MATH 204: Principles of Statistics 2

Dr David A. Stephens

Department of Mathematics & Statistics
Room 1235, Burnside Hall

`d.stephens@math.mcgill.ca`
`www.math.mcgill.ca/∼dstephens/204/`

Textbook: McClave and Sincich (2008), *Statistics* (11th Edition), Chapters 10-15.

Note that the 10th Edition of McClave and Sincich contains essentially the same material.

Prerequisites: MATH 203 (or equivalent)

Some statistical computing knowledge useful.

Method of Assessment:

- Assignments
- Mid-Term
- Final

See syllabus handout for precise details.

Course Objectives

- Extensions of MATH 203 topics to other practical experimental contexts
- Introduction to statistical computation using standard software (SPSS)
- Practice in the use of statistical methods, in particular, hypothesis testing and linear modelling.

Three main sections

I. THE ANALYSIS OF VARIANCE AND DESIGNED EXPERIMENTS

II. LINEAR REGRESSION MODELLING

III. NON-PARAMETRIC TESTING

Typical experimental scenario

- two different groups of subjects
- single observation/measurement made on each subject
- scientific question of interest

  ARE THE TWO GROUPS OF SUBJECTS SIGNIFICANTLY DIFFERENT IN TERMS OF THEIR MEASURED VALUES ?

Example: Pre-Natal Care

Objective: To compare the birthweights of babies in two groups of mothers.

- GROUP A: Received five or fewer pre-natal visits
- GROUP B: Received more than five pre-natal visits

Do the GROUP A babies have significantly different birthweights from those from GROUP B ?

Data: Birthweights (grammes)

- GROUP A: 10 subjects

2164	2600	2184	2080	1820
2496	2184	2080	2184	2576

- GROUP B: 7 subjects

3224	2704	2912	2444	3120
2912	3848			

First step in analysis: statistical summary

- GROUP A:
  - Sample size: $n_A = 10$
  - Sample mean: $\overline{x}_A = 2236.8$
  - Sample variance: $s_A^2 = 61190.4$

- GROUP B: 7 subjects
  - Sample size: $n_B = 7$
  - Sample mean: $\overline{x}_B = 3023.429$
  - Sample variance: $s_B^2 = 198679.6$

Recall, for data $x_1, \ldots, x_n$

$$\begin{aligned} \overline{x} &= \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \\ s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \end{aligned}$$

$\overline{x}$ measures the "average" of sample
s^2 measures the amount of variability around the average.

In the birthweight example

$$\overline{x}_A = 2236.8 \qquad \overline{x}_B = 3023.429$$

so it appears that Group B birthweights are higher....

... BUT ARE THEY SIGNIFICANTLY HIGHER ?

i.e. is the difference due to chance alone

- sample sizes quite small
- birthweights quite variable

Statistical Testing

We adopt the following procedure to assess the “significance” of the difference between $\overline{x}_A$ and $\overline{x}_B$.

1. Define a *test statistic*, T, that permits comparison of the two groups
2. *Predict* how T will behave assuming that the two groups are **not** significantly different.
3. Compare the *prediction* with what was actually *observed*.

Formally, we

- assume a **Normal distribution** for the data in the two groups

  i.e. $x_{A1}, \ldots, x_{An_A}$ are drawn from a population of birthweights that is well-modelled by a

  $$Normal(\mu_A, \sigma_A^2)$$

  distribution.

  Similarly

  $$x_{B1}, \ldots, x_{Bn_B} \sim Normal(\mu_B, \sigma_B^2)$$

  We might initially assume that

  $$\sigma_A^2 = \sigma_B^2$$

- consider the two hypotheses

$$\begin{aligned} H_0 :& \quad \mu_A = \mu_B \\ H_a :& \quad \mu_A \neq \mu_B \end{aligned}$$

H_0 is the **NULL HYPOTHESIS**
H_a is the **ALTERNATIVE HYPOTHESIS**

- define the test statistic

$$t = \frac{\overline{x}_A - \overline{x}_B}{s\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

where

$$s^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}$$

s^2 is the estimate of the common population variance

$$\sigma^2 = \sigma_A^2 = \sigma_B^2$$

Here

$$s^2 = \frac{(10-1)61190.4 + (7-1)198679.6}{10+7-2} = 116186.1$$

so that

$$s = 340.8608.$$

Thus

$$t = \frac{2236.8 - 3023.429}{340.8608\sqrt{\frac{1}{10} + \frac{1}{7}}} = -4.683$$

Now, if the null hypothesis was **true**, so that

$$\mu_A = \mu_B$$

the test statistic t should look like an observation from a

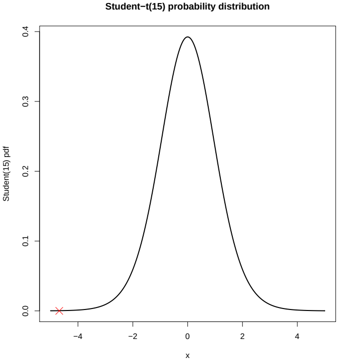
Student-t

distribution with

$$n_A + n_B - 2 = 15$$

"degrees of freedom".

i.e. t should lie somewhere in the “high-probability region” of the Student-t(15) probability distribution



Clearly, in this case, t does not lie in a high probability region.

i.e. we are surprised to see t so far away from zero.

The predicted behaviour of t, under the assumption that H_0 is TRUE, DOES NOT MATCH THE OBSERVED BEHAVIOUR !

Therefore, the assumption that H_0 is true MUST BE INCORRECT and we

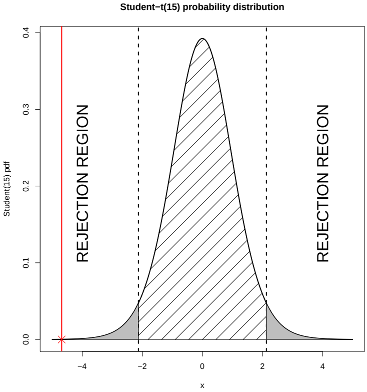
REJECT H_0

How do we quantify the “statistical significance” ?

Two approaches:

1. Define the “high-probability” region, and reject H_0 if t does not lie in this region.
2. Compute the level of “surprise” at observing t under the assumption that H_0 is TRUE.

For 1: Set *significance level* α, with $0 < \alpha < 1$, and find the central $1 - \alpha$ "high-probability" region, between the two values $-C_R$ and C_R (marked by dotted lines).



If $t < -C_R$ or $t > C_R$, REJECT H_0.

Typically, $\alpha = 0.05$ (or 0.01), so for the Student-t(15) distribution

$$C_R = 2.131 \qquad \text{(or 2.947)}$$

The regions $(-\infty, -C_R)$ and (C_R, ∞) form the CRITICAL REGION or REJECTION REGION.

If t lies in the critical region, we reject H_0.

For 2: To compute the level of "surprise", we evaluate the probability of observing a "more extreme" test statistic under the assumption that H_0 is TRUE.

Here, this probability is

$$p = 0.00029.$$

This probability is very small, so we are **very surprised** by the observed result.

p is termed the *p-value* or *observed significance level*.

If $p < \alpha = 0.05$ (or 0.01), we reject H_0.

Some questions:

- How do we choose the test statistic ?
- How do we choose α ?
- Why is the distribution of T (and t) a Student-t(15) distribution ?
- How do we compute C_R and p ?

Equal Variances ?

Is the assumption of equal population variances

$$\sigma_A = \sigma_B$$

fair in this case ?

$$\begin{aligned} s_A^2 &= 61190.4 \\ s_B^2 &= 198679.6 \end{aligned}$$

so that

$$\frac{s_A^2}{s_B^2} = 0.3080.$$

Can we test $\sigma_A = \sigma_B$ formally ?

Yes:

$$H_0: \quad \sigma_A = \sigma_B$$
$$H_a: \quad \sigma_A \neq \sigma_B$$

Test statistic is

$$F = \frac{s_A^2}{s_B^2} = 0.3080$$

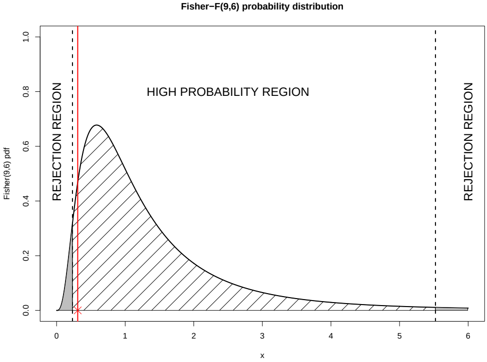
If H_0 is true, F should look like an observation from a

Fisher-F

distribution with

$$(n_A - 1, n_B - 1)$$

"degrees of freedom".



From tables, for $\alpha = 0.05$,

$$C_{R_1} = 0.231 \qquad C_{R_2} = 5.523$$

so the observed value of F **does** lie in the high probability region, and there is no reason to reject H_0 at $\alpha = 0.05$.

Can also compute a 95 % *confidence interval* for $\mu_A - \mu_B$

$$(\overline{x}_A - \overline{x}_B) \pm t_{n_A+n_B-2}(0.975)s\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$$

where

$$t_{n_A+n_B-2}(0.975) = 2.131$$

that is, the 0.975 probability point of the $Student - t(15)$ distribution.

Hence the 95 % confidence interval is

$$(-1144.59, -428.67)$$

- does not contain zero !

NOTE: Significance level α.

$$\begin{aligned} \alpha &= P[H_0 \text{ is rejected, given that } H_0 \text{ is TRUE}] \\ &= P[H_0 \text{ rejected} | H_0 \text{ is TRUE}] \end{aligned}$$

If

- T is the test statistic *random variable*
- $\mathcal{R}$ is the rejection region

then

$$\alpha = P[T \text{ lies in } \mathcal{R} | H_0 \text{ TRUE}] = P[T \in \mathcal{R} | H_0 \text{ TRUE}]$$

that is, α is the probability of committing a

TYPE I ERROR

Part I

Analysis of Variance

In this section

- introduction to the terminology of *designed experiments*
- extension of statistical testing theory to comparison of more than two population means
- THE ANALYSIS OF VARIANCE (ANOVA) F-TEST

1.1 DESIGNED EXPERIMENTS

Data collection studies typically fall into one of two categories:

(i) Observational studies: the experimenter has no control over the variables under study, and can only measure outcomes.

- The IQ of MAC and PC users
- The relationship between environmental exposure to toxins and health status.

i.e. The experimenter does not control the exposure to variables that may cause changes in the outcome of interest.

This type of study is common in medicine and epidemiology as it is relatively cheap to carry out.

Common type of observational study:

CASE-CONTROL STUDY

Example (Smoking and Lung Cancer)

A study (Doll and Hill, 1950) investigated 649 lung cancer cases and 649 matched healthy controls, both drawn from a population of men in the UK. They found out what proportion in each group were smokers.

Neither health status nor smoking status were controlled by the experimenter, but were merely observed.

	Smokers	Non-smokers	Total
Lung cancer	647	2	649
Controls	622	27	649

This type of study can be unreliable, and cannot uncover all the relationships of interest.

A preferred approach involves the experimenter controlling the variables that cause variation in the other variables.

Note that this may not be ethical in a smoking/lung cancer study.

(ii) Designed experiments: the experimenter can the levels of variables that may affect the variable of interest.

Example (Birthweight study)

GROUP A : 5 or fewer visits
GROUP B : More than 5 visits.

at the control of

(a) Mothers ⟶ OBSERVATIONAL STUDY
(b) Doctors ⟶ DESIGNED EXPERIMENT

- after each mother is recruited to take part in the study, they are RANDOMLY assigned to either GROUP A or GROUP B. This is termed a

RANDOMIZED EXPERIMENTAL STUDY

This type of study is preferable, but can be more difficult to implement.

Terminology

- **Response variable** (dependent variable): the variable of interest in the study
- **Factors** : the variables that may have an effect of the response variable
  - quantitative if measured on a numerical scale
  - qualitative otherwise
- **Factor Levels**: the values of the factors utilized in the experiment
- **Treatments**: the factor-level combinations utilized.
- **Experimental Units** (subjects): the objects on which the factors are measured or observed.

Therefore:

- A *designed experiment* is one for which the analyst or experimenter **controls** the specification of treatments and the method of assigning units to treatments.

- An *observational experiment* or study is one for which the analyst simply **observes** the treatments and response on a sample of experimental units.

Example (Birthweight study)

- **Response:** Birthweight (g)
- **Factor:** Pre-natal treatment group
- **Factor levels:** GROUP A or GROUP B

that is, we have a single factor with two factor levels.

Example (SAT scores)

The SAT scores of female and male students in four schools are to be compared.

- **Response:** SAT score
- **Factors:** SEX and SCHOOL (both qualitative)
- **Factor levels:**
  - SEX: Female and Male
  - SCHOOL: A,B,C,D

that is, we have a two factors, SEX with two factor levels and SCHOOL with four factor levels. There are 8 possible treatments:

$$(F, A), (F, B), (F, C), (F, D), (M, A), (M, B), (M, C), (M, D)$$

Example (Pain Relief)

Different pain relief remedies are to be compared : factors are

- REMEDY (quantitative/qualitative, 3 levels)
    - Dose level 0
    - Dose level 1
    - Dose level 2
- AGE GROUP (quantitative/qualitative, 4 levels)
    - 0-16 years
    - 17-40 years
    - 41-65 years
    - 66 years and over
- SEX (qualitative, 2 levels)
    - Female
    - Male

A total of $3 \times 4 \times 2 = 24$ possible treatment combinations; REMEDY is the only factor that can be assigned by the analyst.

Completely Randomized Design

A *completely randomized design* (CRD) is a design for which treatments are randomly assigned to experimental units, or in which random samples of experimental units are selected for each treatment.

The term can be applied to both experimental and observational studies. For example,

- if the treatments are FEMALE/MALE for the factor SEX, a CRD draws independent samples of FEMALES and MALES for the two treatment groups.
- if the treatments are DOSE 0/DOSE 1, a CRD assigns experimental units independently to the two treatment groups at random.

Statistical Objectives

The experimental units assigned to different treatments (factor-level combinations) form

independent samples

from

different populations

in a CRD.

We wish to **compare** treatments: specifically, we wish to compare the treatment means.

A Multiple Group Comparison of Means !

Suppose that there are k treatments:

$$\begin{array}{ll} \text{TMT } 1 & \text{Mean } \mu_1 \\ \text{TMT } 2 & \text{Mean } \mu_2 \\ \vdots & \vdots \\ \text{TMT } k & \text{Mean } \mu_k \end{array}$$

We wish to test the hypotheses

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_a &: \text{At least two of the } k \text{ treatment means are different} \end{aligned}$$

How do we do this ?

What is the relevant test statistic ?

Comparing k Treatments

Suppose

$$\begin{array}{ll} \text{TMT } 1 & \text{has } n_1 \text{ experimental units} \\ \text{TMT } 2 & \text{has } n_2 \text{ experimental units} \\ \vdots & \qquad\qquad \vdots \\ \text{TMT } k & \text{has } n_k \text{ experimental units} \end{array}$$

Denote by x_{ij} the response for unit j in treatment group i, for $j = 1, \ldots, n_i$ and $i = 1, \ldots, k$.

Let

$$\overline{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

denote the sample mean for treatment i, and

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2$$

denote the sample variance for treatment i.

Now we consider pooling, that is, combining all units into a single group.

Define

- the total sample

$$n = n_1 + \cdots + n_k = \sum_{i=1}^{k} n_i$$

- the overall sample mean

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$$

- the overall sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2$$

Finally, consider the pooled sample variance

$$s_P^2 = \frac{1}{n-k}\sum_{i=1}^{k}(n_i - 1)s_i^2$$

- the extension of the pooled estimate of the population variance in a two-sample *t*-test.

Using these quantities, we can derive a test statistic for multiple group comparison.

We wish to compare how much variation is due to the

A DIFFERENCE BETWEEN TREATMENTS

and how much is due to

B RANDOM VARIATION WITHIN TREATMENTS

We measure A using the statistic

$$SST = \sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2$$

SST - Sum of Squares for Treatments

We measure B using the statistic

$$\begin{aligned} SSE &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2 \\ &= (n-k)s_P^2 \end{aligned}$$

SSE - Sum of Squares for Error

NOTE: This measure of random or error variability implicitly assumes that the variability **within** the treatment groups is the **same for each group**. That is, population variances

$$\sigma_1^2, \ldots, \sigma_k^2$$

are equal.

In practice this assumption must be checked.

Finally, we define the test statistic using the mean levels of variability

- MST - Mean Square for Treatments

$$MST = \frac{SST}{k-1} = \frac{1}{k-1} \sum_{i=1}^{k} n_i (\overline{x}_i - \overline{x})^2$$

- MSE - Mean Square for Error

$$MSE = \frac{SSE}{n-k} = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2 = s_P^2$$

Then the test statistic is

$$F = \frac{MST}{MSE} = \frac{\text{Average Variation due to Treatments}}{\text{Average Variation due to Errors}}$$

F large $\Longrightarrow$ Treatments Different !
F small $\Longrightarrow$ Treatments Similar !

The behaviour of F is given by the following Theorem

Theorem (ANOVA F-test to compare k treatments in a Completely Randomized Design)

To test the hypothesis of **equal treatment means**,

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_a &: \textit{At least two of the } k \textit{ treatment means are different} \end{aligned}$$

the test statistic is

$$F = \frac{MST}{MSE}$$

If H_0 is **TRUE**, *then*

$$F \sim \textit{Fisher-F}(k-1, n-k)$$

and the rejection region for a test at significance level α is the region to the right of the $1-\alpha$ probability point of this Fisher-F distribution, C_R.

