

# Experimental Methods Workshop II

Simon Poulding

University of York

June 2008



## Part I

### Introduction

## Primary Objective

To share knowledge and experience of reliable, efficient and principled experimentation, and associated statistical methods

Emphasis on **concepts, terminology and methods** rather than theory . . . with suggestions for further reading

Build on techniques and concepts covered during workshop last year. New themes this year are multivariate statistics and resampling techniques.

# Syllabus

- ① Concepts Refresher
- ② Hypothesis Testing - power, sample size
- ③ Correlation - joint distributions, correlation
- ④ Dimension Reduction - principal component analysis, factor analysis
- ⑤ Resampling - bootstrapping

- Ask questions at any time
- Exercises - as a group
- Plenty of breaks

Part II

Concepts Refresher

# Random Variables

# Random Variables

## Definition

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

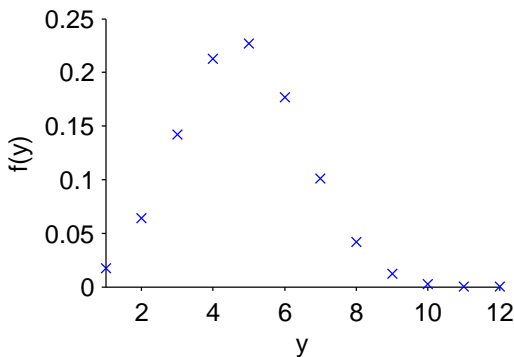


# Probability Mass Function (Discrete Distributions)

# Probability Mass Function (Discrete Distributions)

Probability of random variable  $Y$  having the value  $y$  is denoted  $f(y)$ :

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

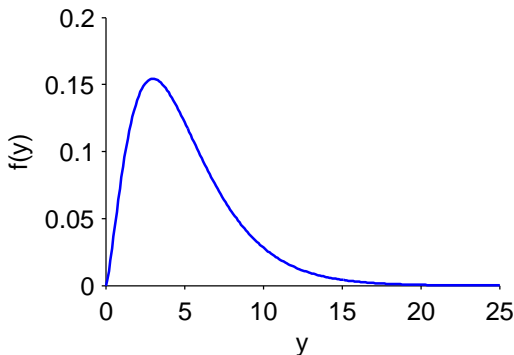


# Probability Density Function (Continuous Distributions)

# Probability Density Function (Continuous Distributions)

Probability of random variable  $Y$  having a value in the interval  $(a, b)$ :

$$\mathbb{P}(a < Y < b) = \int_a^b f(y) dy$$



# Named Probability Distributions

Discrete

Continuous

# Named Probability Distributions

## Discrete

Binomial  
Bernoulli  
Poisson  
Geometric  
Discrete Uniform

## Continuous

Normal (Gaussian)  
Student's  $t$   
 $F$   
Exponential  
Chi-Squared  
Gamma  
Beta  
Rayleigh  
Weibull  
Continuous Uniform

# Population and Sample

# Population and Sample

## Population

Refers to parameters and properties of a theoretical (idealised) distribution

## Sample

Refers to parameters and other values calculated from an empirical sample taken from an (unknown) distribution



# Expectation

## Definition

**Expectation** is the 'average' value of a quantity weighted by the probability of taking each value.

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

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

$$\mathbb{E}(Y) = \int_y y f(y) dy$$

# Population Mean and Variance

## Definition (Population Mean)

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

# Population Mean and Variance

## Definition (Population Mean)

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

## Definition (Population Variance)

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



## Definition

A **statistic** is a function applied to a sample of observed responses

## 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



# Sample Mean and Variance

# Sample Mean and Variance

## Definition (Sample Mean)

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

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

estimator of the population mean,  $\mu$

# Sample Mean and Variance

## Definition (Sample Mean)

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

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

estimator of the population mean,  $\mu$

## Definition (Sample Variance)

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

estimator of the population variance,  $\sigma^2$

# Mean and Variance of Sample Mean

# Mean and Variance of Sample Mean

## Theorem

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

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

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

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

# Distribution of Sample Mean

# Distribution of Sample Mean

## Theorem

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

# Distribution of Sample Mean

## Theorem

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

## Theorem (Central Limit Theorem)

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



## Part III

# Hypothesis Testing

- Example
- Formal Definition
- Hypothesis Tests
- Significance and Power
- Sample Size

# Hypothesis and Experiment

## Hypothesis

70% of adults in the UK prefer tea to coffee.

## Experiment

Pick 10 adults at random and ask if they prefer tea to coffee.  
Count the number who say “yes”.

## Simplifying Assumption

Since number of adults in the UK is much larger than the sample size, assume probability of saying “yes” is the same for each member of the sample.

## Model

$X$  = number of adults who answer “yes”.

$X$  is a random variable with a ? distribution

## Simplifying Assumption

Since number of adults in the UK is much larger than the sample size, assume probability of saying “yes” is the same for each member of the sample.

## Model

$X$  = number of adults who answer “yes”.

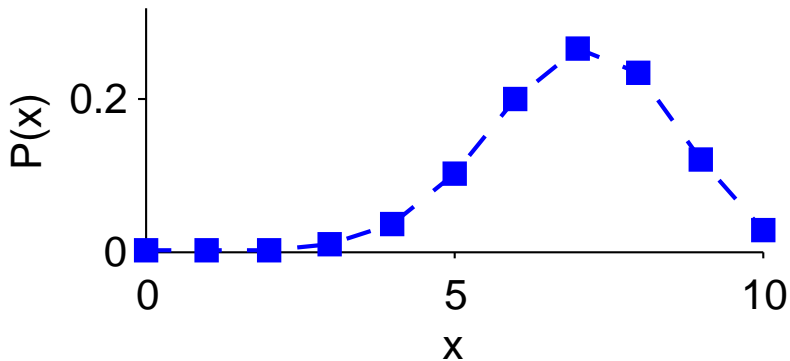
$X$  is a random variable with a **Binomial** distribution

What is the probability that  $X = 4$  if the hypothesis is correct?

What is the probability that  $X = 4$  if the hypothesis is correct?

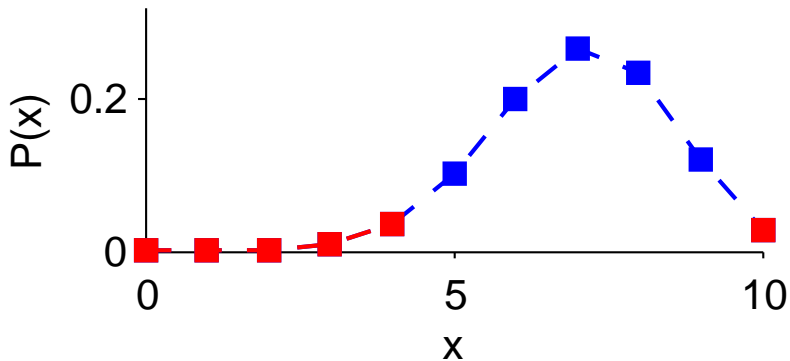
$$\begin{aligned}\mathbb{P}(X = 4) &= \binom{10}{4} (0.7)^4 (0.3)^6 \\ &= \frac{10!}{6!4!} (0.7)^4 (0.3)^6 \\ &\approx 0.0368\end{aligned}$$

# Analysis - Probability Distribution





## Analysis - Critical Region



# Analysis - Critical Region

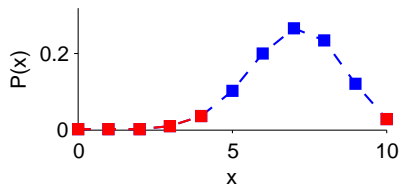
The critical region are those values of  $X$  that are unlikely to occur if the hypothesis is correct.

