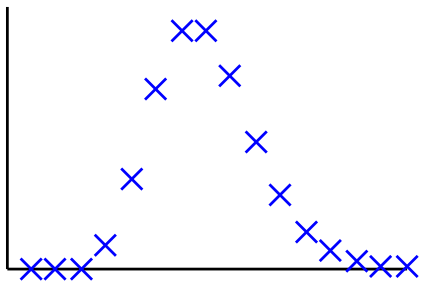
Other Distributions

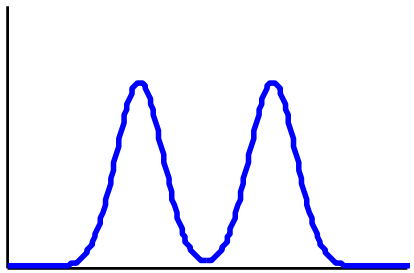
Sometimes good reason to assume Normal distribution. But can be an over-simplification, or simply incorrect.

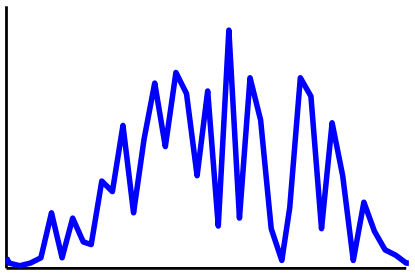
See resources for information about other distributions.

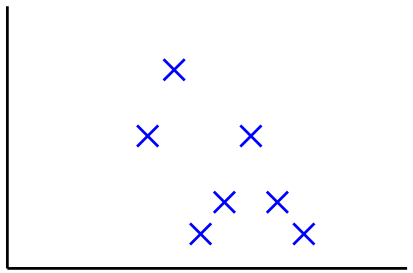












- Statistical Concepts
- Experimental Model
  - Factors and Response
  - Mathematical Model
  - Experimental Hypotheses
- Sampling
- Confidence Intervals

Experimental Model

- A simple, general model of an experiment
- Applicable to all the types of experiments considered in this workshop
- But don't claim that all experiments fit this model
- Will enable us to link statistics to experiments

Response

Definition

The response is the quantity that we measure in an experiment - the 'output' of the experiment.

Examples

Response

Definition

The response is the quantity that we measure in an experiment - the 'output' of the experiment.

Examples

- How long an algorithm takes to find a solution.
- Percentage of runs converging to the global optimum.
- Fitness after a set number of generations.
- Diversity of the solutions on a Pareto front.
- Number of function evaluations made by the algorithm.

Multivariate Response

Definition

A multivariate response consists of more than one quantity.

In this workshop, we consider univariate responses.

Factors

Definition

A factor is any 'input' to the algorithm that might affect the response (output) of the algorithm

Examples

Factors

Definition

A factor is any 'input' to the algorithm that might affect the response (output) of the algorithm

Examples

- Algorithm parameters (e.g. mutation rate, crossover rate, cooling rate, population size).
- Speed of the computer.
- 'Difficulty' of the problem.

Factor Types

algorithmic algorithm parameters
choice of cost function

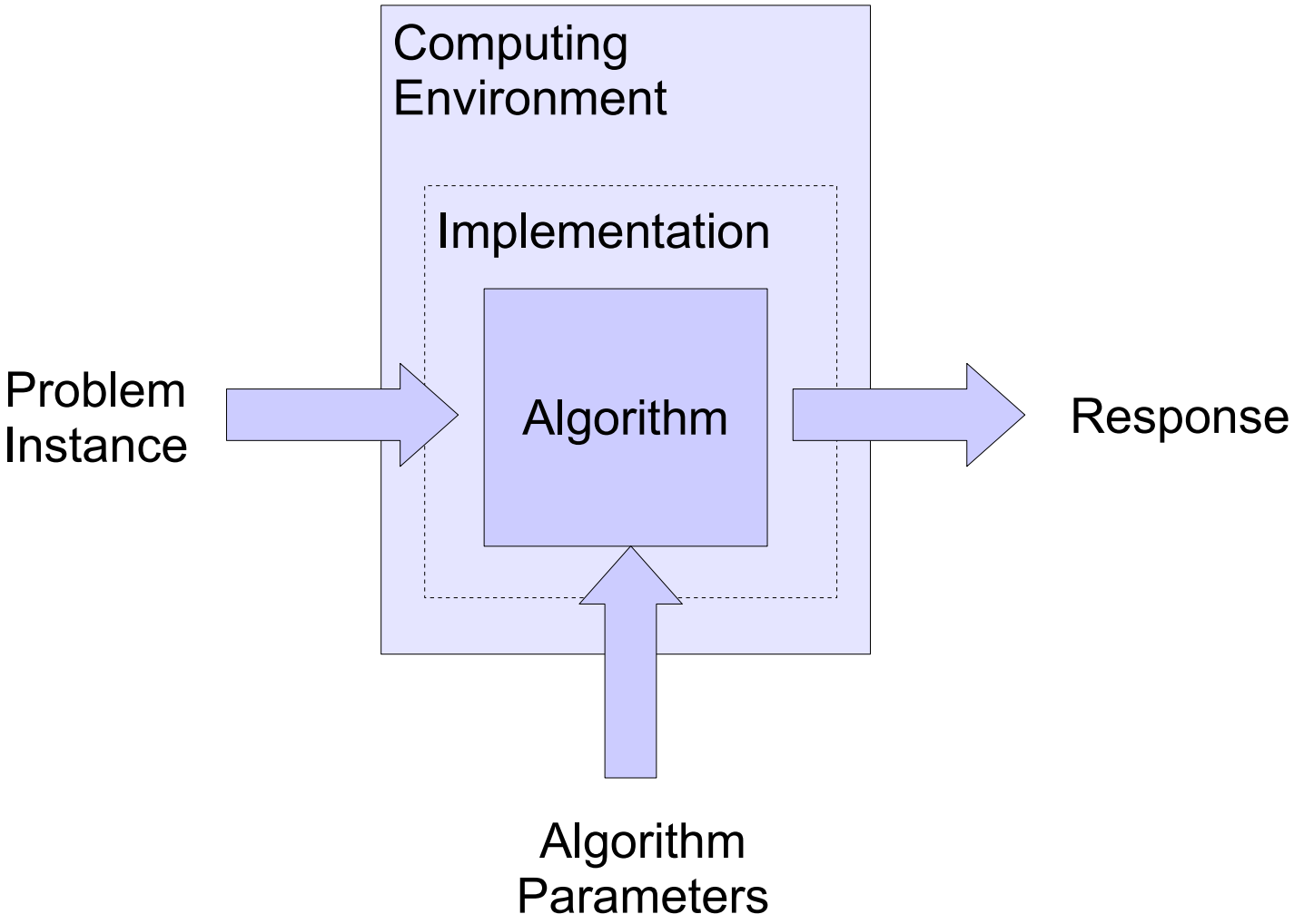
problem characteristics domain-specific properties of the problem that affect its 'difficulty', e.g.:

- size ('scale');
- symmetry (TSP).

environmental factors relating to the computing environment, e.g.:

- computing speed;
- efficiency of algorithm implementation;
- load from other processes/applications.

Factor Types



Random Numbers

Multiple runs of the same stochastic algorithm, on the same problem, using the same computer, produce different responses.

Why?

Different sequence of random numbers are generated each time the algorithm is run

. . . resulting in different initialisation, different choices for crossover, mutation, moves etc. in the algorithm

Therefore, could consider the random number sequence as a factor. Alternatively consider it as causing variance in the response.

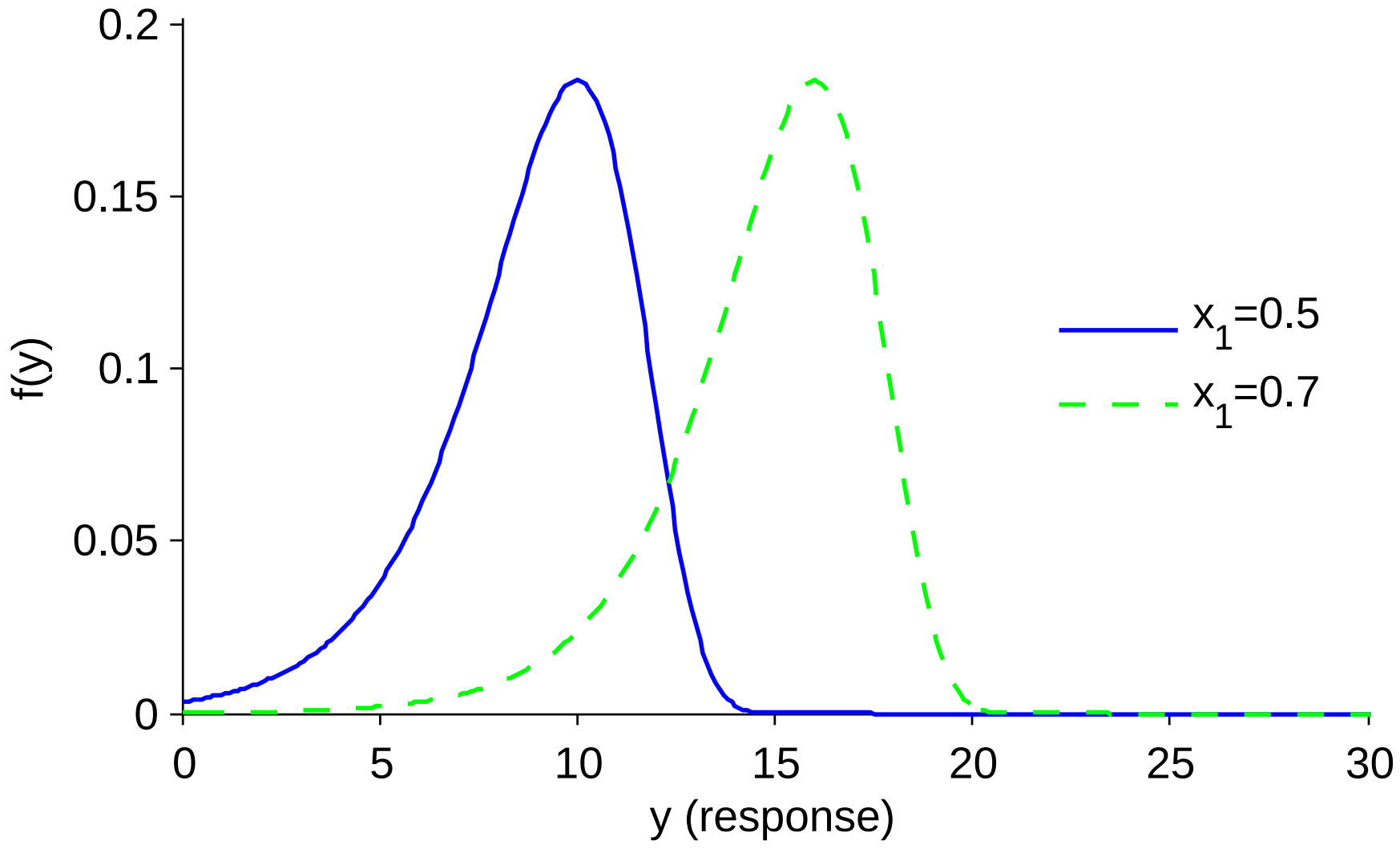
Mathematical Model

Simple Model

$$Y = F(\mathbf{x})$$

- Y is the response.
- $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ are the factors.
- $F(\cdot)$ has a probability distribution whose form is affected by the values of $\mathbf{x}$.
- In particular, mean response (μ) and variance (σ^2) depend on values of $\mathbf{x}$.