NOTE: If

$$SS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2$$

is the *overall* or *total* sum of squares, then

$$SS = SST + SSE$$

so we can decompose the overall variation (SS) into the variation due to treatments (SST) and the variation due to the errors (SSE).

Assumptions behind the ANOVA F-test

1. The samples are randomly selected in an independent manner from the k treatment populations.
   [Satisfied in a CRD]
2. All k populations have distributions that are approximately normal.
3. The k population variances are equal.

$$\sigma_1^2 = \sigma_2^2 = \cdots \sigma_k^2.$$

Example (Milk Quality Data)

The impact on milk protein level of three different diets is being studied.

Data: Measurements of milk protein levels for $n = 1337$ samples.

- **Response:** Milk Protein Level (%)
- **Factor:** DIET
- **Factor levels:** $k = 3$
    - 1: Barley
    - 2: Barley + Lupins
    - 3: Lupins

	TMT 1	TMT 2	TMT 3
n_i	425	459	453
$\overline{x}i$	3.532	3.430	2.312
s_i^2	0.102	0.091	0.114

$$\begin{aligned} SST &= 10.606 \\ SSE &= 136.432 \\ SS &= 147.038 \end{aligned}$$

$$\begin{aligned} k-1 &= 2 \\ n-k &= 1334 \end{aligned}$$

Therefore

$$
\begin{aligned}
MST &= \frac{SST}{k-1} = \frac{10.606}{2} = 5.303 \\
MSE &= \frac{SSE}{n-k} = \frac{136.432}{1334} = 0.102
\end{aligned}
$$

and

$$F = \frac{MST}{MSE} = 51.851$$

If H_0 is true, that is,

$$\mu_1 = \mu_2 = \mu_3$$

then F should look like an observation from a

$$\text{Fisher-F}(k-1, n-k)$$

distribution.

Here we are dealing with the

$$\text{Fisher-F}(2, 1334)$$

distribution. From tables, we discover that if $\alpha = 0.05$, then

$$F_\alpha(2, 1334) = 3.002$$

and thus we

Reject H_0

and conclude that there is a significant impact on milk protein level due to diet.

Note: Tables in McClave and Sincich only give

$$\begin{aligned} F_{0.05}(2,120) &= 3.07 \\ F_{0.05}(2,\infty) &= 3.00 \end{aligned}$$

so we cannot look up $F_{0.05}(2,1334)$. However, we know that

$$3.00 < F_{0.05}(2,1334) < 3.07$$

and here the test statistic is $F = 51.851$.

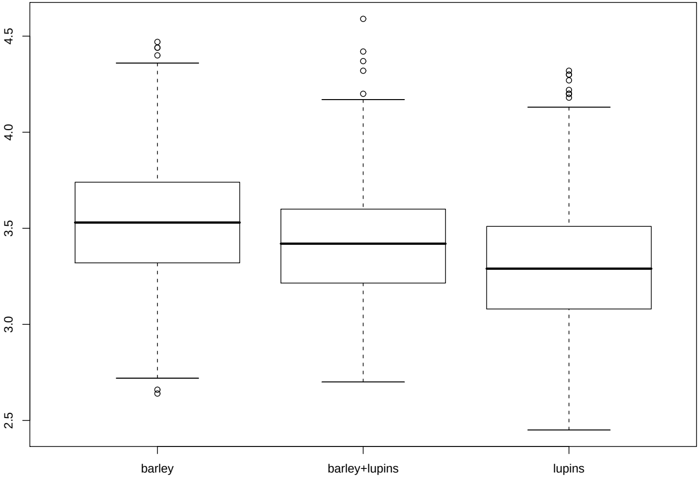
Are the assumptions met ?

1. **Independent samples** : Not possible to tell with current information. In fact, data comprise repeated measurements on 79 cows - potentially not independent, as observations on the same cow are likely to be more similar.
2. **Normal Distributions** : Visual inspection of boxplots indicates that this may be valid.
3. **Equal variances** :

   $$s_1^2 = 0.102 \qquad s_2^2 = 0.091 \qquad s_3^2 = 0.114$$

   so assumption appears to be valid
   - can we test this formally ?

Milk Data: 3 Treatments



Example (Anxiety Response Treatment)

In a study of Alzheimer's disease and care of its sufferers, a medication designed to improve anxiety relief has been developed.

In a lab experiment, $n = 20$ rats were assigned to one of four ($k = 4$) treatment groups corresponding to dose-level of the medication.

A measure of response to a "flee stimulus" was recorded.

- **Response:** Pull response to stimulus (units of force)
- **Factor:** DOSE-LEVEL
- **Factor levels:** $k = 4$
    - Dose 0 (zero units)
    - Dose 1 (one unit)
    - Dose 2 (two units)
    - Dose 3 (three units)

0	1	2	2
27.0	22.8	21.9	23.5
26.2	23.1	23.4	19.6
28.8	27.7	20.1	23.7
33.5	27.6	27.8	20.8
28.8	24.0	19.3	23.9

We find that

$$SST = 140.094 \qquad SSE = 116.324 \qquad SS = 256.418$$

$$MST = 46.698 \qquad MSE = 7.270$$

and

$$F = 6.423$$

which we need to compare with the Fisher-F$(3, 16)$ distribution.

For $\alpha = 0.05$, from McClave and Sincich tables

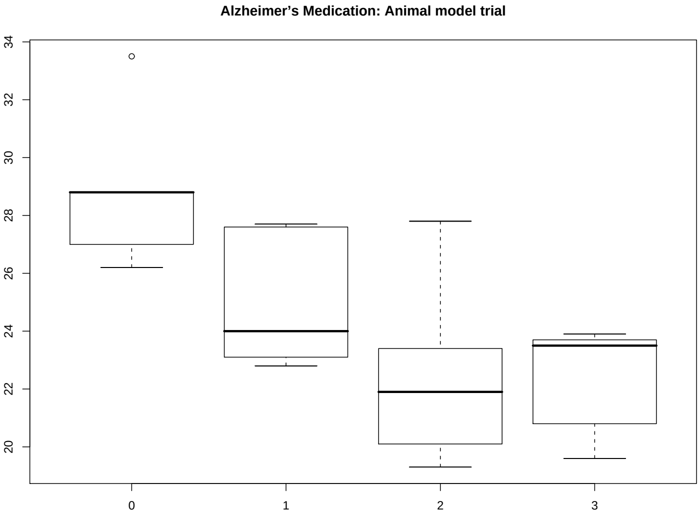
$$F_{0.05}(3, 16) = 3.24$$

and so we

Reject H_0

at $\alpha = 0.05$ and conclude that there is a significant difference between treatment groups.

p-value is 0.0046.


Alzheimer's Medication: Animal model trial
34
32
30
28
26
24
22
20
0
1
2
3

Note: Here

	DOSE 0	DOSE 1	DOSE 2	DOSE 3
s_i^2	8.018	5.873	11.315	3.875

so we might suspect that the treatment variances $\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2$ are not equal. We may test this formally using

LEVENE'S TEST

- SPSS can report this test result.

Note: Visual inspection can give an idea of whether the equal variance assumption is valid, or whether the populations are normal. **But the sample sizes may be small, so that visual inspection or testing may not detect deviations from these assumptions**.

Ideally we would like to be able to relax these assumptions.

The ANOVA Table

For a completely randomized design, we may report the results of the ANOVA F-test in a stylized form, the **ANOVA Table**

SOURCE	DF	SS	MS	F
TREATMENTS	$k-1$	SST	$MST = \frac{SST}{(k-1)}$	$F = \frac{MST}{MSE}$
ERROR	$n-k$	SSE	$MSE = \frac{SSE}{(n-k)}$	
TOTAL	$n-1$	SS		

Note

(i) $(k-1)+(n-k)=(n-1)$

(ii) $SST+SSE=SS$

i.e. we can fill in missing values if they are not given.

Sometimes an extra column is added at the right of the table to give the *p*-value of the ANOVA F-test.

SOURCE	DF	SS	MS	F	p
TMT	$k-1$	SST	MST	$F = \frac{MST}{MSE}$	p-val
ERROR	$n-k$	SSE	MSE		
TOTAL	$n-1$	SS			

where p-val solves

$$\frac{MST}{MSE} = F_{p\text{-val}}(k-1, n-k)$$

and $F_\alpha(\nu_1, \nu_2)$ is the $(1-\alpha)$ probability point of the Fisher-F distribution.

SPSS Handout: Examples

- **DIET**: milk-protein level example (p. 1)
- **DOSE-LEVEL**: pull-strength in Alzheimer's example (p. 3)
- **DIAGNOSIS**: (p. 5)
  - RESPONSE: gut permeability of drug mannitol in AIDS/HIV patients
  - FACTOR: AIDS/HIV Status
  - FACTOR LEVELS: $k = 4$
    - AIDS - Full AIDS
    - ARC - AIDS-related conditions
    - HIV+ - HIV positive
    - HIV- - HIV negative

SPSS Handout: Examples

- **BATCH NUMBER**: bacteria level (per mill.) in different batches of milk (p. 7)
  - RESPONSE: Bacteria level count per million
  - FACTOR: Batch number
  - FACTOR LEVELS: 1,2,3,4,5 ($k = 5$)
- **TREATMENT GROUP**: Post-traumatic stress disorder (PTSD) score in different treatment groups(p. 9)
  - RESPONSE: PTSD score
  - FACTOR: Therapeutic treatment method
  - FACTOR LEVELS: $k = 4$
    - SIT - "Stress Innoculation Therapy"
    - RE - "Relive Experience"
    - SC - "Standard Counselling"
    - WL - "Waiting List" (Control)

Levene's Test

To test

$$\begin{aligned} H_0 &= \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 \\ H_1 &= \text{At least one pair of } \sigma^2 \text{ different.} \end{aligned}$$

Test statistic

$$W = \frac{(n-k)}{(k-1)} \frac{SST_Z}{SSE_Z} = \frac{MST_Z}{MSE_Z}$$

where SST_Z and SSE_Z are the usual sums of squares evaluated for the new data z_{ij} where

$$z_{ij} = |x_{ij} - \overline{x}_i|.$$

If H_0 is true

$$W \sim \text{Fisher-F}(k-1, n-k).$$

Example (PTSD Example (see handout))

$n = 45, k = 4.$

F-statistic	$F = 3.046$
Critical Value	$F_{0.05}(3, 41) \doteq 2.84$
	$F_{0.025}(3, 41) \doteq 3.46$
	$F_{0.01}(3, 41) \doteq 4.31$

Tables give $F_\alpha(3, 40)$.

$$\Longrightarrow \text{Reject } H_0 \text{ at } \alpha = 0.05 \quad (p = 0.039).$$

BUT Levene's Test suggests that the assumption of equal variances is **NOT** valid.

Why do we need the three assumptions ?

- independence
- Normality
- equal variances

- so that we can predict (under H_0) that

$$F \sim \text{Fisher-F}(k-1, n-k)$$

and complete the test (compute p-values and the rejection region).

But our hypothesis of interest is

$$H_0 \; : \; \text{No difference between treatments}$$

Under this hypothesis, the treatment labels

SHOULD NOT MATTER !

i.e. we should be able to exchange the labels, and not notice any major difference in the test statistic.

This leads us to consider **permutation** or **randomization** tests.

i.e. we compute the test statistic for all possible relabellings consistent with H_0, retaining the group sample sizes, and use these values to compute the rejection region.

Randomization/Permutation Tests

Suppose that there are N possible relabellings that give rise to test statistics

$$F_1, F_2, \ldots, F_N$$

Then the rejection region for significance level α is the interval to the right of

$$N(1-\alpha)\text{th largest of the values } F_1, F_2, \ldots, F_N$$

and the p-value is

$$\frac{\text{Number of } \mathrm{F}_1, F_2, \ldots, F_N \geq F}{N}$$

where

$$F = \frac{MST}{MSE}$$

is the true test statistic.

If the group sample sizes are $n_1, n_2, \dots, n_k$ then

$$N = \frac{n!}{n_1!n_2!\dots n_k!}$$

where

$$n! = n(n-1)(n-2)\dots 3.2.1$$

("n factorial") - potentially very large.

Example (PTSD Example)

$$k = 4, n = 45 \qquad (n_1 = 14, n_2 = 10, n_3 = 11, n_4 = 10)$$

There are

$$\frac{45!}{14!10!11!10!} = 2.610 \times 10^{24}$$

possible relabellings: a very big number.

We compute $F = \frac{MST}{MSE}$ for each relabelling. For the real data, $F = 3.046$.

Example (PTSD Example (continued))

Using this approach, we compute for $\alpha = 0.05$

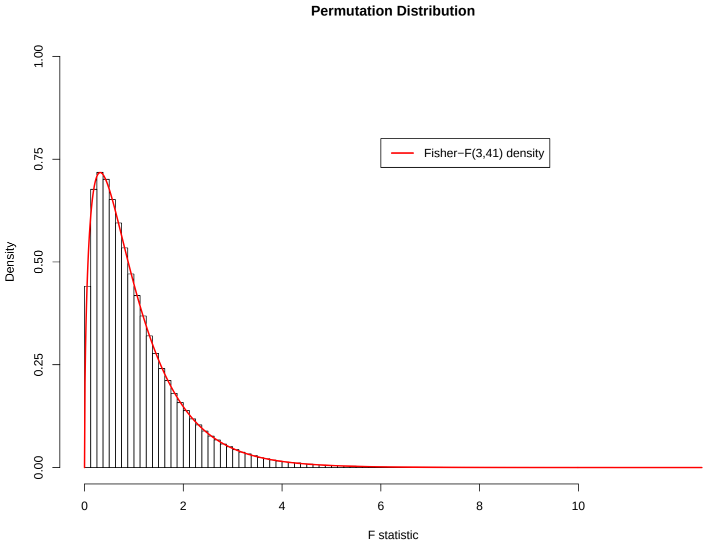
$$\begin{aligned} \text{CRITICAL VALUE} &: \quad C_R = 2.844 \\ p\text{-VALUE} &: \quad p = 0.040 \end{aligned}$$

Compare this with the ANOVA F-test values

$$\begin{aligned} \text{CRITICAL VALUE} &: \quad C_R = 2.833 \\ p\text{-VALUE} &: \quad p = 0.039 \end{aligned}$$

(using the Fisher-F(3,41) distribution.

Thus we obtain virtually identical results; **but the randomization test does not need the assumptions of normality or equal variances.**


Permutation Distribution
Fisher−F(3,41) density
Density
0.00
0.25
0.50
0.75
1.00
0
2
4
6
8
10
F statistic

Example (PTSD Example (continued))

Thus the null hypothesis (of equal means) is

REJECTED

under both procedures at the $\alpha = 0.05$ significance level.

In this case, the computations give similar conclusions. Here the truth or otherwise of the normality/equal variance assumptions **does not matter**.

Final Note on ANOVA F-test for a CRD

If $k = 2$, consider $F = MST/MSE$;

$$\begin{aligned} MST &= \frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2 = n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 \\ &= \frac{n_1 n_2}{n_1 + n_2}(\overline{x}_1 - \overline{x}_2)^2 \end{aligned}$$

$$\begin{aligned} MSE &= \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2 = s_P^2 \\ &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \end{aligned}$$

Therefore

$$F = \frac{\left(\frac{n_1 n_2}{n_1 + n_2}\right)(\overline{x}_1 - \overline{x}_2)^2}{s_P^2} = \left(\frac{(\overline{x}_1 - \overline{x}_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right)^2$$

Thus $F = t^2$, where t is the two-sample t-test statistic.

Thus if $k = 2$, the ANOVA F-test and the two sample t-test are **EQUIVALENT**

$$\begin{aligned} t &\sim \text{Student-t}(n-2) \\ F &\sim \text{Fisher-F}(1, n-2) \end{aligned}$$

and we must get the same conclusion (to reject H_0 or otherwise) using either statistic.

Summary

If the assumptions

- independence (holds by design in a CRD)
- Normal populations
- equal variances

hold, use

ANOVA F-test

If the assumptions **do not hold**

- use Randomization/Permutation test
- use Non-parametric test (see Section 3)

1.3 Multiple Comparison of Means

If the ANOVA F-test null hypothesis

$$H_0 : \mu_1 = \cdots = \mu_k$$

is **rejected**, then it is of interest to discover which of the means are different. For k groups, there are $c = k(k-1)/2$ pairs of group means that can be compared.

Consider a “family” of hypothesis tests - a collection of tests of different hypotheses carried out independently on different data sets. For each test in the family, we consider testing the hypothesis at significance level α.

Notation

Label the tests $i = 1, \ldots, c$, and for each i, label

- the null hypotheses H_{0i}
- the test statistics T_i
- the rejection regions $\mathcal{R}_i$

that are potentially different for each i.

We specify for each i,

$$\alpha = P[T_i \in \mathcal{R}_i | H_{0i} \text{ is } \textbf{TRUE}]$$

which implicitly defines $\mathcal{R}_i$. Note that α is the

"Test Type-I Error Rate" or "Comparisonwise Error Rate"

Now consider the results of all tests in the family; what is the "Familywise" Type-I error rate ?