## Decision Rule

If the observed value of  $X$  lies in the critical region, then **reject** the hypothesis.

Otherwise **accept** the hypothesis.

## Analysis - Potential Errors



But even if hypothesis is correct, there is still a (small) chance that observed  $X$  is in the critical region.

Sum of probabilities in critical region is 0.0756.

- Example
- **Formal Definition**
- Hypothesis Tests
- Significance and Power
- Sample Size

# Summary

- 1 Define hypothesis
- 2 Decide an appropriate statistic (to be calculated from the observed data)
- 3 Assuming hypothesis to be true, determine the statistic's probability distribution
- 4 Identify values for the statistic that are unlikely 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.

## 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

Often pick one of the standard hypothesis-based tests for which the test statistic and its distribution have already been worked out.

### Example

test statistic:  $X$ , the number of people in sample who preferred tea to coffee

## 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 (a 'simple' hypothesis)
- In step 2, choose a test statistic whose distribution is relatively to calculate

If we use a standard test, then distribution of test statistic will be known.

### Example

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

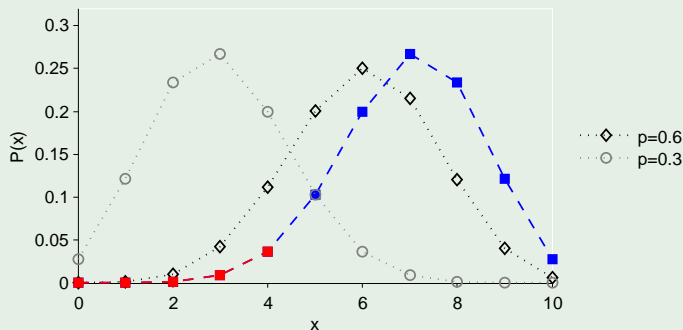


## Step 4 - Unlikely Test Statistic Values

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

### Example

$$H_0: p = 0.7 \quad H_1: p < 0.7$$

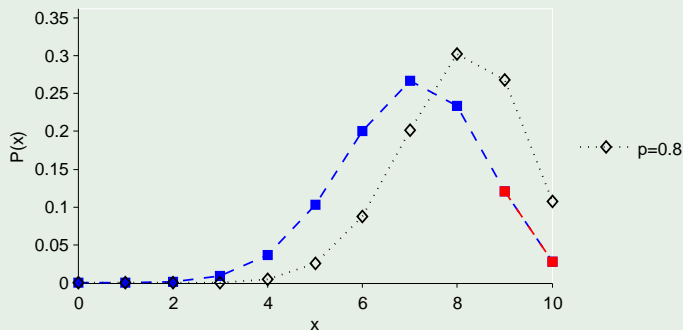


## Step 4 - Unlikely Test Statistic Values

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

### Example

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



## 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  $H_1$ ).

### Interpretation

If test statistic is not in critical region, then it means that there is not sufficient evidence in the observed data to reject it.

However, a very precise null hypothesis (such  $p = 0.7$ ) are unlikely to be absolutely correct. If true value of  $p$  is, say 0.7001, it would be difficult to collect sufficient evidence to disprove the hypothesis.

- Example
- Formal Definition
- Hypothesis Tests
- Significance and Power
- Sample Size

# Hypothesis Test Examples

## Parametric

z-Test

t-Test

Binomial Test

$\chi^2$  Goodness of Fit

ANOVA

## Non-Parametric

Sign Test

Rank Sum Test (Mann-Whitney-Wilcoxon)

Signed Rank Test (Wilcoxon)

Kruskal-Wallis (non-parametric ANOVA)

## Example - Rank Sum Test

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

- Example
- Formal Definition
- Hypothesis Tests
- Significance and Power
- Sample Size

## Definition (Type I Error)

$H_0$  is really **true**, but it is **rejected** by the test.

## Definition (Type II Error)

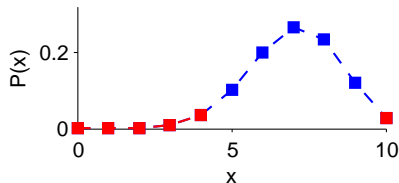
$H_0$  is really **false**, but it is **accepted** by the test.

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  $\alpha$ , then this is the chance of a type I error. It is the **significance** of the test.

When there is a choice, a critical region is often chosen so that the significance is 5%.

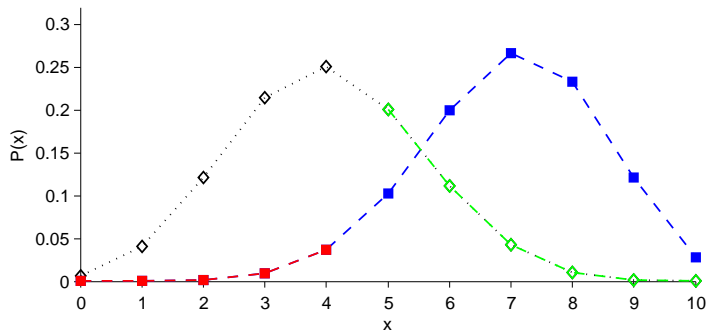
# Type II Errors

Easiest to consider when  $H_1$  is a simple hypothesis:

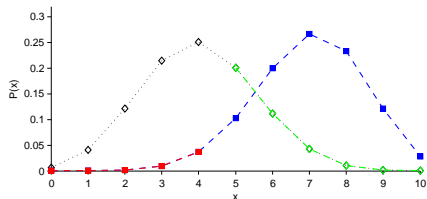
## Example

$$H_0: p = 0.7$$

$$H_1: p = 0.4$$



## Type II Errors and Power



The chance of type II error is denoted  $\beta$ . (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  $\alpha$  is chosen, and a critical region of this size found that minimises the size of  $\beta$  (or, equivalently, maximises the power). This is called the 'best' test.

It can be shown that this optimal critical region are those values for ratio of probability given  $H_1$  to probability given  $H_0$  is the greatest.

- Example
- Formal Definition
- Hypothesis Tests
- Significance and Power
- Sample Size

# Example

## Hypotheses

Denote the IQ of students at KCL as the r.v.  $Y$ , and  $\mu_Y$  is the mean of  $Y$ .

$$H_0: \mu_Y = 100$$

$$H_1: \mu_Y = 110$$

## Assumptions

- $Y$  has a Normal distribution
- variance of  $Y$  is 250

## Design Decisions

- Test at 5% significance level
- Use the sample mean,  $\bar{Y}$  of a sample of size  $N$  as the test statistic

## Exercise

If the sample size is 10, what is the critical region for  $\bar{Y}$ ? What is the power of the test?