Mathematical Model



Experimental Hypotheses

Example (Comparison Experiment)

Hypothesis: A is 'better' than B

Algorithm A Model: $Y_A = F_A(\mathbf{x})$
Algorithm B Model: $Y_B = F_B(\mathbf{x})$

Hypothesis in terms of model: $\mu_A > \mu_B$

Experimental Hypotheses

Example (Scalability)

Hypothesis: response is $O(\log n)$
where n is the measure of scale (e.g. bits in representation)

Treat n as a problem characteristic factor to give:
Algorithm Model: $Y = F_A(n)$

Hypothesis in terms of model: $\mu = a + k \log n$

- Statistical Concepts
- Experimental Model
- Sampling
  - Experiments as Sampling
  - Statistics and Estimators
- Confidence Intervals

Experiments as Sampling

- Experiments provide information about the algorithm.
- . . . experiments provide information about the model $Y = F(\mathbf{x})$.
- To obtain this information we 'sample' the response of the algorithm.
- Statistics used to interpret the obtained information.

Exercise

Measure 10 responses of an algorithm:

Sample									
3.7	12.6	13.4	6.3	18.0	17.9	11.8	13.6	12.9	11.1

Estimate the value of μ for the probability distribution of model $Y = F(\mathbf{x})$.

Exercise

Measure 10 responses of an algorithm:

Sample									
3.7	12.6	13.4	6.3	18.0	17.9	11.8	13.6	12.9	11.1

Estimate the value of μ for the probability distribution of model $Y = F(\mathbf{x})$.

sum $= 121.3$
average value $= \frac{121.3}{10} = 12.13$

Statistics and Estimators

Definition

A statistic is a function applied to (a sample of) observed responses.

Definition

An estimator is a statistic used to estimate a population parameter.

Notation: For population parameter θ, an estimator is often denoted $\hat{\theta}$.

Estimators - Sample Mean and Variance

Definition (Sample Mean)

sample of n responses: $y_1, y_2, \ldots, y_n$

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

estimator of the population mean, μ

Definition (Sample Variance)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2$$

estimator of the population variance, σ^2

Robust Estimators - Median

Definition

If sample responses are ordered, median is the 'middle' value.

Robust measure of the 'central tendency' of the distribution.
It is not (necessarily) equal to the mean.

Example

3 8 4 6 1 5 8

median is ?

Robust Estimators - Median

Definition

If sample responses are ordered, median is the 'middle' value.

Robust measure of the 'central tendency' of the distribution.
It is not (necessarily) equal to the mean.

Example

1 3 4 5 6 8 8

median is 5

Robust Estimators - Median

Definition

If sample responses are ordered, median is the 'middle' value.

Robust measure of the 'central tendency' of the distribution.
It is not (necessarily) equal to the mean.

Example

1 3 4 5 6 8 23

median is 5

Robust Estimators - Interquartile Range

Definition

If sample responses are ordered, interquartile range is the difference between the values 25% and 75% through the list.

Robust measure of the 'dispersion' of the distribution.
It is not (necessarily) equal to the variance or standard deviation.

- Statistical Concepts
- Experimental Model
- Sampling
- Confidence Intervals

Exercise

What is the mean age of people in this room?

Estimators as Random Variables

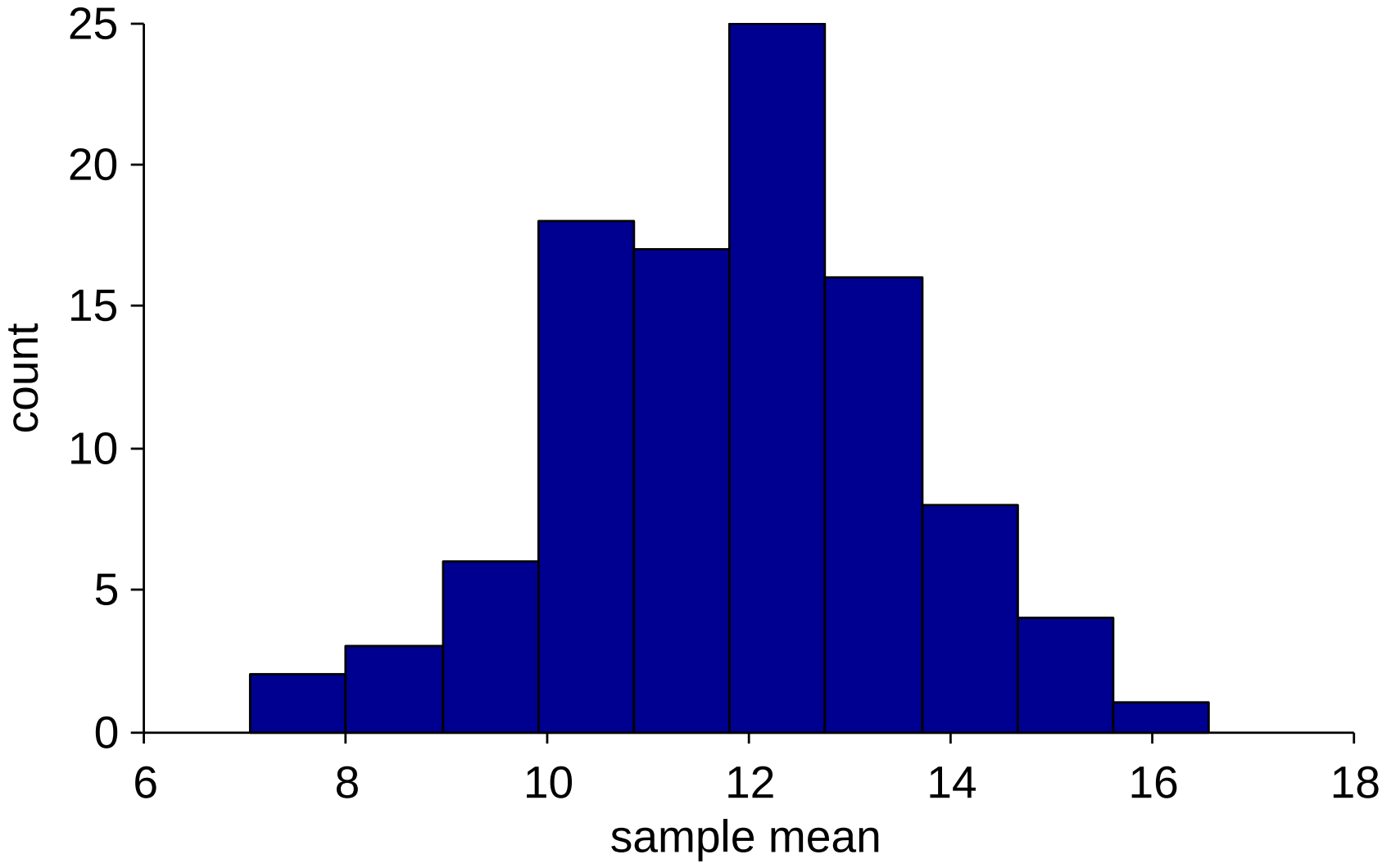
An estimator is also random variable - it will have a different value for each sample.

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
3	14.9 10.2 6.5 18.6 11.9 13.9 11.4 23.3 11.2 11.7	13.35
⋮	⋮	⋮

Estimators as Random Variables



Desirable Properties

Good estimators have:

- A mean (expected value) equal to the population parameter:

$$\text{mean}(\hat{\theta}) = \theta$$

  Such estimators are called unbiased.

- A low variance.

Example - Sample Mean

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

Theorem

For samples of n observations $(y_1, y_2, \ldots, y_n)$ *calculate sample mean* $\overline{Y}$.

$\overline{Y}$ *is a random variable, and:*

$$\mathbb{E}(\overline{Y}) = \mu_Y$$

$$\text{var}(\overline{Y}) = \frac{1}{n}\sigma_Y^2$$

So, variance of the sample mean (considered as a random variable) decreases as the sample size increases.

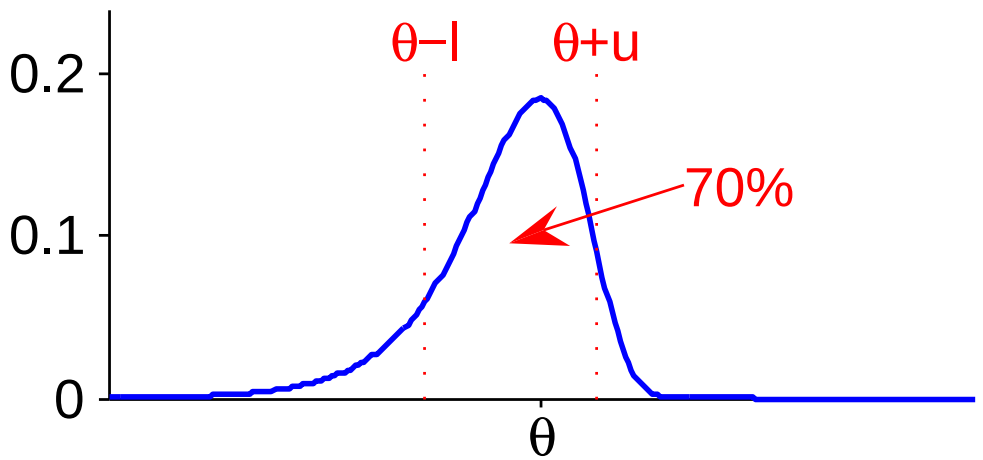
Confidence Intervals

- Since an estimator $\hat{\theta}$ is a random variable, the value calculated for a specific sample is not necessarily the 'true' value of the population parameter θ.
- If we take one sample and calculate the estimator, how close is the estimated value likely to be to the true value θ?
- One method of expressing this is a confidence interval.

Confidence Intervals - Principle

Assume:

1. that the estimator is unbiased, so the mean of the distribution is the same value θ;
2. we know enough about the distribution of the estimator to calculate an interval $[\theta - l, \theta + u]$ which contains (say) 70% of the distribution.



Confidence Intervals - Principle

We obtain sample of responses and calculate a single instance, t, of the estimator, $\hat{\theta}$.