Using the laws of probability

$$P[T_i \in \mathcal{R}_i | H_{0i} \text{ is } \textbf{TRUE}] = \alpha$$

means that

$$P[T_i \notin \mathcal{R}_i | H_{0i} \text{ is } \textbf{TRUE}] = 1 - \alpha$$

giving the probability that the test **does not reject** H_{0i}, if H_{0i} is in fact true, is $1 - \alpha$.

Now we consider all tests together;

$$P[\text{Each } T_i \notin \mathcal{R}_i | \text{Each } H_{0i} \text{ is } \textbf{TRUE}] = (1 - \alpha)^c$$

This is the probability that each test results in the null hypothesis **not** being rejected, that is, the probability that we **never** commit a Type-I error.

Therefore the probability of **at least one** Type-I error is

$$\alpha_F = 1 - (1 - \alpha)^c$$

α_F is the **Familywise Error Rate**.

	$\alpha = 0.05$	$\alpha = 0.01$
c	α_F	α_F
5	0.226	0.049
10	0.401	0.096
50	0.923	0.395
100	0.994	0.634

Therefore, whenever we carry out a "family" of tests, we should not use the traditional choices of $\alpha = 0.05$ or 0.01 on each test.

The Bonferroni Method

To fix $\alpha_F = 0.05$, say, we need to use α on each test where

$$\alpha_F = 1 - (1-\alpha)^c \quad \Longleftrightarrow \qquad \alpha = 1 - (1-\alpha_F)^{1/c}$$

For example, if $\alpha_F = 0.05$ and $c = 10$, use

$$\alpha = 1 - (1-0.05)^{1/10} = 0.0051$$

It can be shown that

$$1 - (1-\alpha)^c \approx c\alpha$$

Therefore, if α_F is the required **familywise error rate**, we must set the **comparisonwise error** rate to be $\alpha = \alpha_F/c$.

α_F/c is known as the Bonferroni Correction.

Confidence Intervals

For the $k = 2$ group comparison of means, a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\overline{x}_1 - \overline{x}_2) \pm t_{\alpha/2}(n_1 + n_2 - 2)s_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $t_\alpha(\nu_1)$ is the $1-\alpha$ probability point of the Student-t distribution with ν_1 degrees of freedom (under the assumptions of independence, Normality and equal group variances).

If we move to a family of c tests, to get simultaneous confidence intervals for the differences in means $\mu_i - \mu_j$ for all pairs of i and j, we should adjust α to α_F when computing the $100(1-\alpha)\%$ confidence interval.

SPSS gives twelve different methods for correcting the confidence interval for use in different experimental situations. For example

- *planned* comparisons $\mu_1 = \mu_3$, $\mu_7 = \mu_{10}$ etc.
- *all* comparisons

Three methods are recommended:

- **Tukey's** Method
- **Bonferroni's** Method
- **Scheffé's** Method

Having selected a multiple comparison correction method, we compute simultaneous confidence intervals for each comparison of means, and identify

- which means are significantly different
- the ranking of differences $\mu_i - \mu_j$ in terms of magnitude.

1.4 Randomized Block Designs

A **randomized block design** used **matched** experimental units organized into sets known as **blocks** and assigns one member from the set to each treatment.

For k treatments

1. Compile b blocks of k experimental units, with each block comprising units that are similar.
2. Assign one unit from each block to each treatment at random.

Then there are a total of $n = bk$ measured responses.

We wish to **compare treatments** whilst acknowledging that there may be **differences between the blocks**.

That is, the observed variation is due to

TREATMENTS and BLOCKS and ERROR

rather than merely

TREATMENTS and ERROR

as in the CRD.

Example (SAT Scores)

- **Response** : Measured SAT Score
- **Factor** : Sex
- **Factor-levels** : $k = 2$ (Female/Male)
- **Blocks** : $b = 5$ (Previous GPA, within same school)

i.e. $k = 2, b = 5 \therefore n = 10$.

	Block	Female SAT	Male SAT
1	A: 2.75	540	530
2	B: 3.00	570	550
3	C: 3.25	590	580
4	D: 3.50	640	620
5	E: 3.75	690	690

Example (SAT Scores (continued))

This design recognizes that GPA score and school are likely to explain some variation in SAT Score, **but that neither is directly related to the "treatment" of interest (SEX - Female/Male).**

i.e. the **blocking** variable removes systematic variation in response that is not of primary interest.

We pick one Female and one Male in each school/GPA category, and pair them.

Example (Treatment for Hypertension)

- **Response** : Blood Pressure (mgHg)
- **Factor** : Drug Type
- **Factor-levels** : $k = 3$ (Drug 1, Drug 2, Drug 3)
- **Blocks** : $b = 4$ Age/Sex combinations
  - Female/Under 50
  - Male/Under 50
  - Female/Over 50
  - Male/Over 50

i.e. $k = 3, b = 4 \therefore n = 12.$

Testing for Equal Treatment Means

An ANOVA F-test can be constructed for a RBD. Let

- $i = 1, \ldots, k$ index **treatments**
- $j = 1, \ldots, b$ index **blocks**

i.e. x_{ij} is the response for the ith treatment in the jth block. Let

- $\overline{x}_i$ be the ith **treatment** mean
- $\overline{x_j^{(\mathrm{B})}}$ be the jth **block** mean
- $\overline{x}$ be the **overall** mean

Let

$$\begin{aligned} SST &= \sum_{i=1}^{k} b(\overline{x}_i - \overline{x})^2 \\ SSB &= \sum_{j=1}^{b} k(\overline{x_j^{(\mathrm{B})}} - \overline{x})^2 \\ SS &= \sum_{i=1}^{k}\sum_{j=1}^{b}(x_{ij} - \overline{x})^2 \end{aligned}$$

SST: Sum of Squares for **Treatments**
SSB: Sum of Squares for **Blocks**
SS: **Total** Sum of Squares

Finally

$$SS = SST + SSB + SSE \quad \therefore \qquad SSE = SS - SST - SSB$$

SSE: Sum of Squares for **Errors**

Test statistic is

$$F = \frac{MST}{MSE}$$

where

$$MST = \frac{SST}{k-1} \qquad MSE = \frac{SSE}{n-b-k+1}$$

ANOVA F-test to compare treatment means in a randomized block design

Theorem (ANOVA F-test for a RBD)

To test

$$\begin{aligned} H_0 &: \mu_1 = \cdots = \mu_k \\ H_a &: \textit{At least one pair of treatment means different.} \end{aligned}$$

use the test statistic

$$F = \frac{MST}{MSE}$$

If H_0 is **TRUE**

$$F \sim \textit{Fisher-F}(k-1, n-b-k+1)$$

- this defines the rejection region for significance level α, and the p-value, in the usual way.

Assumptions:

1. Experimental units (between blocks) are independent, and treatments are allocated at random (within blocks).
2. Normality
3. bk block/treatment combinations correspond to populations with equal variances.

ANOVA Table

SOURCE	DF	SS	MS	F
TMTS	$k-1$	SST	MST	F = MST/MSE
BLOCKS	$b-1$	SSB	MSB	
ERROR	$n-k-b+1$	SSE	MSE	
TOTAL	$n-1$	SS		

After the ANOVA test is complete, and the hypothesis

$$H_0 : \mu_1 = \cdots = \mu_k$$

is **rejected**, we can proceed with the "post-hoc" tests of hypotheses $\mu_i = \mu_j$ for $i \neq j$.

Notes:

1. In a RBD, it is not (in general) possible to estimate individual treatment means, that is, $\overline{x}_i$ does not estimate μ_i as it is an average across blocks, which are believed to be different in terms of response.

2. **Testing the Block Means** However, we can test whether the block means $\mu_1^{(B)}, \ldots, \mu_b^{(B)}$ are significantly different. For

$$H_0 : \mu_1^{(B)} = \cdots = \mu_b^{(B)}$$

we use the F statistic

$$F = \frac{MSB}{MSE}$$

where

$$MSB = \frac{SSB}{b-1}$$

If H_0 is **TRUE**

$$F \sim \text{Fisher-F}(b-1, n-k-b+1)$$

That is, we treat the blocks as levels of another factor, and test to see whether this factor affects response.

Example (Soil Analysis (see handout))

Results of two ANOVA F-tests:

Test of	F	p	Conclusion
SOLVENT	0.673	0.585	No Difference
SOIL	10.568	0.001	Difference

Here SOLVENT is the **treatment** variable, SOIL is the **blocking** variable.

3. Remember to check the assumptions (independence, normality, equal variances in each treatment/block combination)

   Equal variances may be hard to check as we only have one observation per treatment/block comparison.

Comment: The "sum of squares" decompositions

$$
\begin{aligned}
CRD : SS &= [SST] + SSE \\
RBD : SS &= [SST + SSB] + SSE
\end{aligned}
$$

are both of the form

$$
\begin{array}{ccccc}
\text{TOTAL} & = & \text{SYSTEMATIC} & + & \text{RANDOM} \\
\text{VARIATION} & & \text{VARIATION} & & \text{VARIATION}
\end{array}
$$

"SYSTEMATIC" $\begin{cases} \text{For the CRD:} & SST \\ \text{For the RBD:} & SST + SSB \end{cases}$

"RANDOM" For both: SSE

We have studied the

Randomized **Complete** Block Design

where each block/treatment combination has one experimental unit.

An incomplete design could also be considered, where some block/treatment combinations are omitted. However, this design does not lead to straightforward ANOVA analysis.

1.5 Factorial Experiments

Designs studied so far:

- CRD - one factor
- RBD - one factor, plus one blocking variable, so two factors in total, where one (the blocking variable) is a known source of systematic variation.

However, in the RBD, we must assume that the treatments behave in a similar way across blocks.

Let i index treatments ($1 \leq i \leq k$) and consider block j, and two treatment (factor levels) i_1 and i_2.

In an RBD, we assume that

$$E[X_{i_1 j} - X_{i_2 j}] = \mu_{i_1} - \mu_{i_2}$$

which does **NOT** depend on j.

That is, the expected difference in response due to the two treatments does not depend on the block.

But perhaps the difference **does** depend on block; perhaps we have **INTERACTION**.

In the current RBD, we do not have enough data to look for this. We now seek to extend the RBD to allow for tests for interaction; we do this by using **replication**.

RBD with Balanced Replication

Suppose we have r observations per block/treatment combination (termed *replicates*), so that we have $n = bkr$ experimental units in total.

Balanced designs have equal numbers of replicates in each block/treatment combination.

In this design, all the quantities

$$SST, SSB, SSE, SS$$
$$MST, MSB, MSE$$

can be defined, and an ANOVA F-test can be carried out - the only difference is that $n = bkr$.

- Sum of Squares for Treatments (SST)

$$SST = \sum_{i=1}^{k} br(\overline{x}_i - \overline{x})^2$$

- Sum of Squares for Blocks (SSB)

$$SSB = \sum_{j=1}^{b} kr(\overline{x_j^{(\mathrm{B})}} - \overline{x})^2$$

- Overall Sum of Squares (SS)

$$SS = \sum_{i=1}^{k} \sum_{j=1}^{b} \sum_{t=1}^{r} (x_{ijt} - \overline{x})^2$$

and $SSE = SS - SST - SSB$

Third index t indexes the replicates.

The RBD with replication does allow the investigation of interaction. The new test is based on the decomposition

$$SS = SST + SSB + SSI + SSE$$

where SSI is the sum of squares for Interaction.

We have SST, SSB and SS as before, and

$$SSI = \sum_{i=1}^{k} \sum_{j=1}^{b} r(\overline{x}_{ij} - \overline{x}_i - \overline{x_j^{(\mathrm{B})}} + \overline{x})^2$$

where

$$\overline{x}_{ij} = \frac{1}{r} \sum_{t=1}^{r} x_{ijt} \qquad i = 1, \ldots, k,\ j = 1, \ldots, b$$

is the sample mean for replicates in (i, j)th treatment/block combination.

Testing in the RBD with Replication

The three F statistics

$$F = \frac{MST}{MSE} \quad F = \frac{MSB}{MSE} \quad F = \frac{MSI}{MSE}$$

can be used to test for significant Treatment, Block and Interaction effects respectively.

Now

$$MSE = \frac{SSE}{\text{Error d.f.}}$$

But what is "Error d.f." ? It is a constant that dictates how large SSE should be on average.

The general rule for computing the error d.f. for any model is

$$\text{Error d.f.} = n - p$$

where n is the total **sample size** and p is the total **number of parameters fitted**.

How many parameters do we fit ?

- No Interaction

  $$p = 1 + (b - 1) + (k - 1)$$

  that is, the overall mean μ, plus the $b - 1$ differences from μ due to the blocks, plus the $k - 1$ differences from μ due to the treatments.

- Interaction

  $$p = bk$$

  that is, one parameter in each cell of the two-way table of blocks by treatments.

Thus

- No Interaction

  $$p = 1 + (b - 1) + (k - 1) = b + k - 1$$

  parameters, so

  $$\text{Error d.f.} = n - p = n - b - k + 1$$

- Interaction: we fit $p = bk$ parameters, so

  $$\text{Error d.f.} = n - p = n - bk$$

It transpires that if

$$MSI = \frac{SSI}{(b-1)(k-1)}$$

is the Mean Square for Interaction, then

$$F = \frac{MSI}{MSE}$$

yields a test statistic suitable for testing interaction. If there is **no interaction**, then

$$F \sim \text{Fisher-F}((b-1)(k-1), n-bk)$$

where $n = bkr$.

Why $(b-1)(k-1)$? This is the number of **extra** parameters we fit to include the interaction.

For the CRD:

$$\begin{array}{rcl} H_a & & H_0 \\ \text{FULL MODEL} & \longrightarrow & \text{NULL MODEL} \\ k \text{ parameters} & \longrightarrow & 1 \text{ parameter} \end{array}$$

so there are $(k-1)$ extra parameters, and SST varies on $(k-1)$ degrees of freedom.

For the RBD: the (i, j)th treatment/block combination has mean

$$\mu_i + \mu_j^B$$

so for testing for a TREATMENT effect

$$\begin{array}{rcl} H_a & & H_0 \\ \text{FULL MODEL} & \longrightarrow & \text{NULL MODEL} \\ k \text{ parameters} & \longrightarrow & 1 \text{ parameter} \end{array}$$

so there are $(k-1)$ extra parameters, and SST varies on $(k-1)$ degrees of freedom.

$$\mu_1, \ldots, \mu_k \longrightarrow \mu$$

For testing for a BLOCK effect

$$\begin{array}{rcl} H_a & & H_0 \\ \text{FULL MODEL} & \longrightarrow & \text{NULL MODEL} \\ b \text{ parameters} & \longrightarrow & 1 \text{ parameter} \end{array}$$

so there are $(b-1)$ extra parameters, and SSB varies on $(b-1)$ degrees of freedom.

$$\mu_1^{(B)}, \ldots, \mu_k^{(B)} \longrightarrow \mu^{(B)}$$

These models and tests can be fitted and carried out even if we do not have replication.

With replication, we can investigate the interaction, that is the model where the (i,j)th treatment/block combination has mean

$$\mu_i + \mu_j^B + \mu_{ij}$$

rather than the model where

$$\mu_i + \mu_j^B$$

that is, we wish to test

$$\begin{aligned} H_0 &: \mu_{ij} = 0 \qquad \text{for all } i \text{ and } j \\ H_a &: \mu_{ij} \neq 0 \end{aligned}$$

In the **full interaction** model: we fit bk parameters

In the **restricted**, **no interaction** model: we fit

$$1 + (b-1) + (k-1) = b + k - 1$$

parameters. Therefore the differences is

$$bk - (b + k - 1) = bk - b - k + 1 = (b-1)(k-1)$$

and SSI varies on $(b-1)(k-1)$ degrees of freedom.

ANOVA Table

SOURCE	DF	SS	MS	F
TMTS	$k-1$	SST	MST	F_T
BLOCKS	$b-1$	SSB	MSB	F_B
INTERACTION	$(b-1)(k-1)$	SSI	MSI	F_I
ERROR	$(n-bk)$	SSE	MSE	
TOTAL	$n-1$	SS		

where

$$MST = \frac{SST}{k-1} \qquad MSB = \frac{SSB}{b-1}$$

$$MSI = \frac{SSI}{(b-1)(k-1)} \qquad MSE = \frac{SSE}{n-bk}$$

and

$$F_T = \frac{MST}{MSE} \qquad F_B = \frac{MSB}{MSE} \qquad F_I = \frac{MSI}{MSE}$$

Example: Batteries Data (see handout)

Dependent Variable: Battery Life

Source	Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	59,154.000	8	7,394.250	11.103	0.000
Intercept	398,792.250	1	398,792.250	598.829	0.000
material	10,633.167	2	5,316.583	7.983	0.002
temp	39,083.167	2	19,541.583	29.344	0.000
material * temp	9,437.667	4	2,359.417	3.543	0.019
Error	17,980.750	27	665.954		
Total	475,927.000	36			
Corrected Total	77,134.750	35			

R Squared = .767 (Adjusted R Squared = .698)

For $\alpha = 0.05$, there is a significant **temp** effect ($p < 0.001$), and a significant **material** effect ($p = 0.002$), and a significant interaction ($p = 0.019$)

NB: If we do not have replication, we CANNOT fit the interaction. Recall that

$$\text{Error d.f.} = n - bk$$

but if $r = 1$, $n = rbk = bk$, so the error d.f. is zero.

In fact, $SSE = 0$ also, so the MSE is not defined.