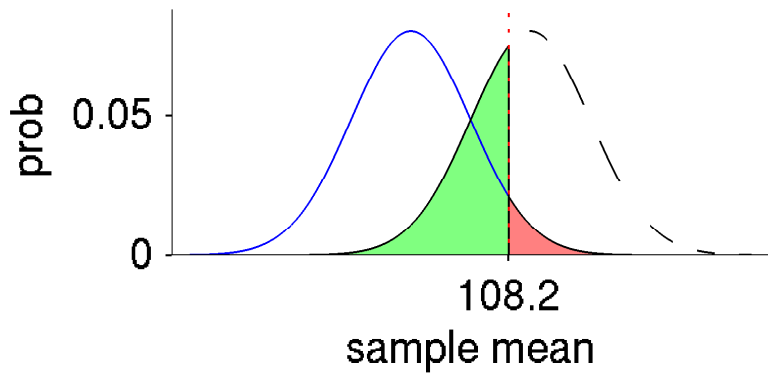
## Exercise

If the sample size is 10, what is the critical region for  $\bar{Y}$ ? What is the power of the test?

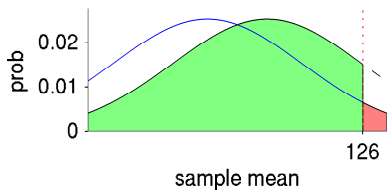
- 1 Since  $Y$  is normally distributed, then so is  $\bar{Y}$ .
- 2 Variance of  $\bar{Y}$  is  $\frac{250}{10} = 25$ .
- 3 If  $H_0$  is true, the mean of  $\bar{Y}$  is 100.
- 4 Since  $H_1$  is more likely to be true than  $H_0$  when the value of  $\bar{Y}$  is large, we choose the right-hand tail for the critical region.
- 5 Using MATLAB, R, tables etc., find that 5% critical region on right tail for normal distribution with mean 100, variance 25 is when  $\bar{Y} \geq 108.2$ .
- 6 If  $H_1$  is true, then  $\bar{Y}$  has a normal distribution with mean 110 and variance 25. The probability that a variable with this distribution has a value less than 108.2 (i.e. outside the critical region) is 0.361, so the power is  $1 - 0.361 = 0.639$ .



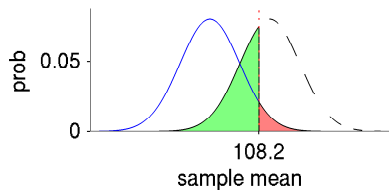
## Exercise



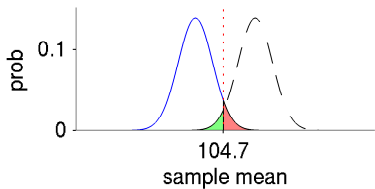
# Different Sample Sizes



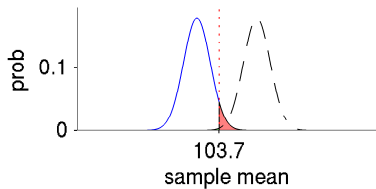
$$N = 1 \quad \beta = 0.844$$



$$N = 10 \quad \beta = 0.361$$

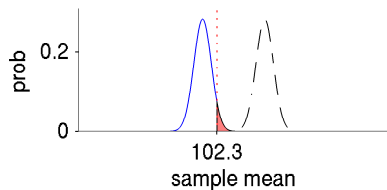


$$N = 30 \quad \beta = 0.0344$$

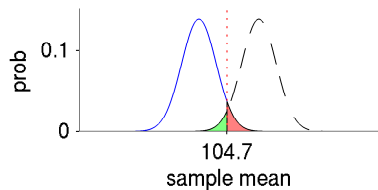


$$N = 50 \quad \beta = 0.00235$$

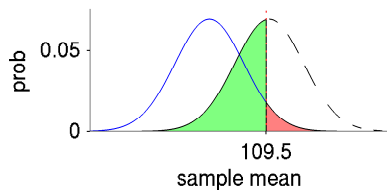
# Different Variances ( $N = 30$ )



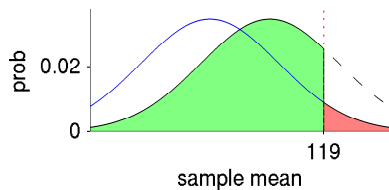
$$\sigma_Y^2 = 60 \quad \beta = 2.88 \times 10^{-8}$$



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

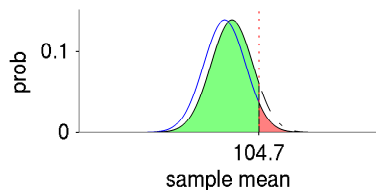


$$\sigma_Y^2 = 1000 \quad \beta = 0.467$$

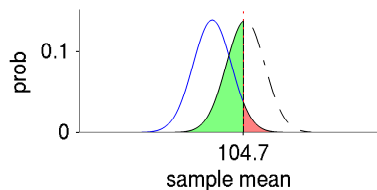


$$\sigma_Y^2 = 4000 \quad \beta = 0.782$$

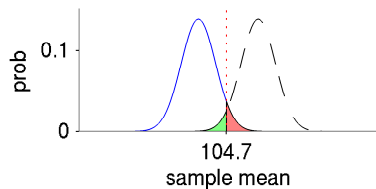
# Different Effect Sizes ( $N = 30, \sigma_Y^2 = 250$ )



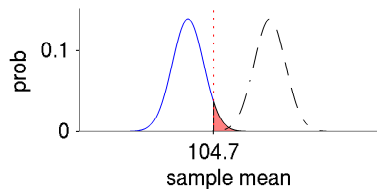
$$H_1: \mu_Y = 101 \quad \beta = 0.903$$



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

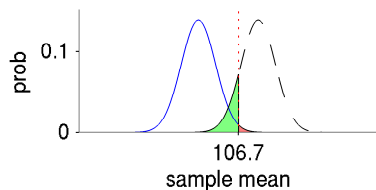


$$H_1: \mu_Y = 110 \quad \beta = 0.0344$$

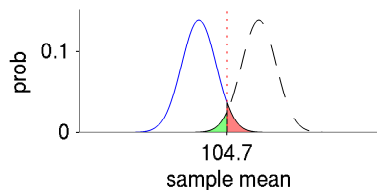


$$H_1: \mu_Y = 115 \quad \beta = 0.000192$$

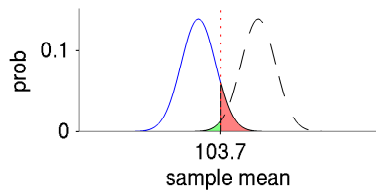
# Different Significances ( $N = 30, \sigma_Y^2 = 250, H_1: \mu_Y = 110$ )



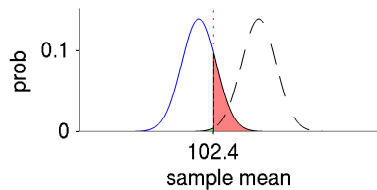
$$\alpha = 0.01 \quad \beta = 0.128$$



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



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



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

# 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.

# Example - Calculating Sample Size and Power

R

```
> power.t.test(n=NULL,power=0.90,delta=10,sd=16,sig.level=0.05,type="one.sample",alt="one.sided")
```

```
One-sample t test power calculation
```

```
      n = 23.34489
  delta = 10
     sd = 16
sig.level = 0.05
  power = 0.9
alternative = one.sided
```

```
> power.anova.test(groups=4,n=10,between.var=10,within.var=20,sig.level=0.05,power=NULL)
```

```
Balanced one-way analysis of variance power calculation
```

```
  groups = 4
        n = 10
between.var = 10
within.var = 20
sig.level = 0.05
  power = 0.882607
```

NOTE: n is number in each group



Part IV

Correlation

- Multivariate Probability Distributions
- Correlation Coefficient
- Sample Correlation
- Rank Correlation

# Multivariate Probability Distributions

Last workshop considered distribution of a single variable - univariate probability distributions.