70% of the time, t lies in the range, so that:

$$\theta - l \leq t \leq \theta + u$$

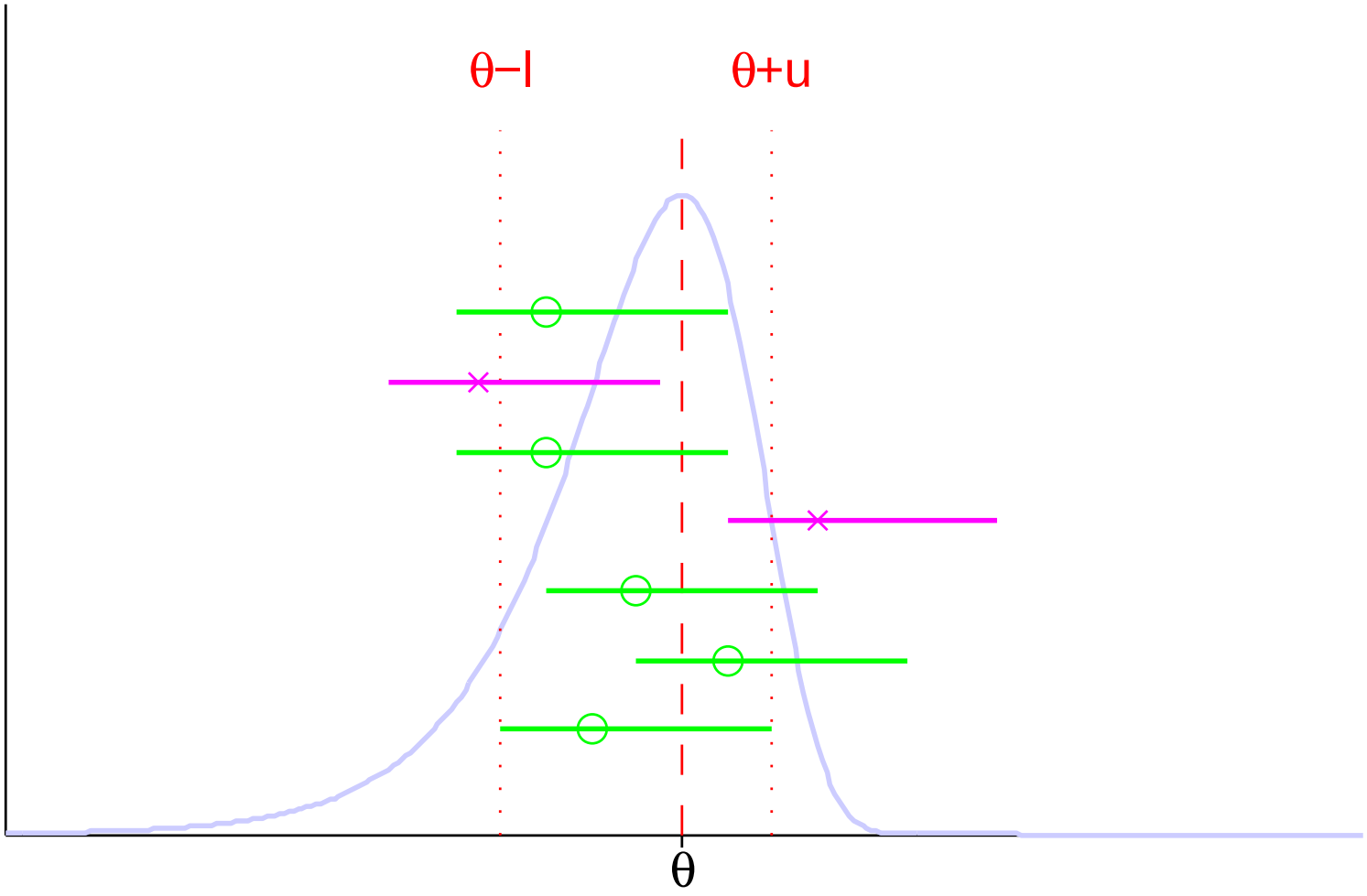
Inequality can be written as:

$$t - u \leq \theta \leq t + l$$

So, 70% of the time, the true value of the parameter, θ, lies in the range $[t - u, t + l]$ where t is a single estimate of the parameter.

$[t - u, t + l]$ is the 70% confidence interval of the estimate t.

Confidence Intervals - Principle



Confidence Intervals - Principle

Often interval is symmetrical, so expressed as $t \pm c$.

In many cases, c is taken to be (a multiple of) the standard deviation.

For an estimator with a Normal distribution, $\mu \pm \sigma$ is approximately 68% of the distribution.

On graphs, confidence intervals expressed as error bars.

Confidence Intervals - Example

- Assume Y has a normal distribution with mean μ_Y and variance $\sigma^2_Y = 25$
- Want to estimate μ_Y, and so use sample mean, $\overline{Y}$, as estimator
- $\overline{Y}$ is unbiased, so has mean equal to μ_Y and variance $\frac{1}{n}\sigma^2_Y$
- Can be shown that when Y has a Normal distribution, so does the sample mean $\overline{Y}$

Confidence Intervals - Example

Sample									
3.7	12.6	13.4	6.3	18.0	17.9	11.8	13.6	12.9	11.1

sample mean: $\overline{Y} = 12.13$

variance of $\overline{Y}$: $\frac{1}{n}\sigma_Y^2 = \frac{1}{10} \times 25 = 2.50$

so standard deviation of $\overline{Y}$: $\sqrt{2.50} = 1.58$

For a Normal distribution, 68% confidence interval is approximately 1 standard deviation either side of mean.

Therefore, 68% confidence interval for μ_Y is approximately:
12.13 ± 1.58

Importance of Giving Confidence Intervals

If instead of sample of 10, just had first 4 responses:

Sample			
3.7	12.6	13.4	6.3

estimate: 9.00 ± 2.50

If there were sample of 100:

estimate: 12.03 ± 0.50

Thus, the confidence interval indicates the accuracy of the parameter estimate.

Central Limit Theorem

Have used result:

Theorem

When Y has a Normal distribution, so does the sample mean $\overline{Y}$.

This is not generally true of other distributions.
But . . .

Theorem (Central Limit Theorem)

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

Summary

- Random variables have an associated probability distribution, with properties such as population mean and variance.
- The response is the algorithm output and it depends on input factors.
- For stochastic algorithms, the response is a random variable.
- General mathematical model of algorithm response:

$$Y = F(\mathbf{x})$$

  where the distribution of $F(\cdot)$ depends on the factors $\mathbf{x}$.
- Experiments provide sample responses, from which properties of $F(\cdot)$ can be estimated (e.g. the mean).
- Confidence intervals illustrate the accuracy of a parameter estimate.

Part III

Comparative Experiments

Comparative Experiments

Experiments often compare algorithms:

- two different algorithms;
- the same algorithm with two different sets of parameter settings;
- an algorithm against a known benchmark.

This part describes simple comparisons using hypothesis testing techniques.

- Hypothesis Testing
- Parametric Tests
- Non-Parametric Tests

- Hypothesis Testing
  - Hypothesis Testing Method
  - Significance
  - Hypothesis Testing Interpretation
- Parametric Tests
- Non-Parametric Tests

Hypothesis Testing

Definition (Null Hypothesis)

Usually the current state of knowledge - retained unless there is good evidence to the contrary.

Definition (Alternative Hypothesis)

The hypothesis put forward as new knowledge.

The objective of hypothesis testing is to determine whether there is sufficient evidence to show that the alternative hypothesis is true.

Hypothesis Testing Example

Hypothesis: The existing best algorithm takes, on average, 10 seconds to find a solution. Is my new 'YORKIE2' algorithm faster?

Null Hypothesis

H_0: mean performance of YORKIE2 $= 10$ seconds (no difference)

Alternative Hypothesis

H_1: mean performance of YORKIE2 < 10 seconds (YORKIE2 is faster)

If YORKIE2 is faster: H_1 true (and so H_0 false)

Hypothesis Testing Example

Restate hypotheses in terms of experimental model:

Hypotheses

H_0: $\mu = 10$ $\quad\quad$ H_1: $\mu < 10$

Decide to use sample mean, $\overline{Y}$ as estimator for μ.
$\overline{Y}$ is the test statistic.

For some reason, we know that Y has a Normal distribution with a variance of 4.

Hypothesis Testing Example

Take sample of 10 responses:

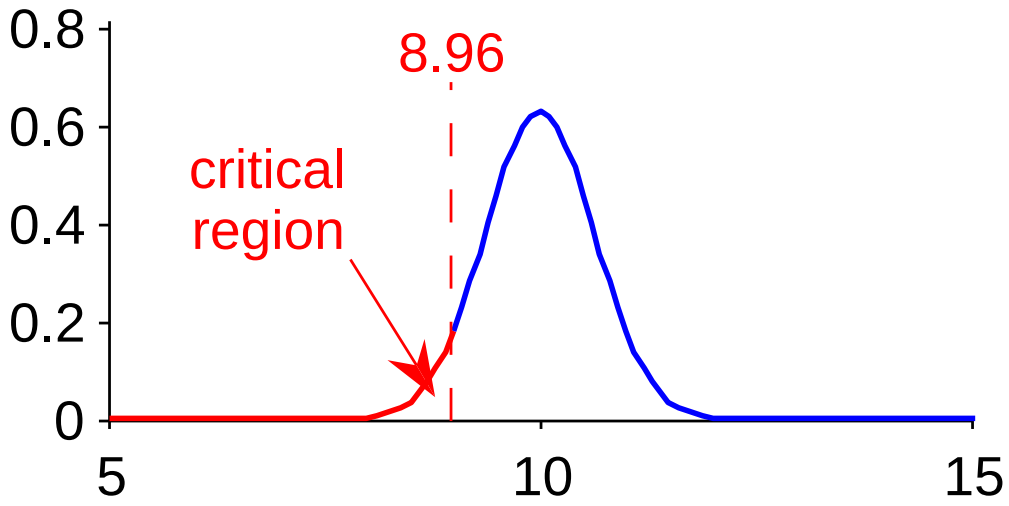
Sample									
9.6	6.3	10.4	12.2	7.6	10.7	11.5	5.8	6.1	10.1

To test hypothesis, we assume H_0 is true.

If H_0 true, $\overline{Y}$ has a Normal distribution with mean 10 and variance 0.4.

Hypothesis Testing Example

1. Determine a critical region of this distribution where observed values of $\overline{Y}$ would appear if H_1 were true.
   This is the region of lowest values of $\overline{Y}$.
2. Fix the size of the critical region so that it is (say) 5% of the distribution.
   For $\mathcal{N}(10, 0.4)$, it is the region $\overline{Y} < 8.96$.



Hypothesis Testing Example

Decision Rule

If test statistic falls in critical region: accept H_1
Otherwise: accept H_0

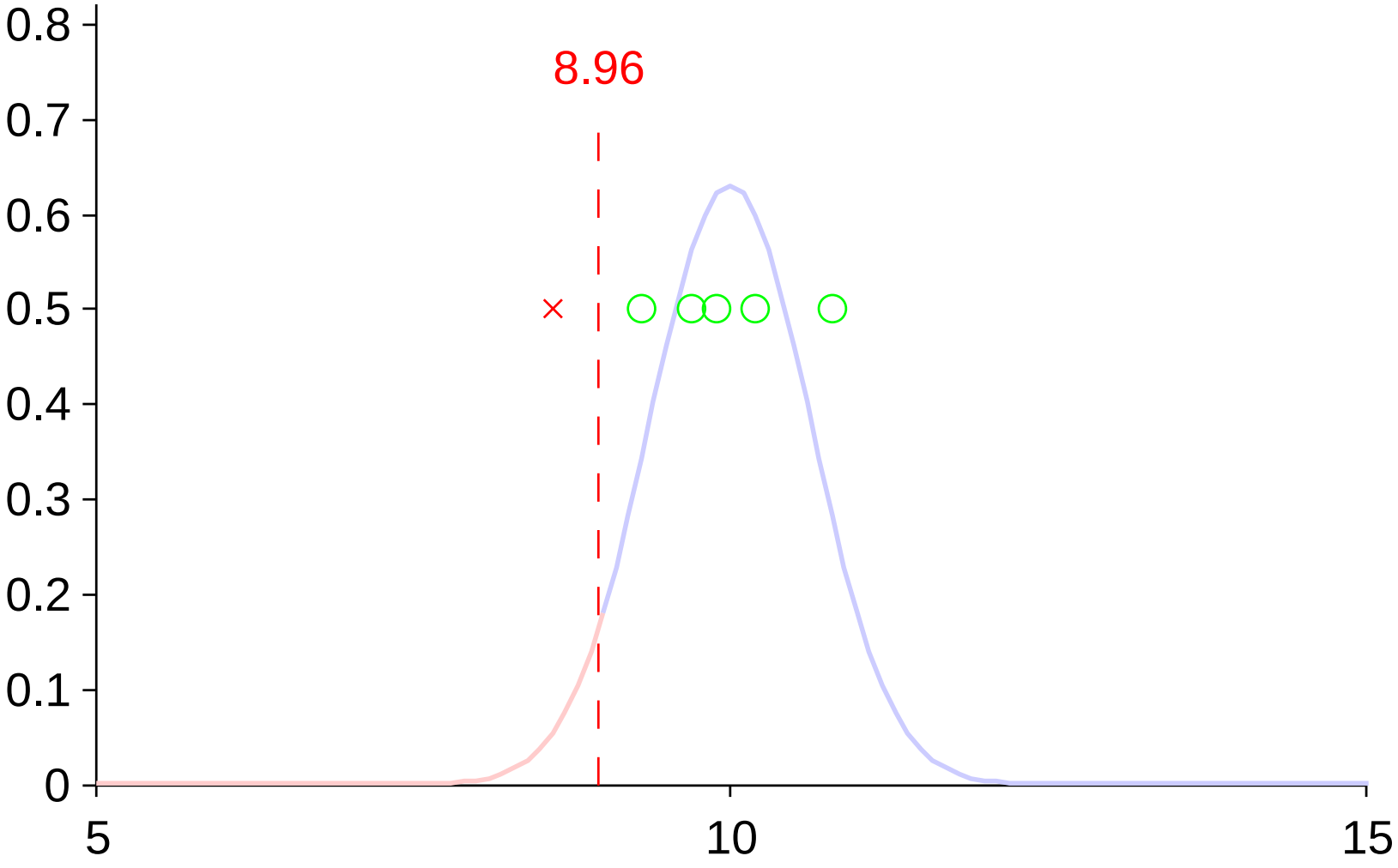
observed sample mean, $\overline{Y}$: 9.03
critical region: $\overline{Y} < 8.96$