We now study multifactor designs, to assess the effects and interactions of several factors simultaneously.

We consider all possible combinations of

$$\begin{array}{lll} \text{FACTOR A} & \text{with} & a \text{ levels} \\ \text{FACTOR B} & \text{with} & b \text{ levels} \\ \text{FACTOR C} & \text{with} & c \text{ levels} \\ & \vdots & \end{array}$$

to define the treatments in a factorial design.

Factorial Experiments

A complete factorial experiment is one in which every combination of a number of factors is utilized.

i.e. the number of treatments is equal to the total number of factor-level combinations.

We focus on two factor experiments

FACTOR A with a levels
FACTOR B with b levels

so there are ab treatments in total.

A two-way layout with $a = 3$ and $b = 5$.

Factor B

Factor A

	1	2	3	4	5
1					
2					
3					

This design is very similar to the RBD, but now the second factor is not a blocking factor;

- that is, the *ab* treatment populations are constructed independently from the same base population, or from populations not necessarily believed to be systematically different.
- individuals from the same base population are assigned at random to one of the *ab* treatments.

In this design we can study the effect of Factor A and Factor B (**main effects**) as well as the **interaction** provided we have (balanced) replication.

We construct ANOVA F-tests based on the decomposition

$$SS = SST_A + SST_B + SSI_{AB} + SSE$$

- Sum of Squares for Treatments due to factor A ($\mathrm{SST_A}$)

$$SST_A = \sum_{i=1}^{a} br(\overline{x}_{i.} - \overline{x}_{..})^2$$

- Sum of Squares for Treatments due to factor B ($\mathrm{SST_B}$)

$$SST_B = \sum_{j=1}^{b} ar(\overline{x}_{.j} - \overline{x}_{..})^2$$

- Sum of Squares for Interaction ($\mathrm{SSI_{AB}}$)

$$SSI_{AB} = \sum_{i=1}^{a} \sum_{j=1}^{b} r(\overline{x}_{ij} - \overline{x}_{i.} - \overline{x}_{.j} + \overline{x}_{..})^2$$

New notation:

- sample mean for Factor A level i

$$\overline{x}_{i\cdot} = \frac{1}{br}\sum_{j=1}^{b}\sum_{t=1}^{r} x_{ijt} \qquad i = 1, \ldots, a$$

- sample mean for Factor B level j

$$\overline{x}_{\cdot j} = \frac{1}{ar}\sum_{i=1}^{a}\sum_{t=1}^{r} x_{ijt} \qquad j = 1, \ldots, b$$

- sample mean for replicates in (i, j)th factor combination

$$\overline{x}_{ij} = \frac{1}{r}\sum_{t=1}^{r} x_{ijt} \qquad i = 1, \ldots, a,\ j = 1, \ldots, b$$

- overall sample mean

$$\overline{x}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{t=1}^{r} x_{ijt}$$

These allow computation of

$$SST_A, SST_B, SSI_{AB}, SS, SSE$$

$$MST_A, MST_B, MSI_{AB}, MSE$$

using the degrees of freedom identical to those in the RBD with replication.

Tests for

- significant effect for Factor A
- significant effect for Factor B
- significant interaction

will be carried out as before using an ANOVA table.

ANOVA Table

SOURCE	DF	SS	MS	F
FACTOR A	$a-1$	SST_A	MST_A	F_A
FACTOR B	$b-1$	SST_B	MST_B	F_B
INTERACTION	$(a-1)(b-1)$	SSI_{AB}	MSI_{AB}	F_{AB}
ERROR	$(n-ab)$	SSE	MSE	
TOTAL	$n-1$	SS		

If Factor A is not influential (H_0 specifying no difference between responses at different levels of factor A), then

$$F_A \sim \text{Fisher-F}(a-1, n-ab)$$

Similarly,

$$\begin{aligned} \text{No effect of Factor B} \quad &: \quad F_B \sim \text{Fisher-F}(b-1, n-ab) \\ \text{No Interaction} \quad &: \quad F_{AB} \sim \text{Fisher-F}((a-1)(b-1), n-ab) \end{aligned}$$

SEE EXAMPLES HANDOUT

Note: For two factors A and B, the **main effects plus interaction** model can be written

$$\mathsf{A} + \mathsf{B} + \mathsf{A.B}$$

whereas the **main effects only** can be written

$$\mathsf{A} + \mathsf{B}$$

The models

$$\mathsf{A} + \mathsf{A.B} \qquad\qquad \mathsf{B} + \mathsf{A.B}$$

do not make sense.

For a two factor design, the only models that should be considered and or reported are

MODEL	FACTOR	INTERACTION
NULL	NONE	NONE
A	A	NONE
B	B	NONE
A+B	A,B	NONE
A+B+A.B	A,B	YES

that is, if the interaction is significant, the only model you should report is

$$A + B + A.B$$

Note: ANOVA analysis for the RBD and FD (both with replication) are identical. The only difference lies in the interpretation of the factors

- RBD: one blocking, one treatment factor
- FD: two treatment factors

“Blocking” factors are known or strongly believed to have a significant effect on the response.

Estimating Effect Size

In multifactor designs, parameter estimation can be carried out in different parameterizations

For the CRD (one-way layout):

- Natural parameters: $\mu_1, \ldots, \mu_k$
- Contrast parameters: $\beta, \beta_0, \ldots, \beta_{k-1}$ where

$$\beta = \mu_k \qquad \beta_i = \mu_i - \mu_k, \quad i = 1, \ldots, k-1$$

that is, differences from baseline.

For the two-factor designs (RBD/FD): In the two-way layout, with cells (i, j), $i = 1, \ldots, a$, $j = 1, \ldots, b$. The cell means are m_{ij}, where

$$m_{ij} = \mu_{i.} + \mu_{.j} + \mu_{ij}$$

where $\mu_{i.}$ gives the Factor A contribution, $\mu_{.j}$ gives the Factor B contribution, and μ_{ij} gives the interaction.

The parameterization used by SPSS is the contrast parameterization is

$$
\begin{aligned}
m_{ij} &= \beta_0 && i = a, j = b \\
&= \beta_0 + \beta_i^{(A)} && i = 1, \ldots, a-1, j = b \\
&= \beta_0 + \beta_j^{(B)} && i = a, j = 1, \ldots, b-1 \\
&= \beta_0 + \beta_i^{(A)} + \beta_j^{(B)} + \gamma_{ij}^{(AB)} && \\
& && \begin{array}{r} i = 1, \ldots, a-1 \\ j = 1, \ldots, b-1 \end{array}
\end{aligned}
$$

where

$$
\begin{aligned}
\beta_i^{(A)} &: \text{ contrasts for factor A} \\
\beta_j^{(B)} &: \text{ contrasts for factor B} \\
\gamma_{ij}^{(AB)} &: \text{ interaction}
\end{aligned}
$$

SPSS takes the *a*th level of factor A and the *b*th level of factor B as the baseline, and looks at differences compared to this baseline.

The *ab* parameters are

β_0	1
$\beta_1^{(A)}, \ldots, \beta_{a-1}^{(A)}$	$(a-1)$
$\beta_1^{(B)}, \ldots, \beta_{b-1}^{(B)}$	$(b-1)$
$\gamma_{ij}^{(AB)}, i = 1, \ldots, a-1, j = 1, \ldots, b-1$	$(a-1)(b-1)$
Total	ab

For example: $a = 3, b = 4$.

		Factor B			
		1	2	3	4
Factor A	1	①	②	③	$\beta_0 + \beta_1^{(A)}$
	2	④	⑤	⑥	$\beta_0 + \beta_2^{(A)}$
	3	$\beta_0 + \beta_1^{(B)}$	$\beta_0 + \beta_2^{(B)}$	$\beta_0 + \beta_3^{(B)}$	β_0

where

$$\textcircled{1} \quad = \quad \beta_0 + \beta_1^{(A)} + \beta_1^{(B)} + \gamma_{11}^{(AB)}$$

$$\textcircled{6} \quad = \quad \beta_0 + \beta_2^{(A)} + \beta_3^{(B)} + \gamma_{23}^{(AB)}$$

and so on.

Estimation is still straightforward:

PARAMETER	ESTIMATE
β_0	$\overline{x}_{ab}$
$\beta_i^{(A)}$	$\overline{x}_{i.} - \overline{x}_{ab}$
$\beta_j^{(B)}$	$\overline{x}_{.j} - \overline{x}_{ab}$
$\gamma_{ij}^{(AB)}$	$\overline{x}_{ij} - \overline{x}_{i.} - \overline{x}_{.j} + \overline{x}_{ab}$

for $i = 1, \ldots, a$, $j = 1, \ldots, b$.

Other parameterizations can be used.

Final Note on ANOVA

We have studied the simplest design scenarios: extension to

- *incomplete*
- *unbalanced*
- *nested*
- *random effect*

designs are possible.

Furthermore SPSS has greater functionality: for example, it has the capability to carry out ANOVA-like analyses even for the case of non-equal variances (when Levene's test rejects the hypothesis of equal variances).

Part II

Linear Regression Modelling

2. Linear regression Modelling

In the previous section, we attempted to explain the variation in an observed response variable by fitting models with one or more factors.

Factors are **discrete** variables taking different levels; in this section we will now utilize **continuous** variables that can similarly explain variation in an observed response.

2.1 Simple Linear Regression

We will investigate models relating two quantities x and y through equations of the form

$$y = ax + b$$

where a and b are constants (that is, a straight-line).

Variables x and y will not be treated exchangeably - we will regard y as being a function of x.

Such models are deterministic, that is, if we know x (and the values of the constants), we can compute y exactly without error.

A more useful model allows for the possibility that the system is not observed perfectly, that is, we do not observe (x, y) pairs that are always consistent with a simple functional relationship.

Example (Pharmacokinetic Model)

If a dose of drug is taken at time $x = 0$, the amount (concentration) of drug still in the bloodstream at time x is often well-modelled by a simple equation. Let

- D denote the amount of drug taken at $x = 0$
- x time
- $y^\star$ is the amount (concentration per unit volume) in the bloodstream.

Then

$$y^\star = \frac{D}{V} \exp\{-\lambda x\}$$

where

- λ is the elimination rate
- V is the volume of bloodstream.

Example (Pharmacokinetic Model (continued))

Taking logs of both sides, setting $y = \log y^\star$, then

$$y = -\lambda x + \log(D/V) = -\lambda x + (\log D - \log V)$$

that is, $y = ax + b$ where

- $a = -\lambda$
- $b = (\log D - \log V)$

However, in practice, when we measure concentration, we do so with random error.

2.1.1 Probabilistic Models

In a **probabilistic** model, we allow for the possibility that y is observed with random error, that is,

$$y = ax + b + ERROR$$

where *ERROR* is a random term that is present due to imperfect observation of the system due to (i) measurement error or (ii) missing information.

Note that we do not treat x and y exchangeably; x is a fixed observed variable that is measured *without error*, whereas y is an observed variable that is measured *with random error*.

We model the variation in y as a function of x. We observe pairs $(x_i, y_i), i = 1, \ldots, n$.

A Basic Probabilistic Model

Terminology:

- y - *Dependent variable* or *response variable*
- x - *Independent variable*, or *predictor*, or *covariate*

The model we study takes the form

$$y = \beta_0 + \beta_1 x + \epsilon$$

where ϵ is a random error term, a random variable with mean zero and finite variance ($E[\epsilon] = 0$, $Var[\epsilon] = \sigma^2$); it represents the error present in the measurement of y.

- β_0 - *Intercept* parameter
- β_1 - *Slop*e parameter

- $\beta_1 > 0$ - increasing y with increasing x
- $\beta_1 < 0$ - decreasing y with increasing x
- $\beta_1 = 0$ - no relationship between x and y

Note:

$$E[Y|x] = \beta_0 + \beta_1 x$$

where $E[Y|x]$ is the expected value of Y for fixed value of x.

Recall the notation

- Y - a random variable with a probability distribution
- y - a fixed value that the variable Y can take.

Fundamental Problem: If we believe the straight-line model with error is correct, how do we find the values of parameters β_0 and β_1. We only have the observed data $\{(x_i, y_i), i = 1, \ldots, n\}$.

2.1.2 Least Squares Fitting

We select the best values of β_0 and β_1 by minimizing the *error in fit*. For two data points (x_1, y_1) and (x_2, y_2), the errors in fit are

$$\begin{aligned} e_1 &= y_1 - (\beta_0 + \beta_1 x_1) \\ e_2 &= y_2 - (\beta_0 + \beta_1 x_2) \end{aligned}$$

respectively. But note that, potentially, $e_1 > 0$ and $e_2 < 0$ so there is a possibility that these fitting errors cancel each other out. Therefore we look at **squared** errors (as a large negative error is as bad as a large positive error)

$$\begin{aligned} e_1^2 &= (y_1 - (\beta_0 + \beta_1 x_1))^2 \\ e_2^2 &= (y_2 - (\beta_0 + \beta_1 x_2))^2 \end{aligned}$$

For n data, we obtain n misfit squared errors

$$e_1^2, \ldots, e_n^2$$

We select β_0 and β_1 as the values of the parameters that minimize SSE, where

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

We wish to make the total misfit squared error as small as possible.

SSE - sum of squared errors - is similar to the SSE for ANOVA. We could write

$$SSE = SSE(\beta_0, \beta_1)$$

to show the dependence of SSE on the parameters.

Minimization of $SSE(\beta_0, \beta_1)$ is achieved **analytically**.

Two routes: (i) calculus and (ii) geometric methods. It follows that the best parameters $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are given by

$$\widehat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

where

- Sum of Squares SS_{xx}:

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

- Sum of Squares SS_{xy}:

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

$\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the **least-squares estimates**

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

is the **least-squares line of best fit**. The **fitted-values** are

$$\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \qquad i = 1, \ldots, n$$

and the **residuals** or **residual errors** are

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \qquad i = 1, \ldots, n$$

2.1.3 Model Assumptions for Least-Squares

To utilize least-squares for the probabilistic model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

we make the following assumptions

1. The expected error $E[\epsilon]$ is zero so that

$$E[Y] = \beta_0 + \beta_1 x$$

2. The variance of the error, $Var[\epsilon]$, is constant and does not depend on x.
3. The probability distribution of ϵ is a symmetric distribution about zero (a stronger assumption is that ϵ is Normally distributed).
4. The errors for two different measured responses are independent, i.e. the error ϵ_1 in measuring y_1 at x_1 is independent of the error ϵ_2 in measuring y_2 at x_2.

2.1.4 Parameter Estimation: Estimating σ^2

Using the LS procedure, we can construct an estimate of the *error* or *residual error* variance

Recall that

$$Var[\epsilon] = \sigma^2$$

An estimate of σ^2 is

$$\widehat{\sigma}^2 = \frac{SSE(\widehat{\beta}_0, \widehat{\beta}_1)}{n-2} = s^2$$

say.

Note that the denominator $n-2$ is again a *degrees of freedom* parameter of the form

$$\begin{array}{ccc}\text{TOTAL NUMBER} & - & \text{NUMBER OF PARAMETERS} \\ \text{OF DATA} & & \text{ESTIMATED}\end{array}$$

or $n-p$, where in the simple linear regression, $p=2$ ($\widehat{\beta}_0$ and $\widehat{\beta}_1$). Note also that

$$SSE(\widehat{\beta}_0, \widehat{\beta}_1) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SS_{yy} - \widehat{\beta}_1 SS_{xy}$$

where

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

Estimation and Testing for Slope

In the model where

$$E[Y] = \beta_0 + \beta_1 x$$

it is of interest to test the hypothesis

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_a &: \beta_1 \neq 0 \end{aligned}$$

i.e. H_0 implies that there is no systematic contribution of x to the variation of y.

To test H_0 vs H_a we us the test statistic

$$t = \frac{\widehat{\beta}_1}{\text{e.s.e}(\widehat{\beta}_1)} = \frac{\widehat{\beta}_1}{s_{\widehat{\beta}_1}}$$

where $\text{e.s.e}(\widehat{\beta}_1)$ is the *Estimated Standard Error* of $\widehat{\beta}_1$, computed as

$$\text{e.s.e}(\widehat{\beta}_1) = \frac{s}{\sqrt{SS_{xx}}}$$

where s is the estimate of σ defined previously.

If H_0 is true, and $\beta_1 = 0$, then

$$t = \frac{\widehat{\beta}_1}{s/\sqrt{SS_{xx}}} \sim \text{Student}(n-2)$$

so we can carry out a significance test at level α in the usual way (use a p-value, or construct the rejection region).

Note: we might also consider a one-sided test, where $H_a : \beta_1 > 0$, say.

- If $H_a : \beta_1 \neq 0$, we use the *two-sided* rejection region, with critical values
$$C_R = \pm t_{n-2}(\alpha/2)$$
- If $H_a : \beta_1 > 0$, we use the *one-sided* rejection region, with critical value
$$C_R = +t_{n-2}(\alpha)$$
- If $H_a : \beta_1 < 0$, we use the *one-sided* rejection region, with critical value
$$C_R = -t_{n-2}(\alpha)$$

Note: To test

$$\begin{aligned} H_0 &: \beta_1 = b \\ H_a &: \beta_1 \neq b \end{aligned}$$

for any b, the test statistic is

$$t = \frac{\widehat{\beta}_1 - b}{s/\sqrt{SS_{xx}}}$$

(for example, $b = 1$ may be of interest. If H_0 is true

$$t \sim \text{Student}(n-2)$$

Confidence Interval

A $100(1-\alpha)\%$ confidence interval for β_1 is

$$\widehat{\beta}_1 \pm t_{n-2}(\alpha/2) \times s_{\widehat{\beta}_1}$$

where

$$\begin{aligned} t_{n-2}(\alpha/2) &: \quad \alpha/2 \text{ prob. point of Student}(n-2) \text{ distn.} \\ s_{\widehat{\beta}_1} &: \quad \text{Estimated standard error of } \widehat{\beta}_1 \end{aligned}$$

Note: we could perform a similar analysis for β_0, but this is generally of less interest.