In this section (and later), we consider distributions of two or more variables - [multivariate probability distributions](#).

## Example

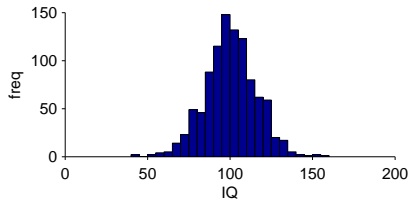
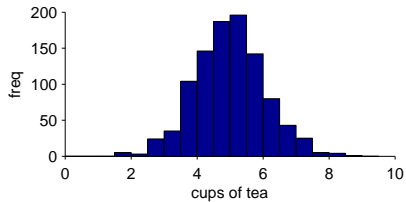
I'm interested in the relationship between IQ and the amount of tea a person drinks.

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

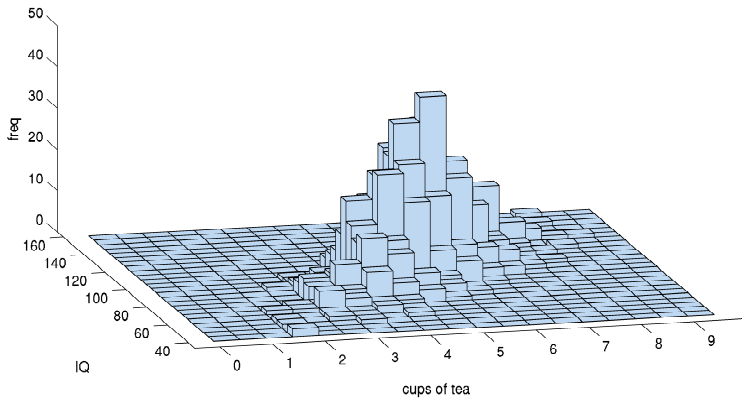
### Sample Data

| <u>cups of tea</u> | <u>IQ</u> |
|--------------------|-----------|
| 6.3                | 109       |
| 3.2                | 73        |
| 4.7                | 95        |
| 4.9                | 114       |
| 5.0                | 90        |
| ⋮                  | ⋮         |

# Frequency Histograms



# Frequency Histograms



# Joint Probability Distribution

## Univariate

Probability Mass Function:

$$\mathbb{P}(X = x) = f(x)$$

Probability Density Function:

$$\mathbb{P}(a < X < b) = \int_{x=a}^b f(x) dx$$

## Multivariate

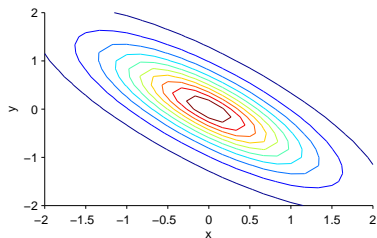
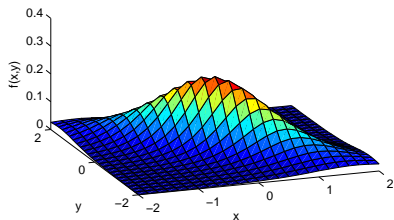
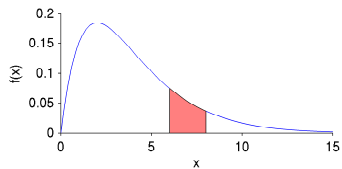
Joint Probability Mass Function:

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

Joint Probability Density Function:

$$\mathbb{P}(a < X < b, c < Y < d) = \int_{x=a}^b \int_{y=c}^d f(x, y) dx dy$$

# Joint Probability Distribution





# Marginal Distributions and Independence

## Marginal Distributions

$$g(x) = \sum_y f(x, y)$$

$$h(y) = \sum_x f(x, y)$$

## Definition (Independence)

X, Y are independent if and only if:

$$f(x, y) = g(x)h(y)$$

where  $g(\cdot)$  and  $h(\cdot)$  are the marginal distributions

- Multivariate Probability Distributions
- Correlation Coefficient
- Sample Correlation
- Rank Correlation

# Covariance and Correlation

## Definition (Variance)

$$\text{var}(X) = \sigma_X^2 = \mathbb{E}[(X - \mu_X)^2]$$

## Definition (Covariance)

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

## Definition (Correlation Coefficient (Pearson))

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Note:  $-1 \leq \rho \leq 1$

## Exercise

X and Y are discrete random variables. Each can take only the values -1 and 1. Their joint probability function is defined by:

$$f(x, y) = \begin{cases} 0.4 & (x, y) = (1, 1) \\ 0.2 & \text{otherwise} \end{cases} \quad (1)$$

$$f(x, y) = 0.25 \quad (2)$$

$$f(x, y) = \begin{cases} 0.4 & (x, y) = (1, -1) \\ 0.2 & \text{otherwise} \end{cases} \quad (3)$$

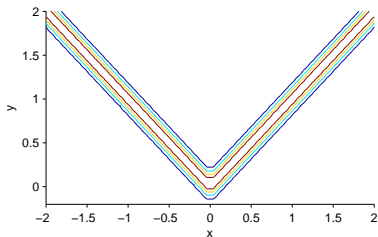
- 1 determine the marginal distributions for X and Y
- 2 are X and Y independent?
- 3 calculate the mean and variance of the marginal distributions
- 4 calculate the covariance of X, Y
- 5 calculate the correlation coefficient of X, Y

# Exercise

|                      | (1)    | (2) | (3)     |
|----------------------|--------|-----|---------|
| $\mathbb{P}(X = -1)$ | 0.4    | 0.5 | 0.4     |
| $\mathbb{P}(Y = -1)$ | 0.4    | 0.5 | 0.6     |
| independent?         | No     | Yes | No      |
| $\mu_X$              | 0.2    | 0   | 0.2     |
| $\mu_Y$              | 0.2    | 0   | -0.2    |
| $\text{var}_X$       | 0.96   | 1   | 0.96    |
| $\text{var}_Y$       | 0.96   | 1   | 0.96    |
| $\text{cov}(X, Y)$   | 0.16   | 0   | -0.16   |
| $\rho(X, Y)$         | 0.1667 | 0   | -0.1667 |

$$\rho = 0$$

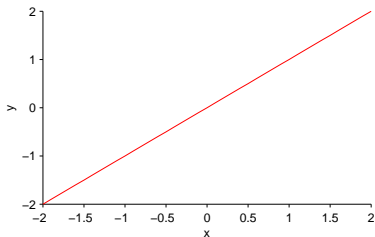
- if  $X$  and  $Y$  are independent, then  $\rho = 0$
- converse does **NOT** hold— $X, Y$  can be dependent but nevertheless have  $\rho = 0$ :