Conclusion: H_0 true, i.e. YORKIE2 is no faster than existing algorithm.

Hypothesis Testing Summary

1. Assume H_0 to be true.
2. Based on this assumption, identify a region of the test statistic's distribution where values are most indicative of H_1 being true.
3. Take a sample and calculate the test statistic.
4. Accept H_1 (reject H_0) only if the value of test statistic falls in the critical region. Otherwise accept H_0.

Hypothesis Testing Summary



Type I and II Errors

If H_0 is true, test statistic could still fall in the critical region.

We chose 5% of the distribution to be in the critical region, so 5% chance of this occuring if H_0 is actually correct.

Definition (Type I Error)

Rejecting H_0 when it is actually true.

Definition (Type II Error)

Accepting H_0 when it is actually false.

Significance Level

Normally choose size of critical region (probability of Type I error) to be 5%, 10%, or 1%
This is the significance level of the test.

Note

The significance level is not the probability of the result of the hypothesis test being wrong.

It is the probability of incorrectly rejecting H_0 given that it is correct.

Good hypothesis tests have low probabilities of both Type I and II errors.

p-Value

Definition

p-value is the probability of obtaining the observed data if H_0 is true.

It is therefore the probability of making a Type I error if we reject H_0 based on this data.

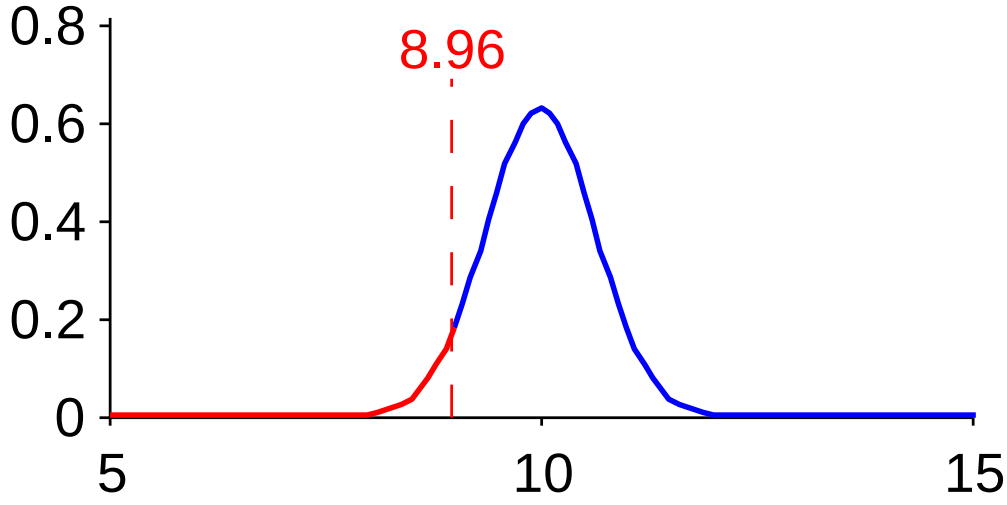
Thus, to accept H_1 at 5% significance, must have a p-value ≤ 0.05.

Analysis tools often report the p-value.
Best practice is to decide on significance level before performing experiment.

One-Tailed and Two-Tailed Tests

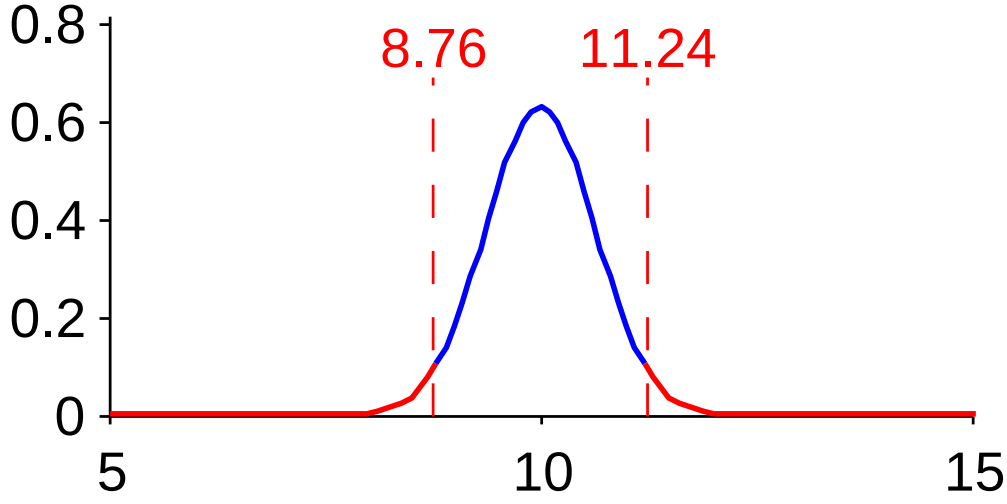
One-Tailed Test

H_0: $\mu = 10$
H_1: $\mu < 10$



Two-Tailed Test

H_0: $\mu = 10$
H_1: $\mu \neq 10$



Interpretation

When H_0 is accepted, doesn't necessarily mean H_1 is incorrect, only that there is insufficient evidence that show that it is correct.

In YORKIE2 example above, H_1 was actually correct, but insufficient evidence to show this.

Criticism of Hypothesis Testing

- Null hypothesis such as '$\mu = 10$' – almost always possible to reject given large enough sample.
- Why choose arbitrary 5% significance?
- Misunderstanding of the meaning of the significance level.

Exercise

I now have a third version of my algorithm, YORKIE3.
Ran one hypothesis test that accepts that performance of YORKIE3 is better than 8 seconds at the 5% significance level.

Hypotheses

H_0: YORKIE3 performance $= 8$ seconds
H_1: YORKIE3 performance < 8 seconds

If I obtained a new sample, what is the probability that the same significance test will again accept H_1?
(Assume Type II error probability is 30%.)

Exercise

I now have a third version of my algorithm, YORKIE3.
Ran one hypothesis test that accepts that performance of YORKIE3 is better than 8 seconds at the 5% significance level.

Hypotheses

H_0: YORKIE3 performance $= 8$ seconds
H_1: YORKIE3 performance < 8 seconds

If I obtained a new sample, what is the probability that the same significance test will again accept H_1?
(Assume Type II error probability is 30%.)

Depends on whether H_1 is really correct or not:

- If H_0 true ($= 8$ seconds): probability is 5% that the new sample will also accept H_1
- If H_1 true (< 8 seconds): probability is 70% that the new sample will also accept H_1

- Hypothesis Testing
- Parametric Tests
- Non-Parametric Tests

Parameteric Tests

- Tests for performing algorithm comparison.
- Use specific test statistics to test a hypothesis.
- Assume response has a particular (parameterised) probability distribution.

Z Test

In hypothesis testing example above, we somehow 'knew' that our test statistic $\overline{Y}$:

1. was Normally distributed;
2. had a specific population variance.

This information enabled us to apply hypothesis test using a critical region of a Normal distribution. This was a (effectively) a Z test.

Student's t Distribution

What if we don't know the variance of the sample mean, $\overline{Y}$? (But still know it is Normally distributed.) For large samples, could

use sample variance, S^2, as an estimate.

For small samples, more accurate to use a new test statistic:

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}}$$

to test:

Hypotheses

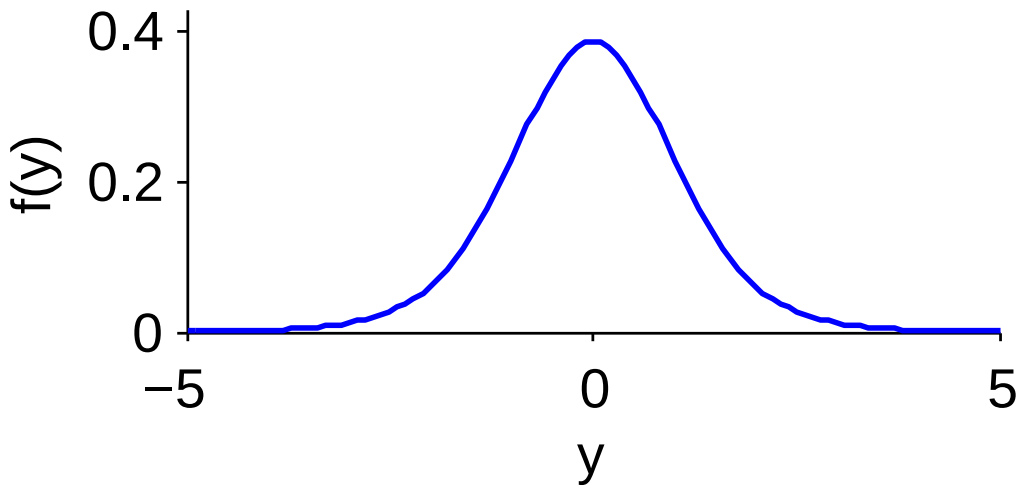
H_0: mean of Y is μ
H_1: mean of Y is not μ

Student's t Distribution

If,

1. Y has a Normal distribution
2. H_0 is true (mean of Y is μ)

then, T has Student's t distribution with $n-1$ 'degrees of freedom'



Two-Sample t Test

assumptions Y_A, Y_B both Normally distributed (with means μ_A, μ_B),
both have same variance,
sample size n for both

hypotheses H_0: $\mu_A = \mu_B$
H_1: $\mu_A \neq \mu_B$

statistic

$$T = \frac{\overline{Y_A} - \overline{Y_B}}{\sqrt{(S_A^2 + S_B^2)/n}}$$

distribution If H_0 true: T has Student's t distribution, $2n - 2$ degrees of freedom

Two-Sample t Test - Example

R

```
> a = c(9.3,5.0,6.5,4.7,8.2,8.0,3.0,7.9,5.8,7.0)
> b = c(7.1,5.9,7.2,7.8,13.0,8.4,6.7,8.7,7.6,9.7)
> t.test(a,b,paired=FALSE,var.equal=TRUE)

        Two Sample t-test

data:  a and b
t = -1.9038, df = 18, p-value = 0.07305
alternative hypothesis: true difference in means is not
 equal to 0
95 percent confidence interval:
 -3.5129263  0.1729263
sample estimates:
mean of x mean of y
     6.54      8.21
```

Assumptions

Parametric Test Assumptions

Important to verify assumptions when applying a parametric test.