The only quantity that needs attention is the estimated standard error of $\widehat{\beta}_0$. It can be shown that

$$\text{e.s.e.}(\widehat{\beta}_0) = s_{\widehat{\beta}_0} = \sqrt{\frac{1}{n}\left(1 + \frac{n\overline{x}^2}{SS_{xx}}\right)}$$

2.1.5 The Coefficient of Correlation

To measure the *strength of association* between the two variables x and y we can use the

Pearson Product Moment Coefficient Of Correlation

or *correlation coefficient* which measures the strength of the **linear** relationship between x and y.

The coefficient, r, is defined by

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

where

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 \quad SS_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

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

- If r is close to 1, there is a strong linear relationship between x and y where y **increases** with x.
- If r is close to -1, there is a strong linear relationship between x and y where y **decreases** with x.

Note: In the model

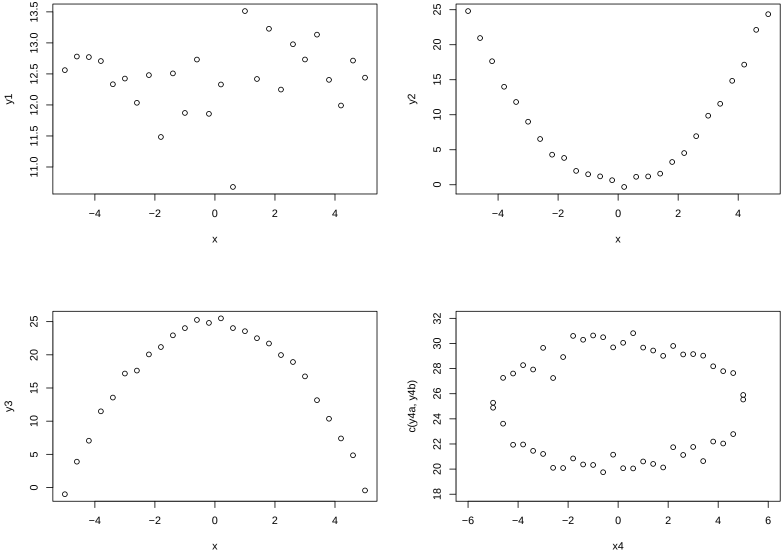
$$y = \beta_0 + \beta_1 x$$

$\beta_1 = 0 \Longrightarrow r \approx 0$, so tests for $\beta_1 = 0$ can also be used to deduce a lack of correlation between the variables.

Notes

1. A strong linear relationship is not necessarily a **causal** relationship, that is, just because $r \approx 1$ does not mean that x **causes** changes in y (we may have a *spurious* correlation).
2. Just because $r \approx 0$ does not mean that that x and y are unrelated, merely that they are **uncorrelated**. That is, it is possible to construct examples where x and y have a strong functional relationship, but where $r = 0$.

Examples where $r \approx 0$.



Testing Correlation

We use ρ to denote the **true** correlation between X and Y.

We can test the hypothesis that $\rho = 0$ (that is, that X and Y are uncorrelated using r. For testing

$$\begin{aligned} H_0 &: \rho = 0 \\ H_a &: \rho \neq 0 \end{aligned}$$

we can use the test statistic

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}}$$

If H_0 is true, then approximately

$$t \sim \text{Student}(n - 2)$$

Alternately, we could use

$$z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$$

and then, if H_0 is true, as (approximately)

$$Z \sim N\left(\frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n-3}\right)$$

when $\rho = 0$, so that (approximately)

$$\sqrt{n-3}\, Z \sim N(0,1)$$

A related quantity is the

Coefficient of Determination

or **$\mathbf{R^2}$ Statistic**

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

Note that the *total variation* in y is recorded via

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

and the *random variation* is recorded via

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Therefore the **variation explained by the linear regression** is

$$SSR = SS_{yy} - SSE \qquad \text{as} \qquad SS_{yy} = SSR + SSE$$

Thus

$$r^2 = \frac{SSR}{SS_{yy}} = \frac{\text{Variation explained by Regression}}{\text{Total Variation}}$$

r^2 is a measure of model adequacy, that is, if $r^2 \approx 1$, then the linear model is a **good fit**.

Example (Blood Viscosity vs PCV)

We have

- $n = 32$
- $r = 0.879$
- $R^2 = r^2 = (0.879)^2 = 0.772$

Test of $\rho = 0$:

$$t = \frac{r}{\sqrt{(1 - r^2)/(n-2)}} = 10.087$$

We compare with a Student$(n-2) \equiv$ Student(30) distribution; the p-value is 3.73×10^{-11}, so there is strong evidence that $\rho \neq 0$.

2.1.6 Prediction

After the linear model is fitted, it can be used for **forecasting** or **prediction**. That is, given a new x value we can predict the corresponding y.

As before, we see that at any value of x_p, the prediction $\hat{y}_p$ is

$$\hat{y}_p = \widehat{\beta}_0 + \widehat{\beta}_1 x_p$$

This is the best predictor of y at this x value.

We can also compute the standard error of this prediction; if the value of the random error variance σ^2 is known, then

$$\text{s.e.}(\hat{y}_p) = \sigma \sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

If σ is unknown, we estimate σ by $\widehat{\sigma} = s$ as defined previously

$$s^2 = \frac{SSE(\widehat{\beta}_0, \widehat{\beta}_1)}{n-2}$$

so that

$$\text{e.s.e.}(\hat{y}_p) = s \sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

Note: This prediction is the expected value of y at $x = x_p$. That is, we have worked out

$$Var[\widehat{Y}_p] = Var[\widehat{\beta}_0 + \widehat{\beta}_1 x_p]$$

to compute the s.e. for $\widehat{Y}_p$.

But we can actually predict an **error corrupted** version of $\widehat{Y}_p$, $\widehat{Y}_p^\star$ say, where

$$\widehat{Y}_p^\star = \widehat{Y}_p + \epsilon_p$$

where ϵ_p is a new random error.

But

$$Var[\widehat{Y}_p^\star] = Var[\widehat{Y}_p + \epsilon_p] = Var[\widehat{Y}_p] + Var[\epsilon_p] = Var[\widehat{Y}_p] + \sigma^2$$

that is, there is an **extra** piece of variation due to ϵ_p.

Thus

$$\text{e.s.e.}(\hat{y}_p^\star) = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}} > \text{e.s.e.}(\hat{y}_p)$$

Prediction Intervals

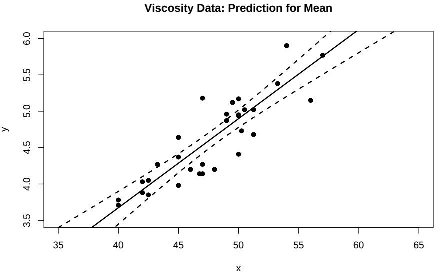
A $100(1-\alpha)\%$ prediction interval for the **mean** value at $x = x_p$ is

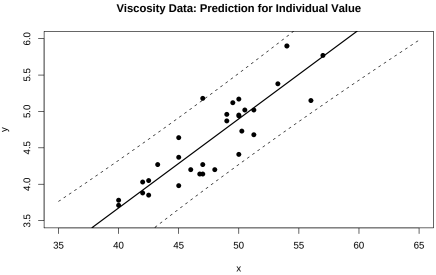
$$\hat{y}_p \pm t_{n-2}(\alpha/2) s \sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

whereas for an individual new value (predicted with error) at $x = x_p$ is

$$\hat{y}_p \pm t_{n-2}(\alpha/2) s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

Prediction Intervals





ANOVA-F test in Regression

An ANOVA-F test can be constructed to test overall (*global*) fit of the linear regression model.

The decomposition of sums of squares for regression takes the form

$$SS = SSR + SSE$$

where

- $SS = SS_{yy}$: overall or total sum of squares
- SSR: sum of squares due to Regression
- SSE: sum of squares due to Error

$$\begin{aligned} SS &= \sum_{i=1}^{n}(y_i - \overline{y})^2 \\ SSR &= \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \\ SSE &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \end{aligned}$$

where

$$\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \qquad i = 1, \ldots, n$$

Degrees of Freedom

- TOTAL: $n - 1$
- REGRESSION: 1
- ERROR: $n - 2$

(error d.f. is $n - p$, here $p = 2$).

The ANOVA Table

SOURCE	DF	SS	MS	F
REGRESSION	1	SSR	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
ERROR	$n-2$	SSE	$MSE = \frac{SSE}{(n-2)}$	
TOTAL	$n-1$	SS		

The test of the hypothesis

$$\begin{aligned} H_0 &: \quad E[Y] = \beta_0 \\ H_a &: \quad E[Y] = \beta_0 + \beta_1 x \end{aligned}$$

can be completed by using the test statistic

$$F = \frac{MSR}{MSE}$$

If H_0 is true

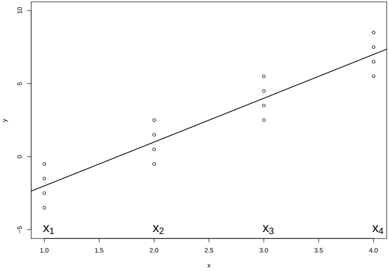
$$F \sim \text{Fisher-F}(1, n-2)$$

This is just like the ANOVA in the one-way layout (CRD) with n groups, but where

$$\mu_i = \beta_0 + \beta_1 x_i$$

That is, the group means are **structured**, that is, we have a formula relating the μ_i quantities.

Consider four replicates at x values (x_1, x_2, x_3, x_4) in a regression;



Then for group i, $\mu_i = \beta_0 + \beta_1 x_i$, $i = 1, 2, 3, 4$.

Checking the Local Fit

A plot of the *residuals*

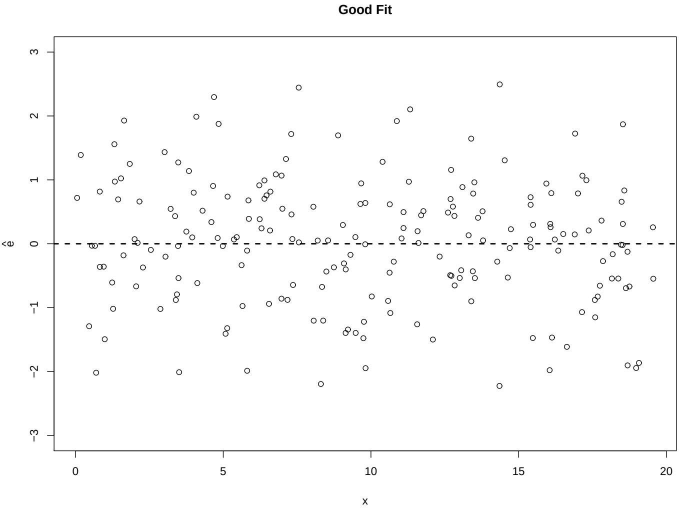
$$\hat{e}_i = y_i - \hat{y}_i$$

can reveal model inadequacies. We should observe that in plots of

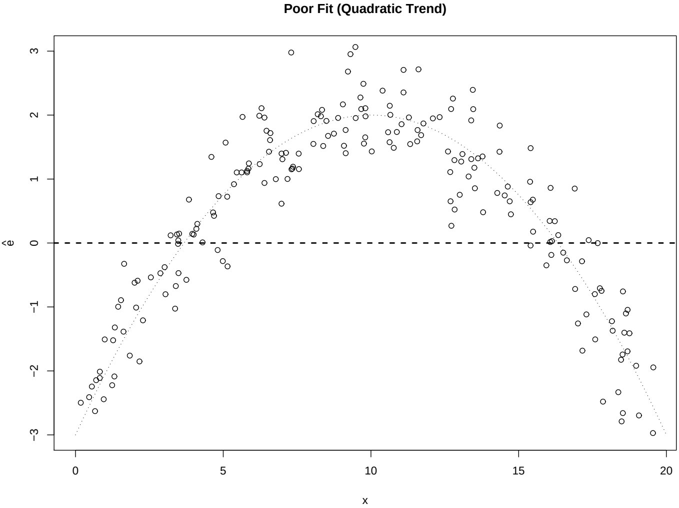
- x vs $\hat{e}$
- y vs $\hat{e}$
- $\hat{y}$ vs $\hat{e}$

there is no discernible pattern

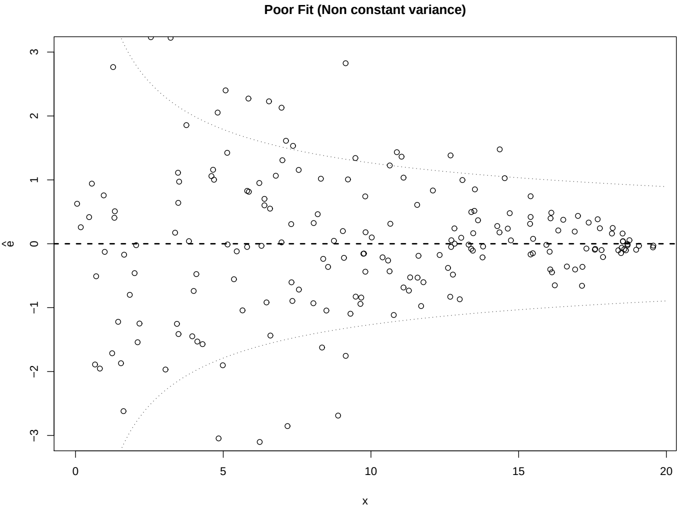
Checking the Local Fit: Good Fit



Checking the Local Fit: Poor Fit



Checking the Local Fit: Poor Fit



R^2 and adjusted R^2

SPSS reports both the R^2 statistic

$$R^2 = 1 - \frac{SSE}{SS}$$

and the **adjusted** R^2 statistic

$$R^2 = 1 - \frac{SSE/EDF}{SS/TDF}$$

where

- EDF = error degrees of freedom $= n - 2$
- TDF = total degrees of freedom $= n - 1$

2.1.7 Polynomial Regression

In many practical situations, the simple straight line

$$y = \beta_0 + \beta_1 x$$

is not appropriate. Instead, a model including powers of x

$$x^2, x^3, \ldots, x^k$$

should be considered. For example

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x^j = \beta_0 + \beta_1 x + \cdots + \beta_k x^k$$

The **Polynomial Regression Model**

$$Y = \beta_0 + \beta_1 x + \cdots + \beta_k x^k + \epsilon$$

where ϵ is a random error term as before can be used to model data.

Two immediate problems:

1. How to choose k
2. How to carry out inference
    - estimation
    - testing
    - prediction

We begin by addressing 2. The estimation of parameters can be again carried out using **Least Squares** provided that the model assumptions listed before are valid. Consider $k = 2$.

We choose $\underset{\sim}{\beta} = (\beta_0, \beta_1, \beta_2)^{\mathsf{T}}$ to minimize the **sum of squared errors**

$$SSE(\underset{\sim}{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2$$

that is the fitted values for parameters $\underset{\sim}{\beta}$ are

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

$\underset{\sim}{\widehat{\beta}}$ can be found to minimize SSE using calculus techniques (differentiating with respect to the elements of $\underset{\sim}{\beta}$) to give the minimum SSE

$$SSE(\underset{\sim}{\beta}) = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i - \widehat{\beta}_2 x_i^2)^2$$

We can also compute the estimated **standard errors**

$$s_{\widehat{\beta}_0}, s_{\widehat{\beta}_1}, s_{\widehat{\beta}_2}$$

which allow tests of parameters to be carried out, and confidence intervals calculated.

We can also compute prediction intervals.

The best estimate of the residual error variance σ^2 is

$$\widehat{\sigma}^2 = \frac{SSE(\underset{\sim}{\widehat{\beta}})}{n-3}$$

p is the number of parameters estimated equal to three, so we divide by $n-3$.

We can also compute

- Residuals
    - can be used to assess the fit of the model.
    - the residuals should be *patternless* if the model fit is good.
- R^2, Adjusted R^2 statistics
    - used to assess the global fit of the model.
    - used to compare the quality of fit with other models.

Example (Hooker Pressure Data)

For the Hooker pressure data, a **quadratic** polynomial ($k = 2$) might be suitable.

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2$$

We need to estimate β_0, β_1 and β_2 for these data to see if the model fits better than the straight line model we fitted previously. This can be achieved using SPSS.

It transpires that the quadratic model produces a set of residuals that are patternless, which the straight line model when fitted does not.

See Handout for full details.

Note: It is common to use the **Standardized Residuals**

$$\widehat{z}_i = \frac{\widehat{e}_i}{\widehat{\sigma}} = \frac{y_i - \hat{y}_i}{\widehat{\sigma}}$$

where $\widehat{\sigma}^2$ is the estimate of σ^2 defined previously, as

$$\text{Var}[\widehat{z}_i] \approx 1$$

if the model fit is good, whereas

$$\text{Var}[\widehat{e}_i] \approx \sigma^2$$

which clearly depends on σ. This makes it hard to compare $\widehat{e}_i$ across different models when inspecting residuals.