- ... **unless**  $f(x, y)$  is a multivariate normal distribution

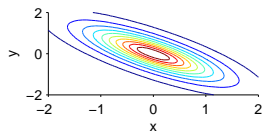
$$\rho = \pm 1$$

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

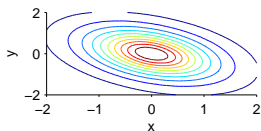


## Other values of $\rho$

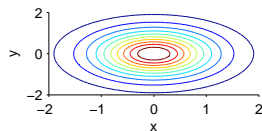
As  $|\rho|$  gets closer to 1, X and Y get closer to a linear dependence:



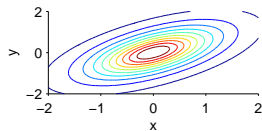
$$\rho = -0.8$$



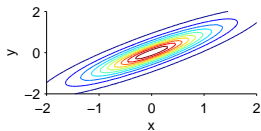
$$\rho = -0.4$$



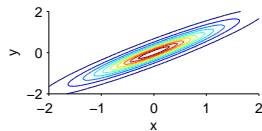
$$\rho = 0$$



$$\rho = 0.6$$



$$\rho = 0.9$$



$$\rho = 0.95$$



- Multivariate Probability Distributions
- Correlation Coefficient
- Sample Correlation
- Rank Correlation

# Sample Estimators

## Mean

Sample Estimator for  $\mu_x$ :

$$\bar{x} = \frac{1}{n} \sum x_i$$

## Variation

Sample Estimator for  $\sigma_x^2$ :

$$s_x^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$$

# Sample Estimators for Covariance and Correlation

## Covariance

$$s_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

## Correlation Coefficient (Pearson)

$$r = \frac{s_{xy}}{s_x s_y}$$

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

# Example

## Sample Data

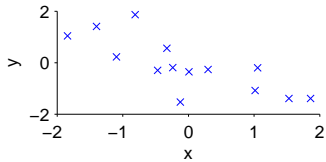
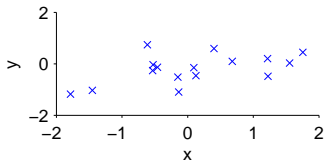
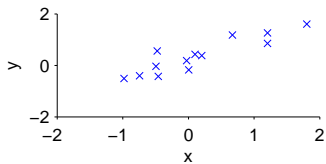
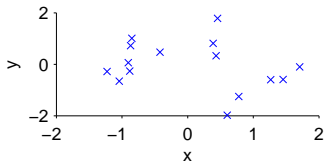
| <u>cups of tea</u> | <u>IQ</u> |
|--------------------|-----------|
| 6.3                | 109       |
| 3.2                | 73        |
| 4.7                | 95        |
| 4.9                | 114       |
| 5.0                | 90        |
| ⋮                  | ⋮         |

## R

```
> tea = c(6.3,3.2,4.7,4.9,5.0)
> iq = c(109,73,95,114,90)
> cor(tea,iq)
[1] 0.7936704
```

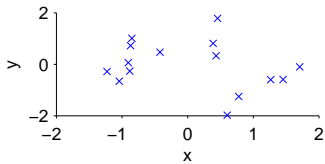
# Exercise

The scatter plots below are for samples which have sample correlation coefficients of  $-0.89$ ,  $-0.11$ ,  $0.54$ , or  $0.94$ . Which sample correlation coefficient belongs to which diagram?

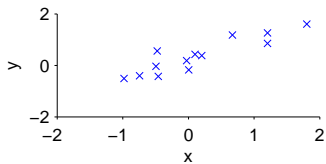


# Exercise

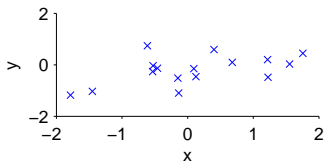
The scatter plots below are for samples which have sample correlation coefficients of  $-0.89$ ,  $-0.11$ ,  $0.54$ , or  $0.94$ . Which sample correlation coefficient belongs to which diagram?



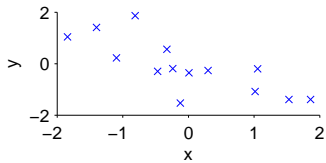
$$r = -0.11$$



$$r = 0.94$$



$$r = 0.54$$



$$r = -0.89$$

# 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
- ... **unless** we know  $X$  and  $Y$  have a normal distribution

## 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 Coefficient
- Sample Correlation
- Rank Correlation



# Spearman's Rank Correlation

- 1 Rank the observations,  $x_i$  of  $X$  in order; denote the rank  $(1, 2, \dots, n)$  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

## Alternative Formula

$$r' = 1 - \frac{6 \sum_i (y'_i - x'_i)^2}{n(n^2 - 1)}$$

where  $n$  is the number of observations

Note: can only be used when there are no ranking ties

# Example

## Sample Data

| <u>cups of tea</u> | <u>IQ</u> |
|--------------------|-----------|
| 6.3                | 109       |
| 3.2                | 73        |
| 4.7                | 95        |
| 4.9                | 114       |
| 5.0                | 90        |
| ⋮                  | ⋮         |

## R

```
> tea = c(6.3,3.2,4.7,4.9,5.0)
> iq = c(109,73,95,114,90)
> cor(tea,iq,method="spearman")
[1] 0.5
```

## Exercise

For the tea and IQ sample data, confirm the calculation of the Spearman Calculative using the alternative formula.

### Sample Data

| cups of tea $x_i$ | rank $x'_i$ | IQ $y_i$ | rank $y'_i$ | $(y'_i - x'_i)^2$ |
|-------------------|-------------|----------|-------------|-------------------|
| 6.3               |             | 109      |             |                   |
| 3.2               |             | 73       |             |                   |
| 4.7               |             | 95       |             |                   |
| 4.9               |             | 114      |             |                   |
| 5.0               |             | 90       |             |                   |

## Exercise

For the tea and IQ sample data, confirm the calculation of the Spearman Calculative using the alternative formula.