- Hypothesis Testing
- Parametric Tests
- Non-Parametric Tests

Non-Parameteric Tests

- Tests for performing algorithm comparison.
- Use specific test statistics to test a hypothesis.
- No assumptions about the probability distribution of the response
- . . . but can be slightly less discriminating because of this.

Paired Samples

t test example assumed that samples for algorithms A and B were just taken randomly (unpaired).

Alternative method: pair the samples, so that each member of the pair taken under equivalent conditions.

For example: give a set of test problem instances for the algorithms, apply A and B to each instance in turn.

Problem	Y_A	Y_B
1	10.4	9.3
2	12.3	11.8
3	8.7	8.9
⋮	⋮	⋮

Paired Sample Tests

Paired sample versions of many parametric and non-parametric tests.

Often paired version calculates difference $D = Y_A - Y_B$ and applies unpaired test to:

Hypotheses

H_0: $\mu_D = 0$
H_1: $\mu_D \neq 0$

Sign Test

Applied to sample of a single distribution.

Hypotheses

H_0: median$(Y) = \eta$
H_1: median$(Y) > \eta$

Rank Sum Test (Mann-Whitney-Wilcoxon)

Applied to unpaired samples of two distributions.

Hypotheses

H_0: Y_A and Y_B have the same distribution
H_1: Y_A and Y_B have different distributions

Almost effective as Student's t test.

Rank Sum Test - Example

R

```
> a = c(9.3,5.0,6.5,4.7,8.2,8.0,3.0,7.9,5.8,7.0)
> b = c(7.1,5.9,7.2,7.8,13.0,8.4,6.7,8.7,7.6,9.7)
> wilcox.test(a,b,paired=FALSE)

        Wilcoxon rank sum test

data:  a and b
W = 29, p-value = 0.123
alternative hypothesis: true mu is not equal to 0
```

Signed Rank Test (Wilcoxon)

Applied to paired samples of two distributions.

Hypotheses

H_0: Y_A and Y_B have the same distribution
H_1: Y_A and Y_B have different distributions

Summary

- Hypothesis testing used to formalise algorithm comparison.
- Significance and p-values
- Examples of hypotheses testing using:
  - parametric tests;
  - non-parametric tests.

Principled Comparison

Have discussed statistical tests for comparing algorithms, but how do we make the comparison fair?

What if the algorithms are very different in nature, e.g. GA vs. SA? How should factors be set when they are not equivalent?

Suggestion

Tune the parameters of both algorithms in the 'same' principled manner before making a comparison.

Part IV

Algorithm Tuning and Factor Investigations

Introduction

Model

$$Y = F(\mathbf{x})$$

Now consider experiments where we want to understand how the response changes in terms of the factors, i.e. the form of $F(\cdot)$.

Examples

- Tuning - finding the algorithmic factors that give the best response.
- Understanding how the response depends on problem characteristics such as scale.

Reliable Tuning of Multiple Factors

Simple approach is one factor at a time:

1. pick a factor x_i;
2. find value of x_i that optimises response;
3. keep factor x_i at optimum, and repeat for other factors.

Not guaranteed to find global optimum.
This section considers a more reliable methodology.

- Linear Models
- Experimental Designs
- Model Fitting (Analysis)
- Nuisance Factors
- Problem Instances
- Model Interpretation

- Linear Models
  - Linear Model
  - Higher Order Linear Models
  - Generalised Linear Models
- Experimental Designs
- Model Fitting (Analysis)
- Nuisance Factors
- Problem Instances
- Model Interpretation

Experimental Model

Simple Model

$$Y = F(\mathbf{x})$$

$F(\mathbf{x})$ could be any function of $\mathbf{x}$.

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

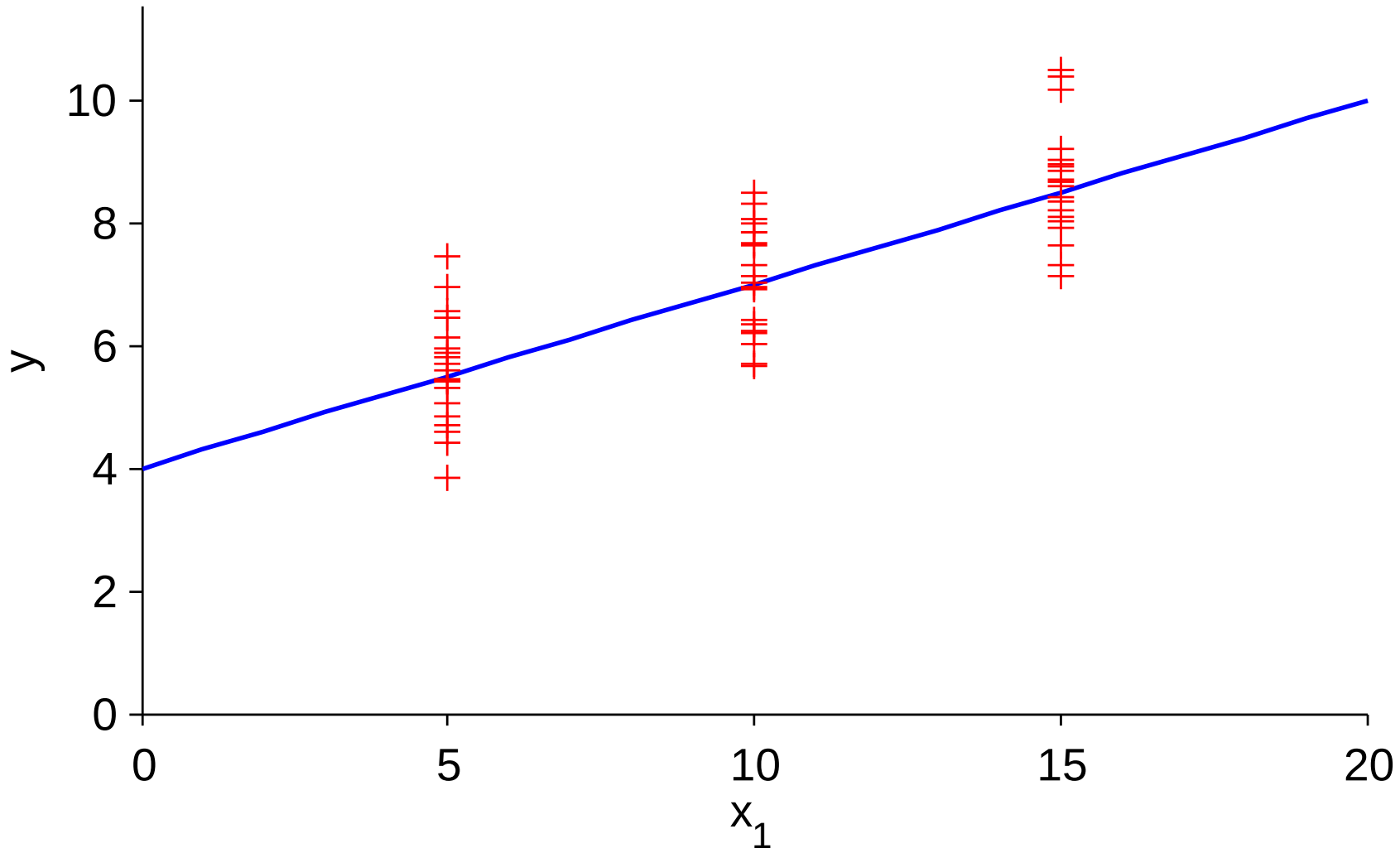
The linear terms ($\beta_0 + \sum_i \beta_i x_i$) explain the mean response.
The noise term (ε) explains the variance in the response.

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



Noise Term

For stochastic search algorithms, noise term accounts for variance in response owing to different random number sequences.

If model doesn't include all factors that affect the response, the noise term may also need to account for the effect of the extra factors.

Noise Term

Example

Response is wall-clock time that a GA takes to converge.

Factors may be mutation rate, crossover rate and population size - we include these in the linear model.

But if there are 5 similar PCs on which we run the GA, the difference in the performance of the PCs may affect the response.

Could include the PC performance as a factor in the model or use the noise term to accommodate the differences owing to PC performace.

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.

Generalised Linear Models

Often assumptions on distribution of noise term are better satisfied by applying a function to the response:

Generalised Linear Model

$$g(Y) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon$$

Example

For example, $g(\cdot)$ may be the logarithm of the response:

$$\log Y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon$$

Model Fitting

Objective is to find values of β_i (and possibly variance of ε).

Many other models, but linear models are frequently used (at least for initial experiments) since they are easy to analyse.

The following general steps apply to other models, not just linear modes.

As for comparison experiments, run experiment trials and use observed responses to give parameter estimates.

Next step: choose which experiments to run to give accurate estimates as efficiently as possible.

- Linear Models
- Experimental Designs
  - Experimental Designs
  - Factorial Design
  - Fractional Factorial Designs
  - Other Designs
  - Replication
- Model Fitting (Analysis)
- Nuisance Factors
- Problem Instances
- Model Interpretation

Experimental Designs

Definition (Experimental Design)

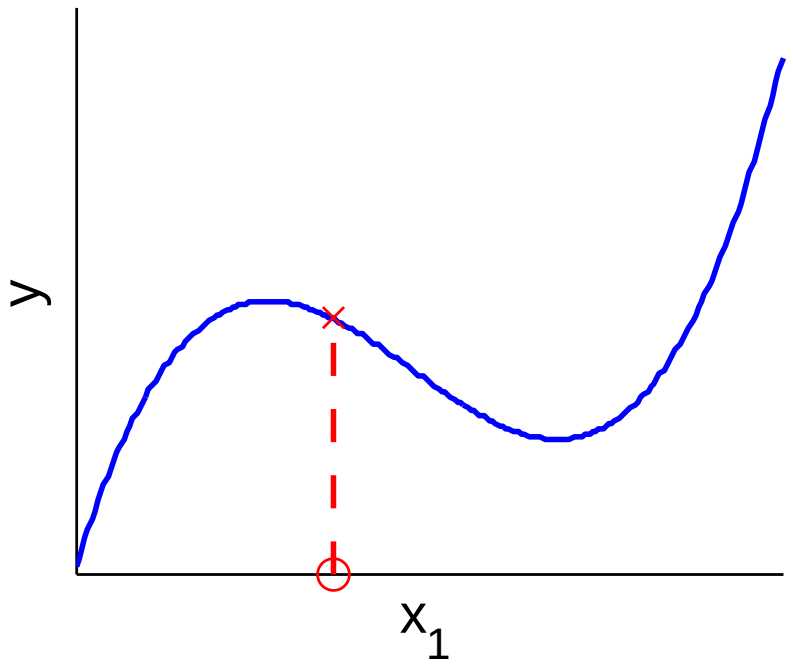
An experimental design is a set of factor settings (design points) for experimental trials.