Note: Although the model based on

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

is **not** linear in x, it **is** linear in the parameters. Because of this, we still term this a *linear model*. It is this fact that makes the least-squares solutions easy to find.

This model is no more difficult to fit than the model

$$y = \beta_0 + \beta_1 \frac{x}{1+x} + \beta_2 (1 - e^{-x})$$

say - it is still a *linear in the parameters model*. It is in the general class of models

$$y = \beta_0 + \beta_1 g_1(x) + \beta_2 g_2(x)$$

where $g_1(x)$ and $g_2(x)$ are general functions of x.

In fact, any model of the form

$$y = \sum_{j=0}^{k} \beta_j g_j(x) + \epsilon \tag{1}$$

can be fitted routinely using least-squares; if we know x, then we can compute

$$g_0(x), g_1(x), \ldots, g_k(x)$$

and plug those values into the formula (1).

Example (Harmonic Regression)

Let

$$\begin{aligned} g_0(x) &= 1 \\ g_1(x) &= \begin{cases} \cos(\lambda_j x) & j \text{ odd} \\ \sin(\lambda_j x) & j \text{ even} \end{cases} \end{aligned}$$

where k is an even number, $k = 2p$ say.

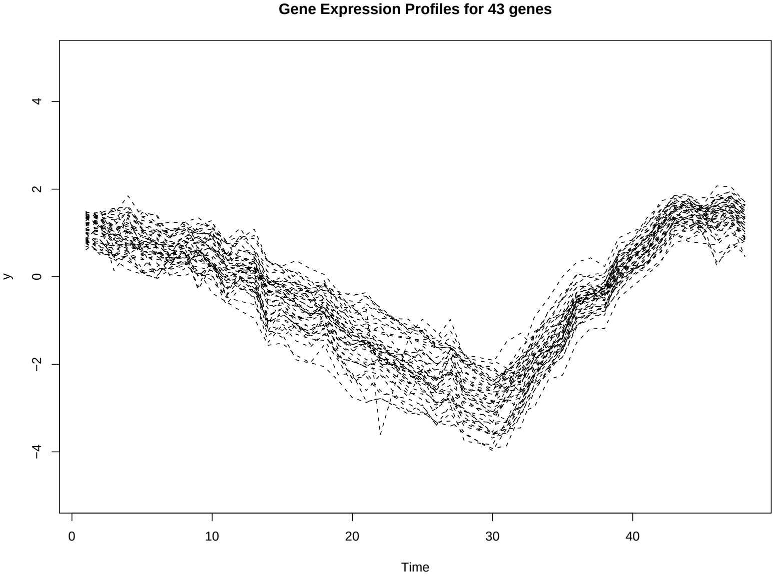
$\lambda_j, j = 1, 2, \ldots, p$ are constants

$$\lambda_1 < \lambda_2 < \cdots < \lambda_p$$

For fixed x, $\cos(\lambda_j x)$ and $\sin(\lambda_j x)$ are also fixed, known values.

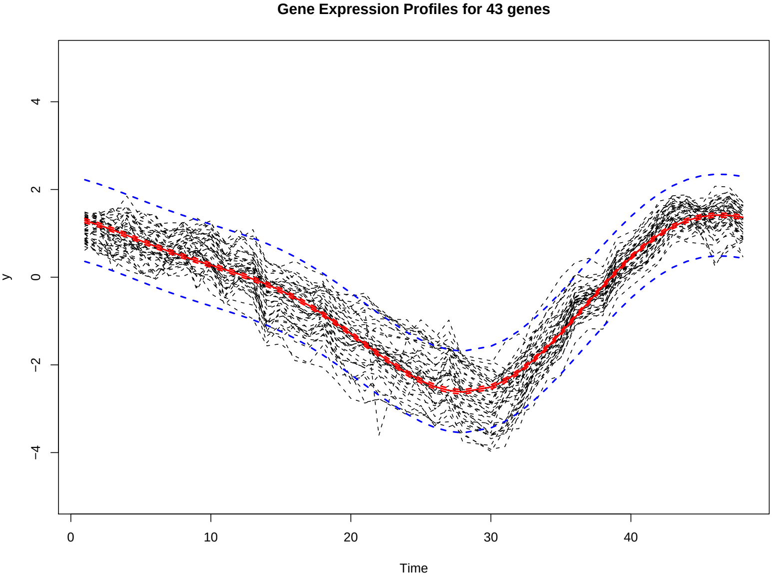
Gene Expression Data Example

Harmonic Regression Fit with $p = 2$.



Gene Expression Data Example

Harmonic Regression Fit with $p = 2$.



Why are things so straightforward ?

- because the system of equations based on the derivatives

$$\frac{\partial}{\partial \beta_j}\left\{SSE(\underset{\sim}{\beta})\right\} = 0 \qquad j = 0, 1, \ldots, k$$

can always be solved routinely, so we can always find $\widehat{\underset{\sim}{\beta}}$.

In the general model (1), simple formulae for

- $\widehat{\underset{\sim}{\beta}}$
- $s.e.(\widehat{\underset{\sim}{\beta}})$
- $\widehat{\sigma}^2$

can be found using a matrix formulation.

See handout on website - NOT EXAMINABLE !

Note: One-way ANOVA can be formulated in the form of model (1). Recall

- k independent groups
- means $\mu_1, \ldots, \mu_k$
- y_{ij} - jth observation in the ith group

Let

$$\begin{aligned} \beta_0 &= \mu_k \\ \beta_t &= \mu_t - \mu_k \qquad t = 1, 2, \ldots, k-1. \end{aligned}$$

Define new data $x_{ij}(t)$ where

$$x_{ij}(t) = \begin{cases} 1 & \text{if } t = i \\ 0 & \text{if } t \neq i \end{cases}$$

Then, using the linear regression formulation

$$y_{ij} = \beta_0 + \sum_{t=1}^{k-1} \beta_t x_{ij}(t) + \epsilon_{ij}.$$

For any i, j, $x_{ij}(t)$ is non-zero for only one value of t, when $t = i$.

We term this a regression on a *factor predictor*; it is clear that $\beta_0, \beta_1, \ldots, \beta_{k-1}$ can be estimated using least-squares.

This clarifies the link between

ANOVA and Linear Modelling

- they are essentially the SAME MODEL formulation.

This link extends to **ALL ANOVA** models; recall that we used the **General Linear Model** option in SPSS to fit two-way ANOVA.

2.2 Multiple Linear Regression

Multiple linear regression models model the variation in response y as a function of **more than one** independent variable.

Suppose we have variables

$$X_1, X_2, \ldots, X_k$$

recording different features of the experimental units. We wish to model response Y as a function of $X_1, X_2, \ldots, X_k$.

2.2.1 Multiple Linear Regression Models

Consider the model for datum i

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

where x_{ij} is the measured value of *covariate* j on experimental unit i. That is

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i$$

where the first two terms on the right hand side are the *systematic* or *deterministic* components, and the final term ϵ_i is the *random* component.

Example ($k = 2$)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

A three parameter model.

Note: We can also include *interaction* terms

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12}(x_{i1} \,.\, x_{i2}) + \epsilon_i$$

where

- The first two terms in x_{i1} and x_{i2} are **main effects**
- The third term in $(x_{i1} \,.\, x_{i2})$ is an **interaction**

This is a four parameter model.

Multiple Linear Regression Examples

SEE HANDOUT

- Multiple regression: Viscosity Example
- Factor Regression:
- Interaction
- Residuals
- SPSS Instructions

Subgroup analysis, with a factor predictor and a continuous covariate, is a form of interaction modelling; the factor predictor *interacts* with the covariate to modify the slope across the subgroups, for example.

We can describe the models using the notation previously introduced for ANOVA; consider the single binary factor predictor and single covariate case;

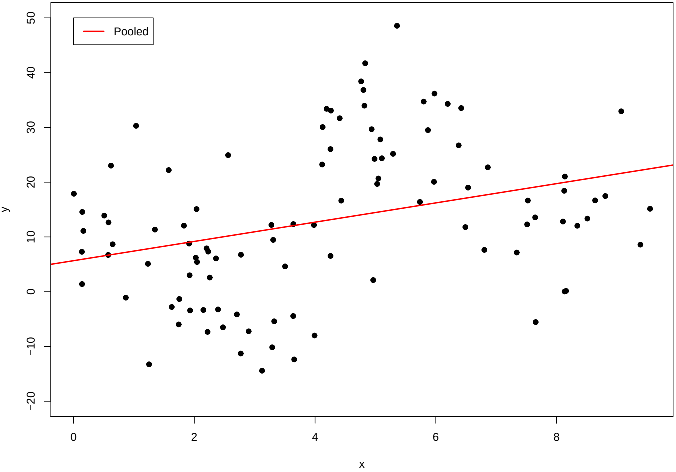
MODEL 0	Single horizontal straight line	1
MODEL 1	Two parallel horizontal straight lines	X_2
MODEL 2	Single straight line, non-zero slope	X_1
MODEL 3	Two parallel straight lines, non-zero slope	$X_1 + X_2$
MODEL 4	Two non-parallel straight lines	$X_1 + X_2 + X_1.X_2$

Note: Always be on the lookout for *lurking* subgroups (subgroups determined by the levels of an unnoticed factor predictor)

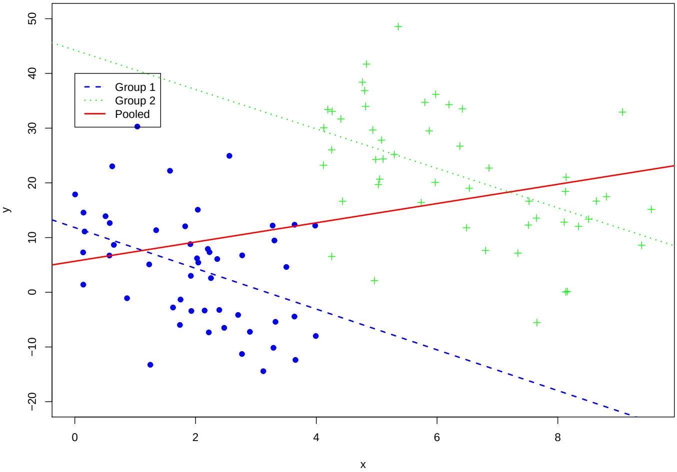
Inferences can change radically when the lurking factor is included in the model

- positive association can be converted into negative association with the continuous covariate.

For example, for factor predictor X_2 taking two levels and continuous covariate X_1. When the pooled data are examined, a **positive association** between Y and X_1 is revealed.



When the pooled data are separated into subgroups, a **negative association** between Y and X_1 in each subgroup is revealed.



$X_2 = 1$ (●) and $X_2 = 2$ (+).

i.e. increasing X_1 decreases response in subgroup 1, and decreases response in subgroup 2, but appears to increase response overall.

This is known as **Simpson's Paradox in Regression**. It illustrates that pooling data over subgroups must be carried out with care !

- you must fit the factor predictor in the model if you suspect subgroup differences exist.

In the example, the problem arises due to **dependence** between X_1 and X_2; all the group with $X_2 = 0$ have **low** values of X_1, whereas all the group with $X_2 = 1$ have **high** values of X_1

Dependence between covariates and factor predictors makes model fitting and results interpretation complicated.

Recap: we can build general models

$$y_i = \beta_0 + \sum_{j=1}^{k} x_{ij} + \epsilon_i$$

to explain the variation of y in terms of covariates and factor predictors $x_1, \ldots, x_k$.

- Simple Linear Regression
- Polynomial Regression
- Multiple Regression
- Factor Predictor Regression
- Interaction Models

We can fit each of these models easily using least-squares to obtain

- estimates $\underset{\sim}{\widehat{\beta}} = (\widehat{\beta}_1, \widehat{\beta}_2, \dots, \widehat{\beta}_k)^{\mathsf{T}}$
- standard errors
- goodness of fit measures R^2 and Adjusted R^2
- residuals for model checking
- predictions

Interpreting $\widehat{\beta}_j$

$\widehat{\beta}_j$ can be interpreted as the amount of increase in response y when x_j increases by one unit when the other predictors

$$x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$$

are held fixed.

We can test the hypothesis

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_0 &: \beta_j \neq 0 \end{aligned}$$

using the usual hypothesis testing approach.

Test statistic:

$$t_j = \frac{\widehat{\beta}_j}{s_{\widehat{\beta}_j}} = \frac{\text{ESTIMATE}}{\text{STANDARD ERROR}}$$

If H_0 is **true**,

$$t_j \sim Student(n-k-1)$$

as we are estimating $k+1$ parameters overall.

Note: In multiple regression, when testing each of

$$\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$$

we should strictly use a **multiple testing correction** (as in post-hoc tests in ANOVA)

2.2.2 Model Checking

Using the General Linear Model approach to regression, we can fit models with different numbers of predictors, and

- assess whether any individual covariate is influential in the model (look at $\widehat{\beta}, s_{\widehat{\beta}}$ and t-statistics
- assess whether there is any explanatory power in the variables combined (look at ANOVA statistics)

For the multiple regression model, the ANOVA table takes the form

SOURCE	DF	SS	MS	F
REGRESSION	k	SSR	MSR	$F = \dfrac{MSR}{MSE}$
ERROR	$n-k-1$	SSE	MSE	
TOTAL	$n-1$	SS		

where

$$MSR = \frac{SSR}{k} \qquad MSE = \frac{SSE}{n-k-1}$$

the F statistic is

$$F = \frac{MSR}{MSE}$$

and if H_0 is true

$$F \sim \text{Fisher-F}(k, n-k-1)$$

Here

$$\begin{aligned} H_0 &: \quad \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_a &: \quad \text{At least one } \beta_j \neq 0 \end{aligned}$$

The model for H_0 has one parameter β_0.
The model for H_a has $k+1$ parameters

$$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$$

Therefore the number of extra parameters for model H_a is

$$(k+1) - 1 = k$$

i.e. to obtain model H_0 from model H_a we constrain k parameters to be zero.

Because we can constrain model H_a by setting some parameters equal to zero to obtain model H_0, we say that

Model H_0 is nested inside Model H_a

The number, k, of constraints $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ explains why the ANOVA table Regression degrees of freedom is k

- the multiple regression brings in k extra parameters.

In addition, we can use the R^2 or Adjusted R^2 statistic to check overall model adequacy

$$R^2 = 1 - \frac{SSE}{SS_{yy}} = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{SSR}{SS}$$

which is equal to

$$\frac{\text{VARIATION EXPLAINED BY THE REGRESSION}}{\text{TOTAL VARIATION}}$$

Also

$$\text{Adj. } R^2 = 1 - \frac{SSE/(n-k-1)}{SS/(n-1)}$$

$R^2 > 0.7$ implies that the model is a good fit, that is, most of the variation observed is explained by the regression model.

We can now fit completely general models in the form of the General Linear Model; if y is the response, and $x_1, \ldots, x_k$ are the covariates or factor predictors, we can include combinations of

- Polynomial Main Effects : $x_j, x_j^2, x_j^3, \ldots$
- Two-way Interactions: $x_{j_1} \,.\, x_{j_2}$
- Three-way Interactions: $x_{j_1} \,.\, x_{j_2} \,.\, x_{j_3}$

etc.

In SPSS, we can use the

General Linear Model → *Univariate*

pulldown menus to build and fit the model.

- To fit factor predictors, we used the *Fixed Factor* option
- To build models, we use the

  Model → *Custom*

  selections on the *Univariate* dialog

Dummy Variables

Recall that we can fit the factor predictor using the Linear Regression pulldown if we create **dummy variables**.

For example, if factor predictor X has L levels, we create L **new** binary predictors $X_1, \ldots, X_L$, where, for $l = 1, \ldots, L$

$$X_l = \begin{cases} 1 & \text{whenever } X = l \\ 0 & \text{otherwise} \end{cases}$$

We can then include $X_1, \ldots, X_L$ in the regression model.

Example ($L = 4$)

X	X_1	X_2	X_3	X_4
3	0	0	1	0
1	1	0	0	0
3	0	0	1	0
4	0	0	0	1
2	0	1	0	0
2	0	1	0	0

See McClave and Sincich, Section 12.7.

2.2.3 Stepwise Model Selection

We seek a method that allows us to compare nested models.

Suppose we want to compare

$$\begin{aligned} \text{MODEL 1} &: \quad y = \beta_0 + \beta_1 x + \beta_2 x^2 \\ \text{MODEL 2} &: \quad y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \end{aligned}$$

Model 1 is nested inside Model 2 as if we set $\beta_3 = 0$ in Model 2, we get Model 1.

If

$$\begin{aligned} \text{MODEL 1} &: \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ \text{MODEL 2} &: \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12}(x_1.x_2) \end{aligned}$$

we can set $\beta_{12} = 0$ in Model 2 to obtain Model 1, so again the models are nested.

We can set up a hypothesis test to assess whether the simplification of Model 2 to Model 1 (by setting one or more parameters equal to zero) is justified by the data.

ANOVA tests for Comparing Nested Models

Terminology

- *Complete Model*

$$E[Y] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- *Reduced Model*

$$E[Y] = \beta_0 + \beta_1 x_1 + \cdots + \beta_g x_g$$

where $g < k$. The reduced model is obtained from the complete model by setting

$$\beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$$

The reduced model is nested inside the complete model.