### Sample Data

| cups of tea $x_i$ | rank $x'_i$ | IQ $y_i$ | rank $y'_i$ | $(y'_i - x'_i)^2$ |
|-------------------|-------------|----------|-------------|-------------------|
| 6.3               | 5           | 109      | 4           | 1                 |
| 3.2               | 1           | 73       | 1           | 0                 |
| 4.7               | 2           | 95       | 3           | 1                 |
| 4.9               | 3           | 114      | 5           | 4                 |
| 5.0               | 4           | 90       | 2           | 4                 |

$$\begin{aligned}r' &= 1 - \frac{6 \sum_i (y'_i - x'_i)^2}{n(n^2 - 1)} \\ &= 1 - \frac{6 \times (1 + 0 + 1 + 4 + 4)}{5 \times (5^2 - 1)} = 1 - \frac{6 \times 10}{5 \times 24} = 1 - \frac{60}{120} = 0.5\end{aligned}$$

Part V

## Dimension Reduction Techniques

- Motivation
- Principal Component Analysis (PCA)
- Factor Analysis

# Motivation - Dimension Reduction

Suppose we have a large number of variables for each sample point:

## Sample Data

| $x_1$ | $x_2$ | $x_3$ | ... | $x_p$ |
|-------|-------|-------|-----|-------|
| 19.3  | 10.9  | 295   |     | 80.2  |
| 45.2  | 18.4  | 351   |     | 93.4  |
| 14.7  | 20.0  | 284   |     | 78.8  |
| 33.0  | 8.7   | 403   |     | 83.2  |
| ⋮     | ⋮     | ⋮     |     | ⋮     |

The number of variables might be too many to conveniently handle. For example, it may make identifying structure in the data (using visualisation, cluster analysis etc.) much more difficult.

If some of the variables are correlated, it might be possible to use a smaller number of variables derived from the observations, while retaining the important features of the data.

- Motivation
- Principal Component Analysis (PCA)
- Factor Analysis



# Principal Component Analysis

Initially have  $n$  observations, and each observation consists of  $p$  variables:  $x_1, x_2, \dots, x_p$ . (Without loss of generality, assume  $\bar{x}_i = 0$ .)

PCA derives  $p$  new variables,  $z_j$ , (called the principal components), that are linear combinations of the  $x_i$ :

$$z_j = a_{j,1}x_1 + a_{j,2}x_2 + \dots + a_{j,p}x_p$$

The variables  $z_j$  are chosen so that they are uncorrelated: this means that each  $z_j$  describes a unique 'dimension' of the observed data.

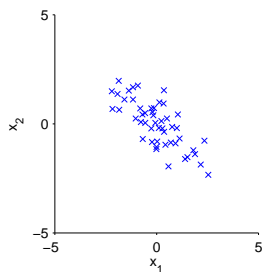
# Principal Component Analysis

PCA also chooses the linear combinations so that  $z_1$  has as large a variance as possible. Given the choice for  $z_1$ ,  $z_2$  is then chosen to have as much variance as possible (which must be  $\leq \text{var}(z_1)$ ), and so on.

The hope is that the first few components will show large variances, and most of the remainder will have low variance. If so, the majority of the variance in the observed data can be 'described' using only these first few components.

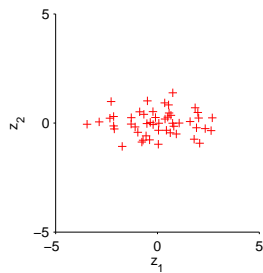
# Example 1

Original Data



$$\text{var}(x_1) = 1.36 \quad \text{var}(x_2) = 1.14$$

Scores



$$\text{var}(z_1) = 2.19 \quad \text{var}(z_2) = 0.30$$

$$z_1 = -0.75x_1 + 0.66x_2$$

$$z_2 = 0.66x_1 + 0.75x_2$$

# Choosing the Number of Components

General rule of thumb is to pick enough of the components such that together they account for about 80% or more of the total variance. (There are other more sophisticated techniques.)

In example above, variance of  $z_1$  accounted for 88% of the total variance.

## Scaling

If original variables have greatly differing scales, then one component might erroneously dominate all others in terms of variance.

Consider scaling the data so that each variable is approximately the same, or use correlation matrix as input to PCA in place of the raw data.

## Example 2 - Eurovision Voting



## Example 2 - Eurovision Voting

### Votes 2008

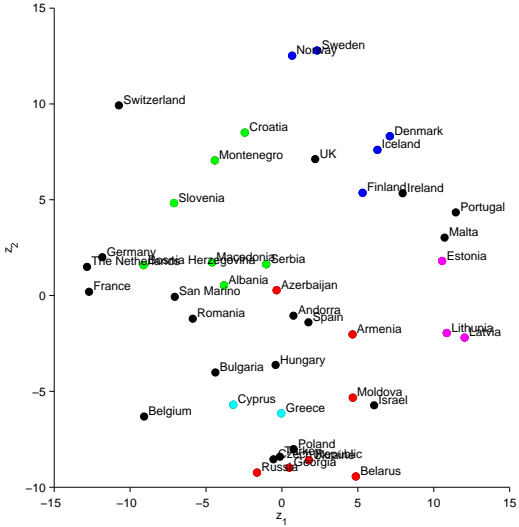
| Voting Country | $x_1$ | $x_2$ | $x_3$ | ... | $x_{16}$ |
|----------------|-------|-------|-------|-----|----------|
| Ukraine        | 12    | 0     | 2     |     | 0        |
| Turkey         | 5     | 8     | 7     |     | 3        |
| Montenegro     | 8     | 4     | 6     |     | 0        |
| Estonia        | 12    | 4     | 1     |     | 0        |
| UK             | 0     | 5     | 12    |     | 1        |
| Belgium        | 3     | 1     | 8     |     | 4        |
| ⋮              | ⋮     | ⋮     | ⋮     |     | ⋮        |

## Example 2 - Eurovision Voting

### Component Variances

| Component | Variance | % Total Variance |
|-----------|----------|------------------|
| $z_1$     | 45.3     | 22.3             |
| $z_2$     | 38.1     | 18.7             |
| $z_3$     | 23.0     | 11.3             |
| $z_4$     | 19.6     | 9.6              |
| $z_5$     | 14.8     | 7.3              |
| $z_6$     | 11.0     | 5.4              |
| $z_7$     | 10.2     | 5.0              |
| $\vdots$  | $\vdots$ | $\vdots$         |
| $z_{15}$  | 2.1      | 1.0              |
| $z_{16}$  | 0.7      | 0.3              |

# Example 2 - Eurovision Voting





## Example 3 - University Performance Indicators

### Data

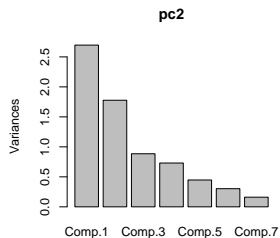
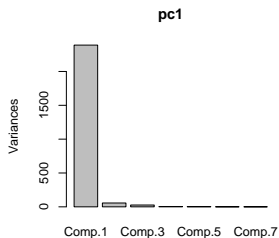
| Institution           | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ |
|-----------------------|-------|-------|-------|-------|-------|-------|-------|
| Oxford                | 91    | 10.0  | 12    | 80    | 7.1   | 511   | 68    |
| LSE                   | 78    | 7.4   | 13    | 85    | 5.6   | 471   | 63    |
| St Andrews            | 93    | 5.6   | 13    | 68    | 7.9   | 445   | 69    |
| Imperial College      | 81    | 9.3   | 10    | 86    | 4.3   | 473   | 50    |
| UCL                   | 84    | 8.0   | 9     | 78    | 6.0   | 437   | 59    |
| York                  | 86    | 6.9   | 13    | 66    | 5.4   | 426   | 57    |
| King's College London | 86    | 7.0   | 12    | 77    | 6.1   | 406   | 56    |
| Birmingham            | 85    | 7.2   | 15    | 67    | 5.3   | 399   | 57    |
| :                     |       |       |       |       |       |       |       |