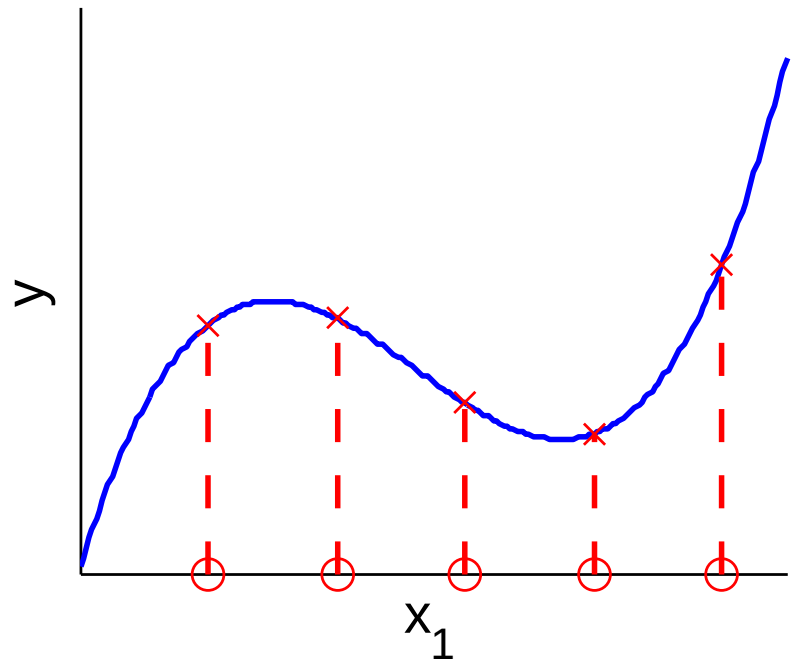
Generally, experimental designs are chosen that:

- 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.

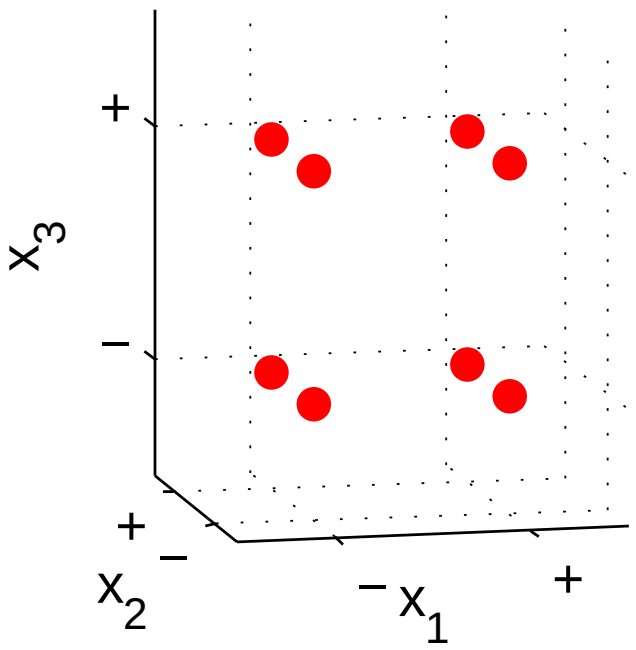
Experimental Designs





Factorial Design

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



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

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

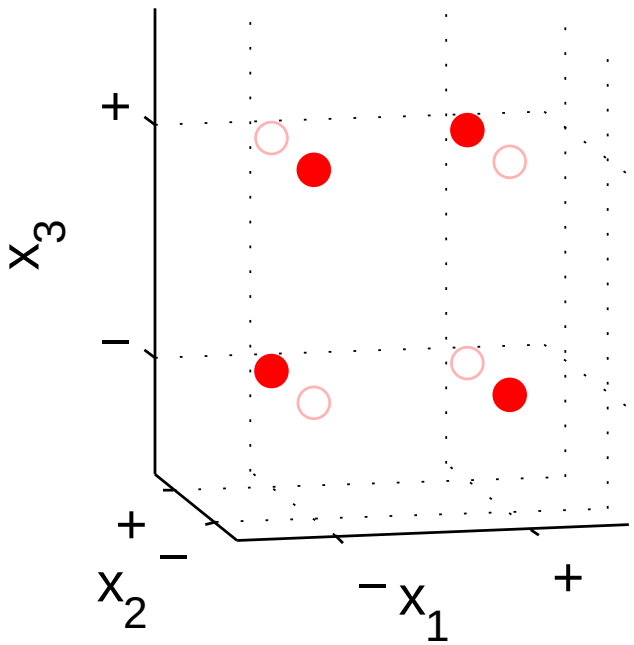
Fractional Factorial Designs

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

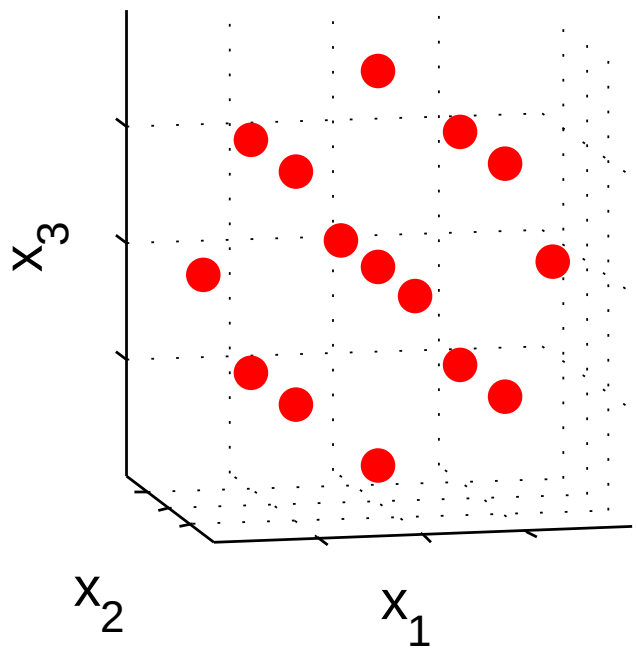
Fractional Factorial Designs



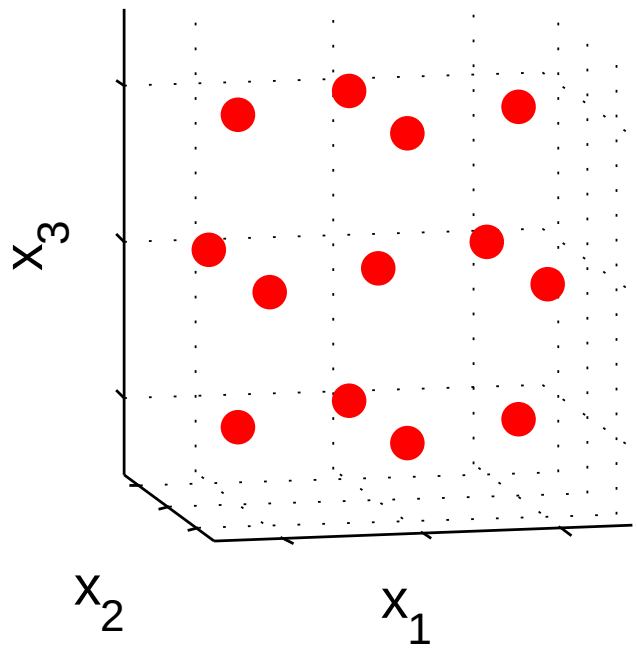
x_1	x_2	x_3
-	-	+
-	+	-
+	-	-
+	+	+

Other Designs - Examples

Central Composite Design



Box-Behnken Design



Replication

Repeating a number of trials at the same design point improves the accuracy of the estimated model parameters.

Most analysis techniques accommodate replication of design points.

Exercise

factor	parameter	good range
x_1	initial temperature	50 – 100
x_2	cooling	10 – 20
x_3	number of samples (integer)	2 – 7

Suggest (a) factorial and (b) fractional factorial designs.

Exercise

factor	parameter	good range
x_1	initial temperature	50 – 100
x_2	cooling	10 – 20
x_3	number of samples (integer)	2 – 7

Suggest (a) factorial and (b) fractional factorial designs.

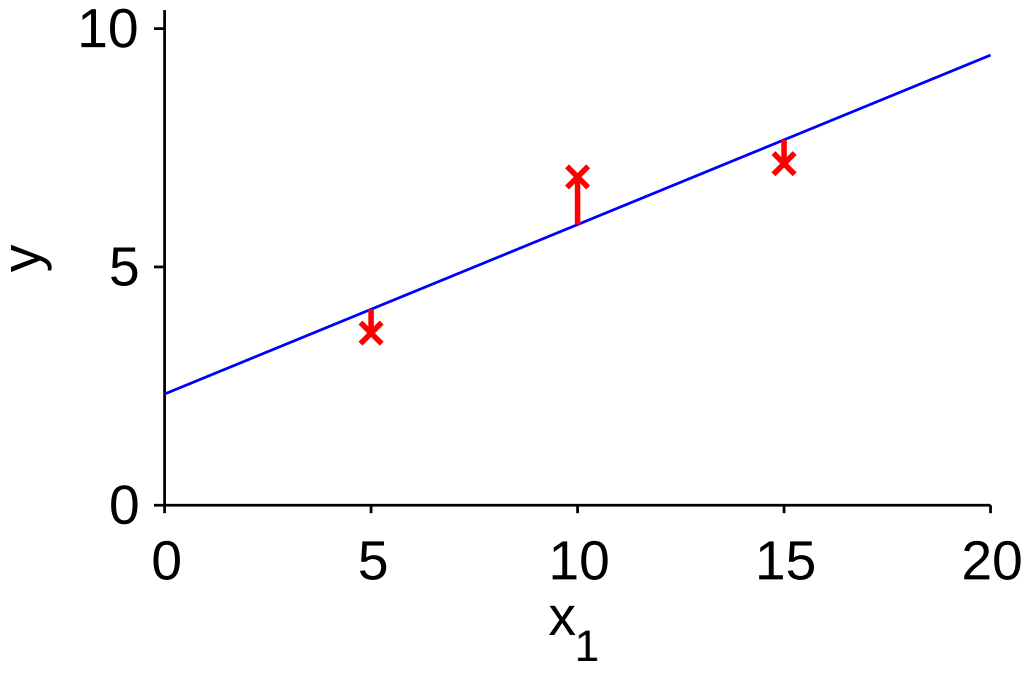
x_1	x_2	x_3
60	12	3
60	12	6
60	18	3
60	18	6
90	12	3
90	12	6
90	18	3
90	18	6

x_1	x_2	x_3
60	12	3
60	18	6
90	12	6
90	18	3

- Linear Models
- Experimental Designs
- Model Fitting (Analysis)
  - Least Squares Linear Regression
  - Maximum Likelihood Estimation
  - Residuals
  - ANOVA
- Nuisance Factors
- Problem Instances
- 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 - Example

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

parameter	estimate
β_0	46.50
β_1	1.98
β_2	-4.92
β_3	0.67

Maximum Likelihood Estimation

Finds parameters $\hat{\beta}$ (and variance $\hat{\sigma}$) that is most likely to have resulted in the observed responses.

For simple linear model, estimates are the same as least squares linear regression.