We wish to test the hypothesis

$$\begin{aligned} H_0 &: \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0 \\ H_a &: \text{At least one of these } \beta_j \neq 0 \end{aligned}$$

We can test this hypothesis by fitting both models, and combining the results; we focus on the sums of squares quantities.

Method

1. Fit the **complete model** and obtain the sum of squared errors, SSE_C, available from the ANOVA table.
2. Fit the **reduced model** and obtain the sum of squared errors, SSE_R, available from the ANOVA table.
3. Form the test statistic

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)}$$

If H_0 is **true**, then $F \sim$ Fisher-F$(k-g, n-k-1)$

Note: $k-g$ is the number of parameters we set equal to zero when moving from complete to reduced model.

Using this F statistic, we can assess whether there is evidence to support the reduced model over the complete model.

Complete Model ANOVA table:

SOURCE	DF	SS	MS	F
COMPLETE MODEL	k	SSR_C	MSR_C	F_C
ERROR$_C$	$n-k-1$	SSE_C	MSE_C	
TOTAL	$n-1$	SS		

Reduced Model ANOVA table:

SOURCE	DF	SS	MS	F
REDUCED MODEL	g	SSR_R	MSR_R	F_R
ERROR$_R$	$n-g-1$	SSE_R	MSE_R	
TOTAL	$n-1$	SS		

The result holds for comparing any two nested models where the standard model assumptions hold:

- ϵ uncorrelated
- ϵ independent of $x_1, \ldots, x_k$
- ϵ has constant variance
- $\epsilon \sim N(0, \sigma^2)$

Note: It does not allow us to compare non-nested models; for example

$$\begin{aligned} \text{MODEL 1} &: \quad y = \beta_0 + \beta_1 x_1 + \epsilon \\ \text{MODEL 2} &: \quad y = \beta_0 + \beta_2 x_2 + \epsilon \end{aligned}$$

- **NOT NESTED !**

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)} = \frac{①/②}{③/④}$$

① - $SSE_R - SSE_C$: this is the improvement in fit when the reduced model is extended to the complete model

② - $k - g$: this is the number of extra parameters needed to extend the reduced model to the complete model

③ - SSE_C
④ - $n - k - 1$

③/④ - this is the best guess we have at the true value of σ^2, that is, the estimate of σ^2 constructed using as much information as possible, once the effects of

$$x_1, \ldots, x_k$$

have been accounted for.

Example (Hooker's Data)

We consider the two models:

$$\begin{aligned} \text{MODEL 1} &: \quad y = \beta_0 + \beta_1 x + \epsilon \\ \text{MODEL 2} &: \quad y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \end{aligned}$$

Here

- MODEL 1: Reduced Model
- MODEL 2: Complete Model

$k = 2$, $g = 1$.

IS THE QUADRATIC TERM NEEDED ?

Example (Hooker's Data)

COMPLETE MODEL	SSR_C	2286.933
	SSE_C	4.382
REDUCED MODEL	SSR_R	2272.474
	SSE_R	18.840

with $n = 31, k = 2, g = 1$

$$\Longrightarrow k - g = 1, n - k - 1 = 28$$

So

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)} = \frac{(18.840 - 4.382)/1}{4.382/28} = 92.383$$

Example (Hooker's Data)

We compare F with the

$$\text{Fisher-F}(k-g, n-k-1) \equiv \text{Fisher-F}(1, 28)$$

distribution.

$$F_{0.05}(1, 28) = 4.20$$

Thus

$$92.383 = F > F_{0.05}(1, 28) = 4.20$$

and $H_0 : E[Y] = \beta_0 + \beta_1 x$ is **REJECTED** in favour of $H_a : E[Y] = \beta_0 + \beta_1 x + \beta_2 x^2$.

i.e. the **quadratic model** fits better than the straight-line model.

NOTE: From the original ANOVA tables, we already know that Model 1 and Model 2 both fit better than the null model

$$\text{MODEL 0} \qquad \begin{aligned} E[Y] &= \beta_0 \\ y &= \beta_0 + \epsilon \end{aligned}$$

where there is no dependence on x.

We have now confirmed that Model 2 fits better than Model 1.

Example (Diabetes Data)

Factor Predictor: **group** (X_2)
Continuous Covariate: **loggluf** (X_1)
Response: **logglut** (Y)

We have five models to confirm:

$$
\begin{aligned}
\text{MODEL 0} &: 1 \\
\text{MODEL 1} &: X_2 \\
\text{MODEL 2} &: X_1 \\
\text{MODEL 3} &: X_1 + X_2 \\
\text{MODEL 4} &: X_1 + X_2 + X_1.X_2
\end{aligned}
$$

Example (Diabetes Data)

MODEL 4 us the most complex model with 6 parameters

$$\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}, \beta_{30}, \beta_{31}$$

MODEL 4:

$$E[Y] = \begin{cases} \beta_{10} + \beta_{11} x_1 & \text{GROUP 1} \\ \beta_{20} + \beta_{21} x_1 & \text{GROUP 2} \\ \beta_{30} + \beta_{31} x_1 & \text{GROUP 3} \end{cases}$$

All of the other models are nested inside Model 4; we can obtain them all by setting parameters equal to zero.

Example (Diabetes Data)

In the SPSS parameterization:

β_{30}, β_{31}	Group 3 Intercept and Slope
$\beta_{10} = \beta_{30} + \delta_{10}$	Changes in the Intercepts in
$\beta_{20} = \beta_{30} + \delta_{20}$	Groups 1 and 2 are δ_{10} and δ_{20}
$\beta_{11} = \beta_{31} + \delta_{11}$	Changes in the Slopes in
$\beta_{21} = \beta_{31} + \delta_{21}$	Groups 1 and 2 are δ_{11} and δ_{21}

Thus the six new parameters are

$$\beta_{30}, \beta_{31}, \delta_{10}, \delta_{20}, \delta_{11}, \delta_{21}$$

MODEL 0 $\beta_{31} = 0$
$\delta_{10} = \delta_{20} = \delta_{11} = \delta_{21} = 0$

MODEL 1 $\beta_{31} = \delta_{11} = \delta_{21} = 0$

MODEL 2 $\delta_{10} = \delta_{20} = \delta_{11} = \delta_{21} = 0$

MODEL 3 $\delta_{11} = \delta_{21} = 0$

Note: $\beta_{31} = 0 \Longrightarrow \delta_{11} = \delta_{21} = 0$, as X_1 is not included in the model.

Counting Parameters

- Whenever we remove a **continuous covariate**, from a model, we set **one** parameter to zero.
- Whenever we remove a **factor predictor** with L levels from a model, we set $L - 1$ parameters to zero.
- Whenever we remove a two-way interaction between these variables from a model, we set $1.(L - 1) = L - 1$ parameters to zero.

Models 0,1,2,3 are nested inside Model 4.

Two approaches to finding the best model are used:

1. Start with Model 0 and try to add terms that improve the model fit (**Forward Selection**)
2. Start with Model 4 and try to remove terms that improve the model fit (**Backward Selection**)

Note:

- Models 0,1 and 2 are nested inside Model 3.
- Model 0 is nested inside Models 1 and 2.

Therefore we can begin with Model 4, or Model 3 or Model 1 or 2, and simplify to a nested model.

Example (Diabetes Data)

Here $n = 144$. From SPSS output handouts:

Model	Description	SSE	p
0	1	28.504	1
1	X_2	4.160	3
2	X_1	3.738	2
3	$X_1 + X_2$	1.472	4
4	$X_1 + X_2 + X_1.X_2$	1.318	6

p is the number of non-zero parameters; k or g is always $p - 1$ in the following analysis.

Backward Selection:

Complete Model : Model 4
Reduced Model : Model 3

Here $k = 5, g = 3$ so $k - g = 2$, and

$$n - k - 1 = 144 - 5 - 1 = 138.$$

We have

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)} = \frac{(1.472 - 1.318)/2}{1.318/138} = 8.062$$

We compare this with the

$$\text{Fisher-F}(k-g, n-k-1) = \text{Fisher-F}(2, 138)$$

distribution; we have $F_\alpha(2, 138) = 3.061$, so we

$$\text{Reject } H_0 \text{ at } \alpha = 0.05$$

i.e. Model 4

$$X_1 + X_2 + X_1.X_2$$

fits **significantly better** than Model 3

$$X_1 + X_2.$$

- we cannot simplify the complete model to the reduced model without the loss of significant explanatory power.

The Interaction is Necessary in the Model

Backward selection stops here; we cannot simplify further from the complete model.

Forward Selection: we start with Model 0 and build up.

Model 1 vs Model 0 $\quad F = 412.568$

Model 2 vs Model 0 $\quad F = 940.846$

It seems that Model 2 is the better improvement, so we try the selection path

$$\text{Model } 0 \longrightarrow \text{Model } 2 \longrightarrow \text{Model } 3 \longrightarrow \text{Model } 4$$

Model	SSE	$SSE_R - SSE_C$
0	28.504	-
2	3.738	24.766
3	1.472	2.266
4	1.318	0.154

ie we work down the table, $28.504 - 3.738 = 24.766$ etc.

Comparison	k	g	SSE_C	$SSE_R - SSE_C$	F
2 vs 0	1	0	3.738	24.766	940.82
3 vs 2	3	1	1.472	2.266	107.76
4 vs 3	5	3	1.318	0.154	8.06

Recall that $n = 144$, and

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)}$$

Under each H_0,

$$F \sim \text{Fisher-F}(k-g, n-k-1)$$

- $F_{0.05}(1, 142) \doteqdot 3.92 < 940.82$
  Therefore Model 0 is **NOT** an adequate simplification of Model 2
- $F_{0.05}(2, 140) \doteqdot 3.07 < 107.76$
  Therefore Model 2 is **NOT** an adequate simplification of Model 3
- $F_{0.05}(2, 138) \doteqdot 3.07 < 8.06$
  Therefore Model 3 is **NOT** an adequate simplification of Model 4

All of the null hypotheses are **rejected**.

Therefore by both forward and backward selection, we select Model 4

$$X_1 + X_2 + X_1.X_2$$

as the most appropriate model.

Note: In this sequence of hypothesis tests, the convention is **not** to correct for multiple testing (we don't know how many tests we are going to do), although a correction could be used.

F-tests for Unbalanced Designs

Example (Potato Damage Data)

The damage to potato plants caused by cold temperatures is to be studied.

In this experimental study, three binary factor predictors were used: we label them A, B and C rather than X_1, X_2, X_3 to recall the link with Factorial Designs in ANOVA. Each factor takes two levels:

	Factor	Levels
A	Potato Variety	0- Variety 1, 1- Variety 2
B	Acclimatization Routine	0- Room Temp, 1- Cold Room
C	Preparation Treatment	0- -4C, 1- -8C

Thus we have a $2 \times 2 \times 2$ three-way factorial design.

F-tests for Unbalanced Designs

However, the design is **unbalanced**; we have different numbers of replicates in each of the 8 factor-level combinations.

This means we cannot use conventional 3-way ANOVA; the lack of balance means that the stated p-values **may be misleading if we perform a standard ANOVA**.

Thus we are forced to use the General Linear Model F-test approach.

We begin with the most complex model and do backward selection.

Here the most complex model can be written

$$A + B + C + A.B + A.C + B.C + A.B.C$$

that is,

- all main effects (terms 1,2 and 3)
- all two-way interactions (terms 4,5 and 6)
- all three-way interactions (term 7)

We may write this model

$$A * B * C$$

which is termed the full factorial model.

Counting the numbers of parameters

Term	Parameters	
A	$(a-1)$	1
B	$(b-1)$	1
C	$(c-1)$	1
$A.B$	$(a-1)(b-1)$	1
$A.C$	$(a-1)(c-1)$	1
$B.C$	$(b-1)(c-1)$	1
$A.B.C$	$(a-1)(b-1)(c-1)$	1
Total		7

where $a = b = c = 2$.

We have 7 parameters in total (excluding the baseline mean) when all terms are considered, so

$$k = 7$$

In the following tables columns are:

Complete Model
Reduced Model
SSE_C
SSE_R
k
g
F (test statistic)
$F_{0.05}(k-g, n-k-1)$

We denote the critical value by F_α and check whether $F > F_\alpha$.

Potato Damage Data: ANOVA-F Tests

We compare four models: M_{R_1}, M_{R_2} and M_{R_3} are nested within the complete model M_C.

$$\begin{aligned} M_C &: A+B+C+A.B+A.C+B.C+A.B.C \\ M_{R_1} &: A+B+C+A.B \\ M_{R_2} &: A+B+C \\ M_{R_3} &: A+B+A.B \end{aligned}$$

COMP.	RED.	SSE_C	SSE_R	k	g	F	F_α
M_C	M_{R_1}	4968.876	5093.746	7	4	0.561	2.76
M_{R_1}	M_{R_2}	5093.746	7183.674	4	3	28.721	3.92
M_{R_1}	M_{R_3}	5093.746	6319.640	4	3	16.846	3.92

Note: The quoted F_α values are approximate as the textbook does not tabulate all Fisher-F distributions. We take $\alpha = 0.05$

Conclusions

Taking the comparisons in order:

1. M_C *vs* M_{R_1} : $F < F_\alpha$. Therefore the result is **not significant**: Model M_{R_1} **is an adequate simplification** of Model M_C, and we choose M_{R_1} over M_C.

   The model M_{R_1} now becomes the complete model.

2. M_{R_1} *vs* M_{R_2} : $F > F_\alpha$. Therefore the result **is significant**: Model M_{R_2} **is not an adequate simplification** of Model M_{R_1}

3. M_{R_1} *vs* M_{R_3} : $F > F_\alpha$. Therefore the result **is significant**: Model M_{R_3} **is not an adequate simplification** of Model M_{R_1}

Thus the final model is

$$A + B + C + A.B$$

i.e. all main effects, plus the interaction between potato variety and acclimatization routine.

We cannot simplify this model further without significant loss in terms of goodness of fit.

Note: $R^2 = 0.631$ and Adjusted $R^2 = 0.610$, so we have a reasonable fit.

Task Distraction Data

Example (Task Distraction Data)

In an experimental study, the number of errors made in performing a specified task was recorded. The experiment investigated the influence of various predictors on the numbers of errors made.

There are two factor predictors (A, B) and one continuous covariate (X).

We have a balanced design with 15 people (replicates) in each factor-level subgroup.

Example (Task Distraction Data)

A	Group	1 : Non-smoker 2 : Delayed smoker 3 : Active smoker
B	Task	1 : Pattern Recognition 2 : Cognitive Task 3 : Driving Simulation
X	Distraction Level	

We compare four models with the **complete** model

Complete Model : $A * B * X$

$$A + B + X + A.B + A.X + B.X + A.B.X$$

Number of parameters

Term	Parameters		Tot.
A	$(a-1)$	$= 3 - 1$	2
B	$(b-1)$	$= 3 - 1$	2
X	(1)		1
$A.B$	$(a-1)(b-1)$	$= 2 \times 2$	4
$A.X$	$(a-1)(1)$	$= 2 \times 1$	2
$B.X$	$(b-1)(1)$	$= 2 \times 1$	2
$A.B.X$	$(a-1)(b-1)(c-1)$	$= 2 \times 2 \times 1$	4
Total			17

For illustration we consider the following sequence of models:

- Reduced Model 1: M_{R_1}

$$A + B + X + A.X + B.X$$

- Reduced Model 2: M_{R_2}

$$A + B + X + B.X$$

- Reduced Model 3: M_{R_3}

$$B + X + B.X$$

- Reduced Model 4: M_{R_4}

$$B + X$$

Task Distraction Data: ANOVA-F Tests

$$\begin{aligned}
M_C &: A + B + X + A.B + A.X + B.X + A.B.X \\
M_{R_1} &: A + B + X + A.X + B.X \\
M_{R_2} &: A + B + X + B.X \\
M_{R_3} &: B + X + B.X \\
M_{R_4} &: B + X
\end{aligned}$$

COMP.	RED.	SSE_C	SSE_R	k	g	F	F_α
M_C	M_{R_1}	5660.010	7627.479	17	9	5.084	2.02
M_{R_1}	M_{R_2}	7627.479	7971.274	9	7	2.817	3.07
M_{R_2}	M_{R_3}	7971.274	8404.654	7	5	3.452	3.07
M_{R_3}	M_{R_4}	8404.654	11154.715	5	3	21.105	3.07

Conclusions

Taking the comparisons in order:

1. M_C *vs* M_{R_1} : $F > F_\alpha$. Therefore the result is **significant**: Model M_{R_1} **is not an adequate simplification** of Model M_C
2. M_{R_1} *vs* M_{R_2} : $F < F_\alpha$. Therefore the result **is not significant**: Model M_{R_2} **is an adequate simplification** of Model M_{R_1}
3. M_{R_2} *vs* M_{R_3} : $F > F_\alpha$. Therefore the result **is significant**: Model M_{R_3} **is not an adequate simplification** of Model M_{R_2}
4. M_{R_3} *vs* M_{R_4} : $F > F_\alpha$. Therefore the result **is significant**: Model M_{R_4} **is not an adequate simplification** of Model M_{R_3}

Follow-up Analysis

In a follow up analysis (see Handout), it transpires that the model

$$A + B + X + A.B + A.X + B.X$$

ie selected.

Note: $R^2 = 0.863$ and Adjusted $R^2 = 0.831$, so we have a good fit.

Note: we must take great care with the sequence of models.

Stepwise Selection in SPSS: Options

It is possible to carry out stepwise selection in SPSS using the *Linear Regression* pulldown menu, and the *Method* pulldown list.