Source: Guardian University Guide 2009

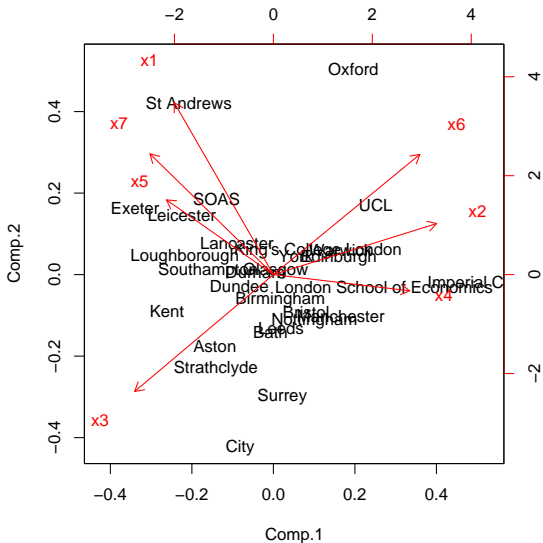
## Example 3 - University Performance Indicators

R

```
> d = read.csv("univleague.csv")  
> pc1=princomp(d)  
> screeplot(pc1)  
> pc2=princomp(d,cor=TRUE)  
> screeplot(pc2)  
> biplot(pc2)
```



# Example 3 - University Performance Indicators



- Motivation
- Principal Component Analysis (PCA)
- Factor Analysis

# Factor Analysis

Initially have  $n$  observation, and each observation consists of  $p$  variables:  $x_1, x_2, \dots, x_p$ . (Without loss of generality, assume  $\bar{x}_i = 0$ .)

Factor analysis finds  $m$  new variables,  $f_1, f_2, \dots, f_m$ , so that:

$$x_j = b_{j,1}f_1 + b_{j,2}f_2 + \dots + b_{j,m}f_m + e_j$$

A number of constraints are applied as to the distributions and independence of  $f_i$  and  $e_j$ .

$f_i$  are common variables

$b_{j,i}$  are factor loadings

$e_j$  are unique (or specific) variables

# Example

R

```
> d = read.csv("univleague.csv")
> fs1 = factanal(d,1)
> fs1
```

Call:

```
factanal(x = d, factors = 1)
```

Uniquenesses:

```
   x1    x2    x3    x4    x5    x6    x7
0.991 0.452 0.450 0.790 0.929 0.410 0.923
```

Loadings:

```
   Factor1
x1
x2 0.740
x3 -0.742
x4 0.458
x5 -0.267
x6 0.768
x7 -0.278
```

```
                Factor1
SS loadings      2.055
Proportion Var   0.294
```

Test of the hypothesis that 1 factor is sufficient.  
The chi square statistic is 35.82 on 14 degrees of freedom.  
The p-value is 0.00111

# Example

R

```
> fs2 = factanal(d,2)
> fs2
```

Call:

```
factanal(x = d, factors = 2)
```

Uniquenesses:

|  | x1    | x2    | x3    | x4    | x5    | x6    | x7    |
|--|-------|-------|-------|-------|-------|-------|-------|
|  | 0.005 | 0.512 | 0.483 | 0.602 | 0.808 | 0.211 | 0.535 |

Loadings:

|    | Factor1 | Factor2 |
|----|---------|---------|
| x1 | 0.166   | 0.984   |
| x2 | 0.620   | -0.321  |
| x3 | -0.719  |         |
| x4 | 0.438   | -0.454  |
| x5 | -0.141  | 0.415   |
| x6 | 0.888   |         |
| x7 | -0.109  | 0.673   |

|                | Factor1 | Factor2 |
|----------------|---------|---------|
| SS loadings    | 1.942   | 1.902   |
| Proportion Var | 0.277   | 0.272   |
| Cumulative Var | 0.277   | 0.549   |

Test of the hypothesis that 2 factors are sufficient.  
The chi square statistic is 8.57 on 8 degrees of freedom.  
The p-value is 0.38

# Example

R

```
> fs3 = factanal(d,3)
> fs3
```

Call:

```
factanal(x = d, factors = 3)
```

Uniquenesses:

|  | x1    | x2    | x3    | x4    | x5    | x6    | x7    |
|--|-------|-------|-------|-------|-------|-------|-------|
|  | 0.005 | 0.382 | 0.332 | 0.005 | 0.791 | 0.286 | 0.509 |

Loadings:

|    | Factor1 | Factor2 | Factor3 |
|----|---------|---------|---------|
| x1 | 0.199   | 0.959   | -0.188  |
| x2 | 0.699   | -0.357  |         |
| x3 | -0.816  |         |         |
| x4 | 0.191   | -0.244  | 0.948   |
| x5 | -0.163  | 0.421   |         |
| x6 | 0.747   |         | 0.389   |
| x7 | -0.127  | 0.687   |         |

|                | Factor1 | Factor2 | Factor3 |
|----------------|---------|---------|---------|
| SS loadings    | 1.831   | 1.761   | 1.097   |
| Proportion Var | 0.262   | 0.252   | 0.157   |
| Cumulative Var | 0.262   | 0.513   | 0.670   |

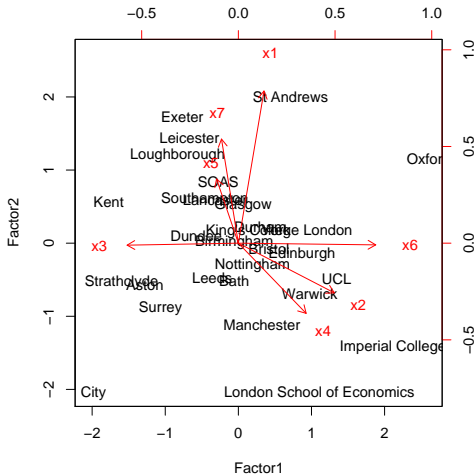
Test of the hypothesis that 3 factors are sufficient.  
The chi square statistic is 1.04 on 3 degrees of freedom.  
The p-value is 0.793



# Example

R

```
> fs2 = factanal(d,2,scores="Bartlett")  
> biplot(x=fs2$scores,y=fs2$loadings)
```



Part VI

Resampling

A statistic calculated from a sample is a random variable. In addition to its value, we often want to know something about its distribution:

- To specify confidence intervals, we may want to know its variance
- For hypothesis tests, we may want to understand the entire distribution

Techniques we can use for this purpose include:

- Theory to derive actual distribution: e.g. Binomial distribution for number of heads when flipping a coin; F distribution for ANOVA statistic
- Assumptions or *a priori* knowledge: e.g. coin is unbiased; distribution is Poisson
- Asymptotic approximations: distribution of a sample mean has normal distribution when the sample size is large
- [Bootstrapping](#)

# General Principle

We want to estimate a population parameter  $\theta$  (e.g. the mean  $\mu$ ) and use an appropriate estimator,  $t$ , calculated from the sample (e.g. the sample mean,  $\bar{x}$ ). Assume the sample is  $x_1, x_2, \dots, x_n$  (size  $n$ ).