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

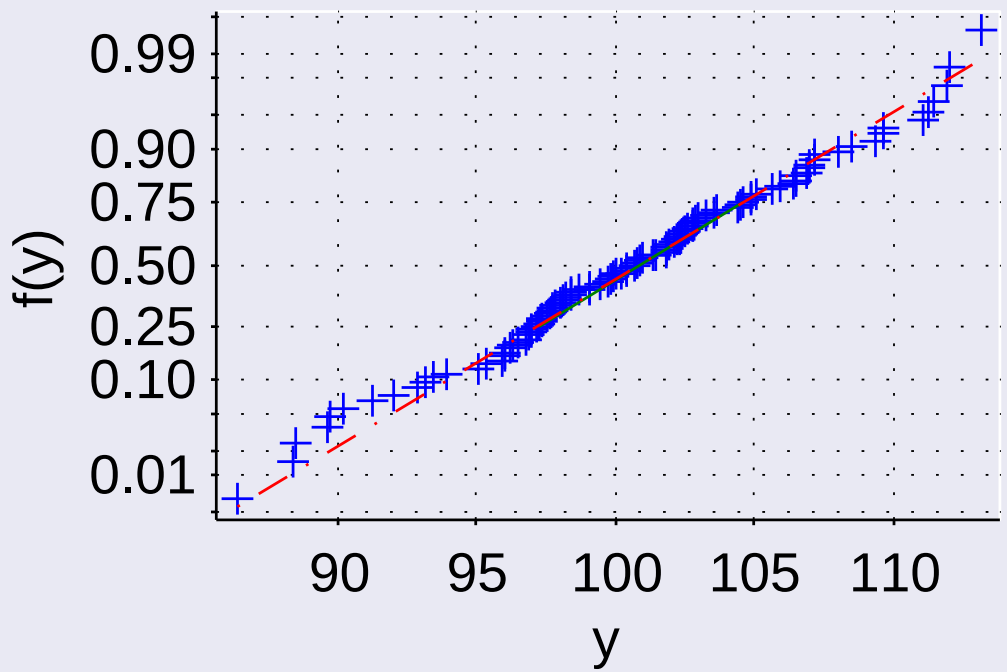
x_1	x_2	x_3	y	$\hat{y}$	$\hat{\varepsilon}$
60	12	3	107	108.5	-1.5
60	12	6	114	110.5	3.5
60	18	3	86	79.0	7.0
60	18	6	72	81.0	-9.0
90	12	3	163	168.0	-5.0
90	12	6	173	170.0	3.0
90	18	3	138	138.5	-0.5
90	18	6	143	140.5	2.5

Residuals

Residuals can be used to analyse effectiveness of the model.

For example, for a linear model, residuals should show a Normal distribution.

Normal Plot



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

(Similar to: H_0: $\beta_i = 0$ $\quad$ H_1: $\beta_i \neq 0$).

Linear regression and MLE also give p-values and/or confidence intervals.

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 - Example

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)
> anova(lm(y ~ x1 + x2 + x3))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 7080.5  7080.5 153.092 0.0002452 ***
x2         1 1740.5  1740.5  37.632 0.0035790 **
x3         1    8.0     8.0   0.173 0.6988276
Residuals  4  185.0    46.2
```

Screening

If too many factors, use a screening experiment with a simple linear model and a small design (e.g. fractional factorial design).

Use (for example) ANOVA to identify the significant factors.

- Linear Models
- Experimental Designs
- Model Fitting (Analysis)
- Nuisance Factors
  - Nuisance Factors
  - Blocking
  - Randomisation
- Problem Instances
- Model Interpretation

Nuisance Factors

Definition

Nuisance factors are (controllable) factors that affect the response other than the ones we're interested in.

Examples

- PC on which an experiment trial is run.
- The problem instance used for a particular trial.
- Time at which the experiment trial occurs (if, for example, performance degrades systematically over time).

Nuisance Factors

Can handle nuisance factors in a number of ways:

1. Eliminate by using a different response is affected by the factor. (Example: function evaluations rather than runtime.)
2. Blocking.
3. Randomisation.
4. Include as part of the model.

Blocking

Definition

Blocking occurs when primary factors occur the same number of times at each level of the nuisance factor.

In effect, the nuisance factor becomes a factor in a systematic design so that its effect on the primary factors is eliminated.

Blocking

Example

Two PCs are available to run trials, but each have slightly different performance.

To make the PC a blocking factor, we could run a fractional factorial design where the PC is a factor in the design.

x_1	x_2	PC
-	-	A
-	+	B
+	-	B
+	+	A

Randomisation

Definition

In randomisation, the value of the nuisance factor is set 'randomly' for each experimental trial.

The intention is that the random assignment minimises the effect of the nuisance factor on the primary factors.

Example

Two PCs are available to run trials, but each have slightly different performance.

Each experimental trial (for the primary factors) is assigned randomly to one of the two PCs.

- Linear Models
- Experimental Designs
- Model Fitting (Analysis)
- Nuisance Factors
- Problem Instances
  - Problem Instances as Nuisance Factors
  - Modelling Problem Instances
- Model Interpretation

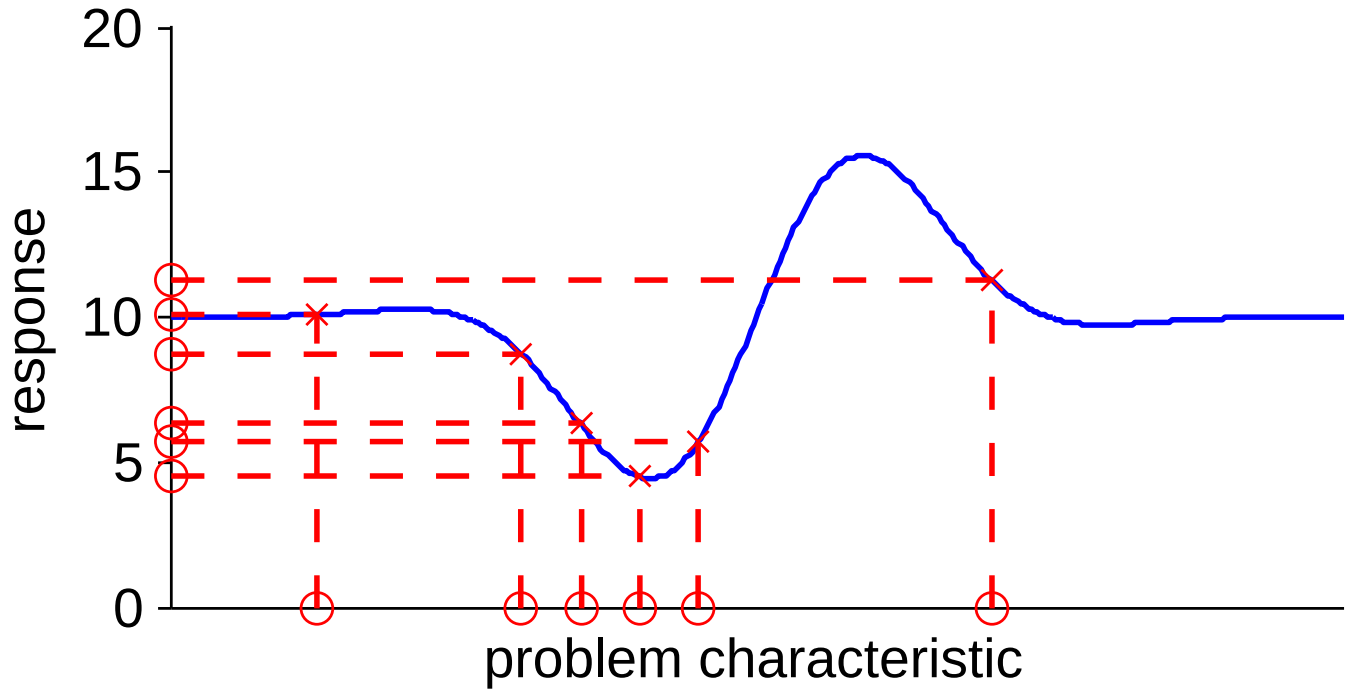
Problem Instances

Problem instances are also factors (have different problem characteristics that affect the response), and may be nuisance factors.

The choice of problem instances can affect the response model estimated by the experiment.

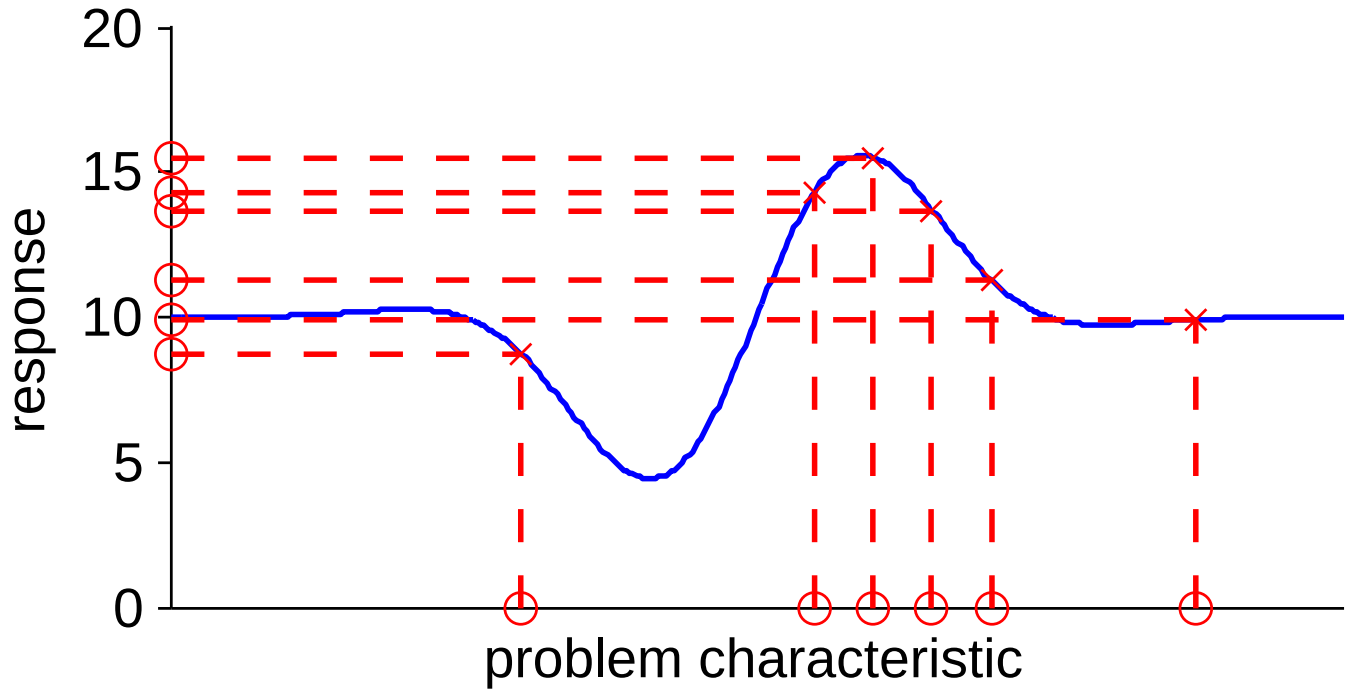
Choice of Problem Instances

Consider an algorithm with no parameters, and one significant problem characteristic.



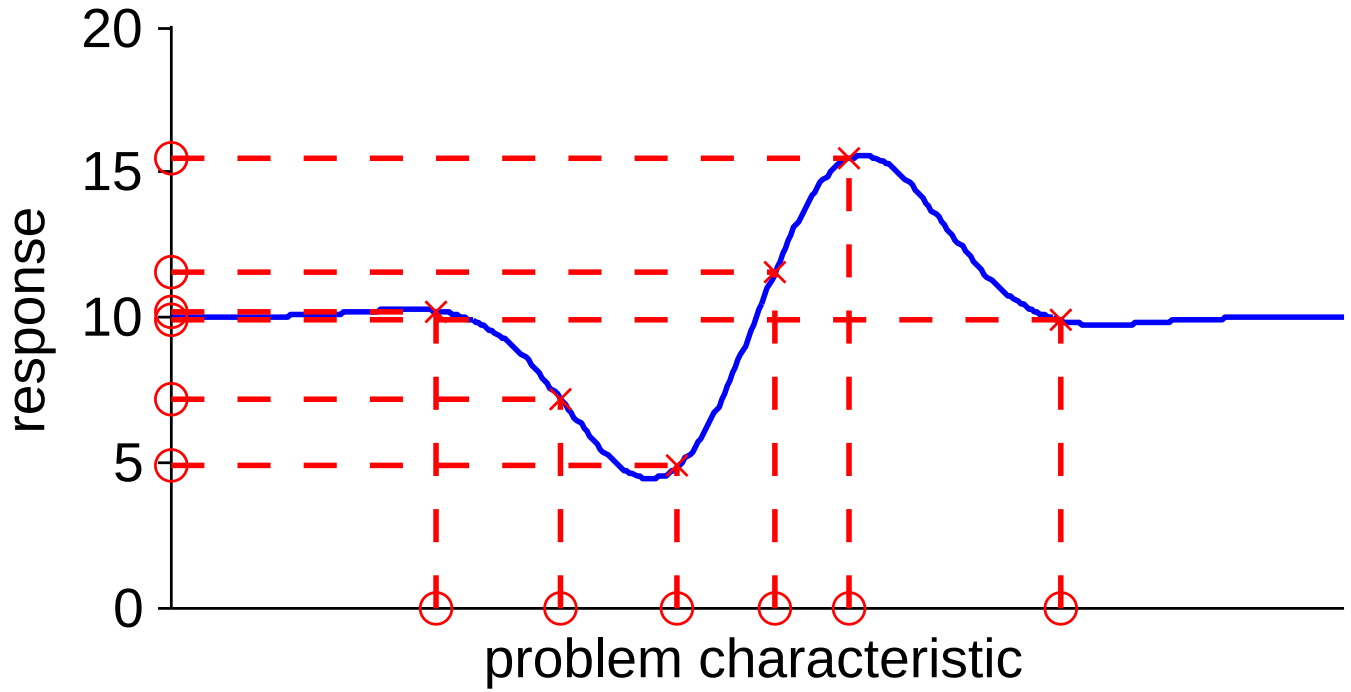
Choice of Problem Instances

Consider an algorithm with no parameters, and one significant problem characteristic.



Choice of Problem Instances

Consider an algorithm with no parameters, and one significant problem characteristic.



Handling Problem Instances

Could handle problem instances using:

Blocking	Consider problem instances as part of design (e.g. run every design point on each problem instance).
Randomisation	Pick random problem instance(s) for each experimental trial.
Include in Model	Include factors that quantify the problem instances in the model.

Problem Instance Sets

Typical methods of creating sets of problem instances suffer from limitations:

random generation	What is correct probability distribution?
benchmarks	Often chosen to demonstrate good performance of existing algorithms.
real world test cases	Are they really a representative?

Modelling Problem Instances - Indicator Variables

Example (Linear Model with Indicator Variables)

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{j=2}^{m} \gamma_j p_j + \varepsilon$$

m problem instances, numbered $1, 2, \ldots, m$

p_j is indicator variable:

$$p_j = \begin{cases} 1 & \text{for problem instance } j \\ 0 & \text{otherwise.} \end{cases}$$

γ_j adjusts the model for the 'difficulty' of problem j compared to problem 1

Modelling Problem Instances - Problem Characteristics

Include problem characteristics as primary factors in the model.

Relevant problem characteristics may be problem size; distance standard deviation (TSP); ratio of tasks to processors (Task Allocation).

Resulting estimated model might express optimal algorithm parameters as functions of the problem characteristics, e.g. mutation rate as a function of problem size.

Modelling Problem Instances - Problem Characteristics

Advantages

- Enables algorithm to be tuned to particular problem instance (using the problem instance's characteristics).
- Can allow algorithm to be compared more accurately to existing algorithm.
- Avoid issue of a representative problem instance set.

Disadvantage

- Very difficult to derive suitable characteristics.

Exercise

For algorithm and problems you have considered, what problem characteristics affect the algorithm response?

- Linear Models
- Experimental Designs
- Model Fitting (Analysis)
- Nuisance Factors
- Problem Instances
- 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).

Model Assessment

A number of statistics exist that tests how well the model fits the observed data.

A poor fit may suggest that a different model should be used, e.g. higher-order linear model or a different form of model.

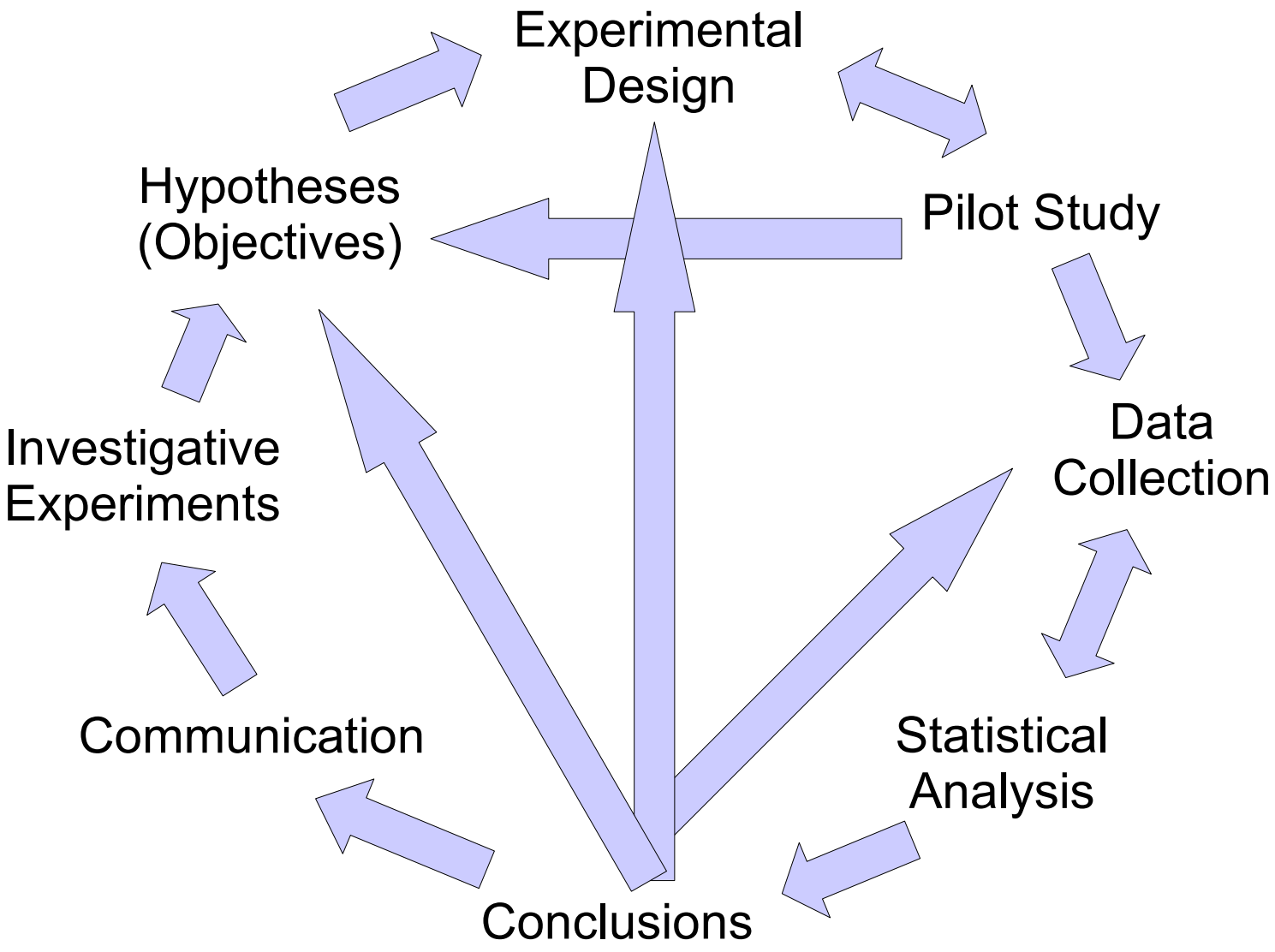
Summary

- Linear models as simple models.
- Experimental designs - factorial, fractional factorial, others.
- Model fitting - parameter estimation using linear regression and MLE.
- Nuisance factors - handled using blocking and randomisation.
- Problem instance - the difficulty of deriving a representative set of instances.
- Model interpretation - for algorithm parameter tuning, optimise using model.

Part V

Experimental Method

Non-Linear / Iterative Experimentation



Assumptions

- Be clear as to assumptions made (e.g. model types).
- Test assumptions before experiment (pilot studies, trial runs).
- Verify assumptions after experiment (e.g. analyse residuals).

Reproducibility

Both by you and by other researchers.

- Store everything necessary to re-run experiments (including problem instances).
- Store all output.
- Where practical, make data available to others.

Choice of Response

- Choose response metric carefully.
- Can eliminate many nuisance factors (e.g. number of function evaluations rather than time).
- May assist with reproducibility.
- Can avoid need to optimise implementation (and therefore allow ‘simulation’).

Visualisation

- Plot data as much as possible.
- Verification of statistical results.
- Provides additional insight (that summary statistics cannot).
- Communicates information effectively.

Communication

- State objectives and hypotheses.
- Communicate meaningful information, such p-values, confidence intervals etc.
- Provide via other means, e.g. website, when paper doesn't have sufficient room.
- Give sufficient details to enable reproduction.
- Do not ignore inconvenient or contradictory results.

Part VI

Resources

Tools

- R - free software, widely used
- MATLAB - requires licence
- SPSS - requires licence
- Others . . .

Resources I

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

book M Berthold, D Hand (eds)
Intelligent Data Analysis (2nd Ed)
Springer, 2003

paper D Johnson
A Theoretician's Guide to the Experimental Analysis of Algorithms
Proceedings of the 5th and 6th DIMACS Implementation Challenges

paper I P Gent, T Walsh
How Not To Do It
AAAI Workshop on Experimental Evaluation of Reasoning and Search Methods, 1994

Resources II

paper J Cohen
The Earth Is Round ($p < .05$)
American Psychologist, 49 (12), 997–1003, 1994

paper J N Hooker
Testing Heuristics: We Have It All Wrong
Journal of Heuristics, 1 (1), 33–42, 1995

paper J N Hooker
Needed: An Empirical Science of Algorithms
Operations Research, 42 (2), 201–212, 1996

workshop T Bartz-Beielstein, M Preuss
Experimental Research in EC
GECCO Workshop (2007 and previous years)

book T Bartz-Beielstein
Experimental Research in Evolutionary Computation
Springer, 2006

Resources III

paper E Ridge, D Kudenko
Analysing Heuristic Performance with Response Surface Models: Prediction, Optimisation and Robustness
GECCO 2007

paper O Kramer, B Gloger, A Goebels
An Experimental Analysis of Evolution Strategies and Particle Swarm Optimisers using Design of Experiments
GECCO 2007

paper S Poulding, P Emberson, I Bate, J Clark
An Efficient Experimental Methodology for Configuring Search-Based Design Algorithms
HASE 2007, to appear

manual MATLAB documentation