- 1 From the original sample, draw a new sample of size  $n$ ,  $x_1^*, x_2^*, \dots, x_n^*$  by picking randomly **with replacement**
- 2 Calculate  $t^*$ , the value of the estimator for this new sample
- 3 Repeat  $m$  times, to give  $t_1^*, t_2^*, \dots, t_m^*$
- 4 The set  $t_1^*, t_2^*, \dots, t_m^*$  gives information about the distribution of the estimator,  $t$ , e.g.:

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

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

## Simple Example

### Original Sample

0.8 0.5 5.3 7.8 9.3 1.3 5.7 4.7 0.1 3.4

$$\bar{x} = 3.89$$

### Bootstrap Samples

| j | bootstrap sample                        | $\bar{x}_j^*$ |
|---|---|---------------|
| 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          |

$$\text{sample mean of } \bar{x}^*: \frac{1}{5} \sum_{i=1}^5 \bar{x}_j^* = 3.18$$

$$\text{sample variance of } \bar{x}^*: \frac{1}{4} \sum_{i=1}^5 (\bar{x}_j^* - 3.18)^2 = 0.702$$

## Example - Distribution of Correlation Coefficient

### Sample Data (Size 50)

| X       | Y       |
|---------|---------|
| -0.8960 | 0.9743  |
| 0.1352  | 0.5965  |
| -0.1390 | 1.3114  |
| -1.1634 | 0.5532  |
| 1.1837  | -0.8961 |
| ⋮       | ⋮       |

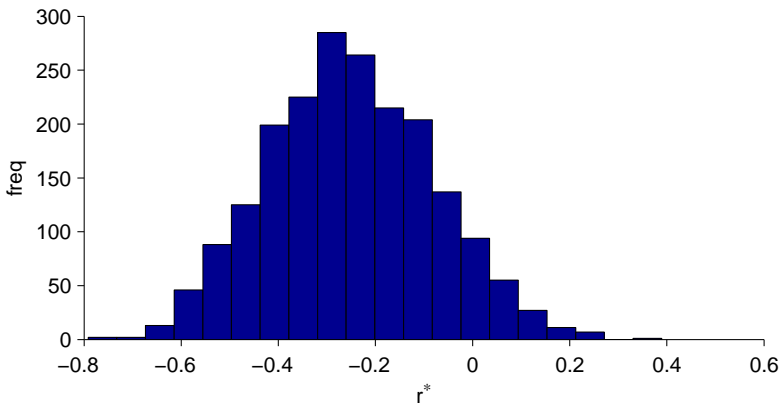
$$r = -0.251$$

Took 2000 bootstrap samples and calculated correlation coefficient,  $r^*$ , for each.

$$\text{sample mean of } r^* = -0.251$$

$$\text{sample variance of } r^* = 0.0288$$

# Example - Distribution of Correlation Coefficient





# 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

## Part VII

### Resources

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

- book** Paul G Hoel  
Introduction to Mathematical Statistics (5th Ed)  
Wiley, 1984
- book** Bryan F J Manly  
Multivariate Statistical Methods - A Primer (3rd Ed)  
Chapman and Hall/CRC, 2002
- book** M R Spiegel, J J Schiller, R A Srinivasan  
Schaum's Outlines: Probability and Statistics (2nd Ed)  
McGraw-Hill, 2000
- book** M Berthold, D Hand (eds)  
Intelligent Data Analysis (2nd Ed)  
Springer, 2003

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

**manual** MATLAB documentation

**manual** R documentation

## Resources III

- 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
- 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

paper J Cohen

The Earth Is Round ( $p < .05$ )

American Psychologist, 49 (12), 997–1003, 1994



- 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

- 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

## Part VIII

Extra Content

- Power Function
- Parametric Hypothesis Tests
- Non-Parametric Hypothesis Tests

# Power Function

If  $H_1$  is composite (e.g.  $p < 0.7$ ), then calculating  $\beta$  is much more difficult since it depends on the (unknown) and actual value of  $p$ .

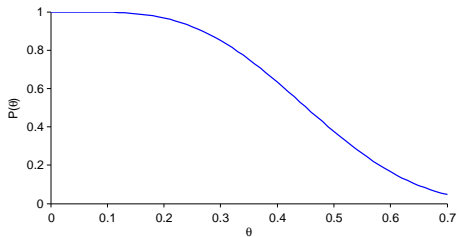
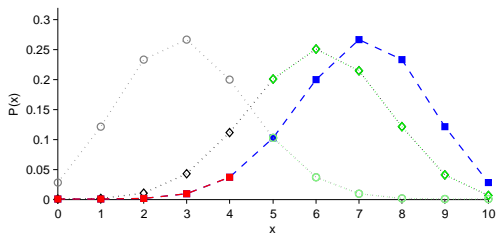
However, if denote the (unknown) actual value of  $p$  as  $\theta$ , then we can calculate  $\beta$  as a function of  $\theta$ .

## Definition (Power Function)

The power function of a test is:  $P(\theta) = 1 - \beta(\theta)$

To minimise the chance of a type II error, we must increase the power function of a test.

# Power Function



- Power Function
- Parametric Hypothesis Tests
- Non-Parametric Hypothesis Tests

# Parametric Tests

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



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

- 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,  $\bar{Y}$ ? (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:

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

to test:

## Hypotheses

$H_0$ : mean of  $Y$  is  $\mu$

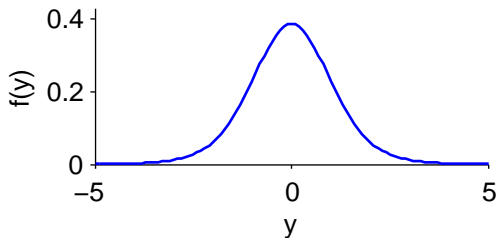
$H_1$ : mean of  $Y$  is not  $\mu$

# Student's $t$ Distribution

If,

- ①  $Y$  has a Normal distribution
- ②  $H_0$  is true (mean of  $Y$  is  $\mu$ )

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  $\mu_A, \mu_B$ ),  
both have **same** variance,  
sample size  $n$  for both

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

**statistic**

$$T = \frac{\bar{Y}_A - \bar{Y}_B}{\sqrt{(S_A^2 + S_B^2)/n}}$$

**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
```

## Parametric Test Assumptions

Important to verify assumptions when applying a parametric test.

- Power Function
- Parametric Hypothesis Tests
- Non-Parametric Hypothesis Tests

# Non-Parametric 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      |
| $\vdots$ | $\vdots$ | $\vdots$ |

# 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

$$H_0: \mu_D = 0$$

$$H_1: \mu_D \neq 0$$

Applied to sample of a single distribution.

## Hypotheses

$$H_0: \text{median}(Y) = \eta$$

$$H_1: \text{median}(Y) > \eta$$

# 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