- **Enter :** All variables in a *block* are entered in a single step.
- **Stepwise :** At each step, the independent variable not in the equation that has the **smallest** p-value in the F-test is entered, if that probability is sufficiently small. Variables already in the regression equation are **removed** if their p-value becomes sufficiently large. The method terminates when no more variables are eligible for inclusion or removal.
- **Remove :** All variables in a block are removed in a single step.

Stepwise Selection in SPSS: Options

- **Backward :** Variables are entered into the equation and then sequentially removed. The variable with the smallest *partial correlation* with the dependent variable is considered first for removal. After the first variable is considered, the variable remaining in the equation with the smallest partial correlation is considered next. The procedure stops when there are no variables in the equation that satisfy the removal criteria.
- **Forward :** Variables are sequentially entered into the model starting from the null model. The first variable considered for entry into the equation is the one with the largest positive or negative correlation with the dependent variable. This variable is entered into the equation only if it satisfies the criterion for entry. If the first variable is entered, the independent variable not in the equation that has the largest partial correlation is considered next. The procedure stops when there are no variables that meet the entry criterion.

2.2.5 Pitfalls of Regression Modelling

Five issues to bear in mind in ANOVA, Regression and General Linear Modelling.

1. Model assumptions
2. Data transformations
3. Model selection
4. Multicollinearity
5. Predicting beyond the range of the covariates

See Handout.

Part III

Non-Parametric Statistics

Non-Parametric Statistics

All of the previous statistical analysis methods studied (t-tests, ANOVA, Regression, General Linear Modelling) have depended heavily on **distributional assumptions**.

i.e. we assume that the data are **Normally distributed**.

We now seek statistical procedures that do not rely on this strong assumption. We term these methods

NON-PARAMETRIC

or

DISTRIBUTION-FREE

They substitute **large sample approximations** for the distributional assumptions.

3.1 Distribution-free tests for Categorical Data

Categorical data are data in which experimental units are allocated to one of a number of categories according to their characteristics. The categories are defined by one or more factors

Examples:

- Female/Male - two categories
- Smoker/Former Smoker/Non Smoker - three categories.

Doll and Hill Data

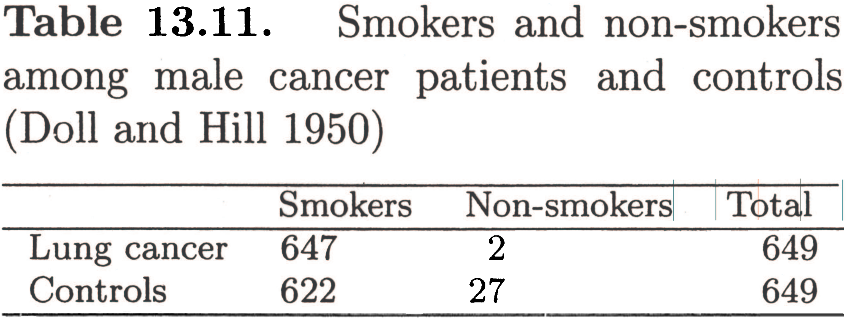


Table 13.11. Smokers and non-smokers among male cancer patients and controls (Doll and Hill 1950)

	Smokers	Non-smokers	Total
Lung cancer	647	2	649
Controls	622	27	649

Juvenile Delinquency and Spectacle-Wearing

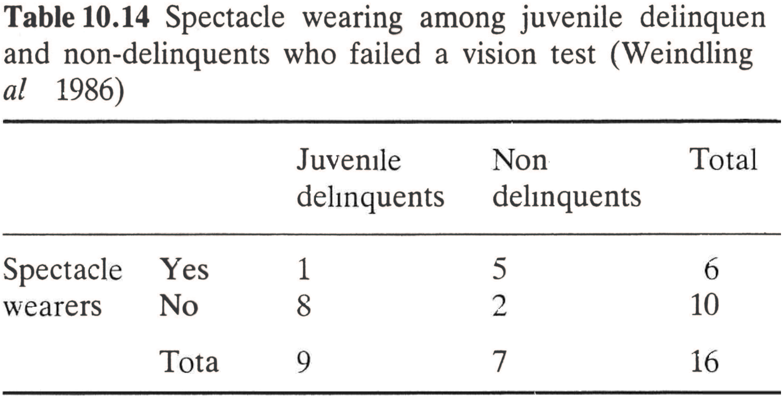


Table 10.14 Spectacle wearing among juvenile delinquen and non-delinquents who failed a vision test (Weindling *al* 1986)

		Juvenile delinquents	Non delinquents	Total
Spectacle wearers	Yes	1	5	6
	No	8	2	10
	Tota	9	7	16

The data are **counts** of experimental units that fall into each category. Suppose

1. There are n experimental units in the study
2. There are k categories
3. The probabilities of the k outcomes are $p_1, \ldots, p_k$, where

$$p_1 + \cdots + p_k = 1$$

4. The experimental units are **independent**
5. The counts in the k categories are $n_1, \ldots, n_k$, where

$$n_1 + \cdots + n_k = n$$

The experimental design is termed a **Multinomial Experiment**

Note: The categories can arise as combinations of factor levels; we can have

- **one-way** classification (categories of a single factor, A)
- **two-way** classification (categories defined by combinations of levels of two factors, A and B)

and so on. The counts table is often called a **contingency table** and the entries in the table are called **cells**.

The idea can be extended to larger numbers of factors $(A, B, C, \ldots)$ to produce a multi-way table. We will focus on at most two-way tables, with r rows and c columns, yielding an $r \times c$ table.

What kinds of tests can be carried out for such data ?

1. Tests about $p_1, \ldots, p_k$
    - $H_0 : p_1 = \cdots = p_k = 1/k$
    - $H_0 : p_1, \ldots, p_k$ determined by some parametric distribution (Normal, Poisson etc.)
2. Tests about the factors A and B
    - are A and B dependent ?
    - i.e. does classification by A influence classification by B.

Chi-Squared Test

For one-way tables: suppose that a null hypothesis **completely specifies** $p_1, \ldots, p_k$, that is, we have H_0 of the form

$$H_0 \ : \ p_1 = p_1^{(0)}, \ldots, p_k = p_k^{(0)}$$

where $p_1^{(0)}, \ldots, p_k^{(0)}$ are fixed probabilities. For example, for $k = 3$,

$$H_0 \ : \ p_1 = p_2 = p_3 = 1/3$$

or

$$H_0 \ : \ p_1 = 1/2, p_2 = p_3 = 1/4$$

To test this hypothesis against the general alternative hypothesis

$$H_a \; : \; H_0 \text{ not true.}$$

we use the test statistic

$$X^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}}$$

If H_0 is true,

$$X^2 \mathrel{\dot{\sim}} \text{Chi-squared}(k-1).$$

that is, X^2 is approximately distributed as Chi-squared$(k-1)$.

In this formula

- n_i is the **observed** count in cell i
- $np_i^{(0)}$ is the **expected** count in cell i if H_0 is **true**.

Sometimes the formula is written

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed count, and E_i is the expected count.

If

$$X^2 > \mathrm{Chisq}_\alpha(k-1)$$

then we reject H_0 at the α significance level, where $\mathrm{Chisq}_\alpha(k-1)$ is the $1-\alpha$ (right-hand) tail critical value of the Chi-squared distribution with $k-1$ degrees of freedom.

This method can be extended in the one-way case to test distribution assumptions, that is, for example

$$H_0 \;:\; \text{Data Normally distributed}$$

or

$$H_0 \;:\; \text{Data Poisson distributed}$$

Unfortunately this facility is not available in SPSS; direct calculation is possible but involved.

For the **two-way** table, we can test the hypothesis

$$H_0 \ : \ \text{Factor A and Factor B levels are assigned independently}$$

that is, classification by factor A is independent of classification by factor B. We use the same test statistic that can be rewritten

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \widehat{n}_{ij})^2}{\widehat{n}_{ij}}$$

where

$$\widehat{n}_{ij} = \frac{n_{i.} n_{.j}}{n} \qquad n_{i.} = \sum_{j=1}^{c} n_{ij} \qquad n_{.j} = \sum_{i=1}^{r} n_{ij}.$$

The terms $n_{i.}$ and $n_{.j}$ are the row and column totals for row i and column j respectively.

If H_0 is true

$$X^2 \stackrel{\cdot}{\sim} \text{Chi-squared}((r-1)(c-1))$$

i.e. the degrees of freedom quantity is $(r-1)(c-1)$. Otherwise the test proceeds as before: we check whether

$$X^2 > \text{Chisq}_\alpha((r-1)(c-1))$$

and if so, we reject H_0.

Example (DNA Sequence Data)

Counts of Nucleotides A,C,G,T in a genomic segment related to the breast cancer gene BRCA2.

Example (Eye and Hair Colour Data)

The assignment of hair and eye colour in a sample of humans

See handout.

Note: For the Chi-squared test to be valid, we need the expected cell counts

$$np_i^{(0)} \qquad i = 1, \ldots, k$$

or

$$\widehat{n}_{ij} \qquad i = 1, \ldots, r, j = 1, \ldots, c$$

to be sufficiently large. The convention is to require the expected value to be greater than **five**.

Note: If $r = c = 2$ we have a 2×2 table, and another **exact** test can be used which does not rely on the large sample approximation

Fisher's Exact Test

- another test for independence of assignment of the row and column factor levels
- test statistic and null distribution are complicated (based on the **hypergeometric distribution**)
- SPSS computes test statistic and p-value.

Example (Juvenile Delinquency and Spectacle Wearing)

Is there any association between the two factors ?

A : Spectacle Wearing (Yes/No)

B : Juvenile Delinquent (Yes/No)

		Delinquent		
		Yes	No	$n_{i.}$
Spectacles	Yes	1	5	6
	No	8	2	10
	$n_{.j}$	9	7	16

Example (Juvenile Delinquency and Spectacle Wearing)

Chi-squared Test:

$$X^2 = 6.112$$

Compare with Chi-squared$((r-1)(c-1))$ = Chi-squared(1); we have

$$\text{Chi-squared}_{0.05}(1) = 3.841$$

and a p-value of 0.013. Therefore we **reject** H_0.

Fisher's Exact Test: p-value is 0.035 (1-sided) or 0.024 (2-sided).

Thus we reject H_0 and we have evidence of association between the factors.

Case-Control Studies

A **case-control** study is an observational study where participants are selected for the study with regard to their **disease status**.

- a sample of **cases** (disease sufferers)
- a sample of **controls** (healthy patients)

We investigate the possible association between disease status and a factor that takes two levels. A 2×2 table of counts is formed for all combinations of disease status/factor level.

Example (BCG Vaccination and Leprosy)

Disease Status : Leprosy Sufferer (Yes/No)
Factor : Vaccination Scar (Yes/No)

		Disease Status		
		Case	Control	
		Yes	No	$n_{i.}$
Scar	Yes	101	554	655
	No	159	446	605
	$n_{.j}$	260	1000	1260

Is there an association ? Does vaccination induce leprosy ?

The Chi-squared test is potentially not valid here because of the design. An alternative test statistic is based on the **odds ratio**

$$\text{O.R.} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \widehat{\psi}$$

say. The test statistic is

$$Z = \frac{\log \widehat{\psi}}{\text{s.e.}(\log \widehat{\psi})}$$

where

$$\text{s.e.}(\log \widehat{\psi}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

That is,

$$Z = \frac{\log n_{11} + \log n_{22} - \log n_{12} - \log n_{21}}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}}$$

Under

$$H_0 \ : \ \text{No association between factor and disease status}$$

it follows that

$$Z \underset{\cdot}{\sim} N(0, 1)$$

Here log means ln or **natural log**.

Example (BCG Vaccination and Leprosy)

$$n_{11} = 101, n_{12} = 554, n_{21} = 159, n_{22} = 446$$

Therefore

$$\widehat{\psi} = \frac{n_{11} n_{22}}{n_{12} n_{21}} = 0.511 \qquad \log \widehat{\psi} = -0.671$$

and

$$\text{s.e.}(\log \widehat{\psi}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} = 0.142$$

so

$$Z = \frac{-0.671}{0.142} = -4.717$$

For a text at $\alpha = 0.05$, the two-sided critical value is ± 1.96, so we

Reject H_0.

Example (Smoking and Lung Cancer)

$$n_{11} = 647, n_{12} = 622, n_{21} = 2, n_{22} = 27$$

Therefore

$$\log \widehat{\psi} = \log \frac{647 \times 27}{2 \times 622} = 2.642$$

and

$$\text{s.e.}(\log \widehat{\psi}) = \sqrt{\frac{1}{647} + \frac{1}{2} + \frac{1}{622} + \frac{1}{27}} = 0.735$$

so

$$Z = \frac{2.642}{0.735} = 3.590$$

For a text at $\alpha = 0.05$, the two-sided critical value is ± 1.96, so we

Reject H_0

and report evidence for association.

3.2 Single Population Tests

We seek non-parametric or distribution-free tests for hypotheses relating to single samples, the equivalents of one-sample Z- or T-tests, which rely on the **normality** of the samples.

Normally these tests are formulated in terms of **ranks** of the data to give

Rank Tests

For example, if the data are

0.24 3.16 1.97 2.10 0.92

we sort them into **ascending** order, and assign ranks in order

	0.24	0.92	1.97	2.10	3.16
Rank	1	2	3	4	5

The tests depend on the behaviour of statistics computed in terms of the ranks, and rely on a **large sample** justification.

Rather than test the **mean**, we test the **median**, x_{MED}, where

$$\Pr[\text{Observation } \leq x_{\mathrm{MED}}] = \frac{1}{2}$$

i.e. the halfway point of the distribution.

The **sample median** is the halfway point of the sorted sample.

Let η denote the population median. We wish to test, for example,

$$H_0 \,:\, \eta = \eta_0$$

See Handout

3.3 Comparing Two Populations : Independent Samples

We seek a non-parametric equivalent to the two-sample *t*-test.

Instead of testing population **means**,

$$H_0 \; : \; \mu_1 = \mu_2$$

we test population **medians**

$$H_0 \; : \; \eta_1 = \eta_2$$

One- and Two- sample tests

- In the **one sample** case we use the

  SIGN TEST

  to test hypotheses about η

- In the **two sample** case we use the

  WILCOXON RANK SUM or **MANN-WHITNEY U** test.

See Handout

Note: For the MWW test

- **Textbook convention :** Label the samples so that $n_1 > n_2$ (i.e. sample 1 is the one with the larger sample size)
- **SPSS convention :** Label the samples such that

  $$x_{\mathrm{MED}_1} < x_{\mathrm{MED}_2}$$

  (i.e. sample 1 is the one with the smaller median) and only test

  $$H_0 \, : \, \eta_1 = \eta_2$$

Other two sample tests are available:

- Kolmogorov-Smirnov Test
- Moses Extreme Reactions Test
- Wald-Wolfowitz Runs Test

None make distributional assumptions, all perform best when the sample size is large.

3.4 Comparing Two Dependent Samples

Suppose we have repeat measurements on the same experimental units.

In this case, the **within-subject** data are **dependent**; we have pairing of observations.

We can use the

Wilcoxon Signed Rank Test

See Handout

3.4 Comparing Three or More Populations

We now seek non-parametric equivalents to ANOVA useful for different designs. We study tests for

(a) the **Completely Randomized Design** (CRD)
(b) the **Randomized Block Design** (RBD)

For (a) we use the

Kruskal-Wallis Test

and for (b) we use the

Friedman Test.

See Handout

Summary of the Non-Parametric Tests

- **Chi-Squared Test :** Goodness of Fit/independence in contingency tables
- **Sign Test :** One Sample (equivalent of one sample t-test)
- **Mann-Whitney-Wilcoxon :** Two Sample (equivalent of two sample t-test)
- **Wilcoxon Signed Rank :** Paired Data
- **Kruskal-Wallis :** one-way layout, multigroup comparison - equivalent of ANOVA for CRD.
- **Friedman :** two-way blocked layout, equivalent of two-way ANOVA for RBD.

Pros:

- No distributional assumptions
- Applicable for most sorts of data
- Large sample approximations make them easy to implement

Cons:

- Low power compared to parametric tests (i.e. often do not reject H_0 when they should - prone to Type II Error)
- Small sample null distributions difficult to compute.

3.6 Rank Correlation

To measure the association between two variables, we previously used the *correlation coefficient*, r; for data $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$,

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

where

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \quad SS_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 \quad SS_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

r is a measure of the linear association between X and Y

Pearson Product Moment Coefficient of Correlation

A more general measure of association is the

Spearman Rank Correlation Coefficient

We compute this as follows:

1. For each sample separately, compute the **ranks** of the data, denote the ranks for the data $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ by $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ respectively.

2. Compute

$$r_S = \frac{SS_{uv}}{\sqrt{SS_{uu} SS_{vv}}}$$

   ie the Pearson correlation between the ranks.

r_S is the **Spearman Correlation**.

Notes:

1. If there are no ties in the data

   $$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

   where $d_i = u_i - v_i$.

2. r_S is potentially a measure of the **non-linear** association between X and Y.

   The calculation can be applied directly to rank data i.e. $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ can be preference ranks given by two observers.

Tests for r_S

To test

$$H_0 \ : \ \rho = 0$$

vs

$$\begin{array}{lccc} (1) & H_a & : & \rho > 0 \\ (2) & H_a & : & \rho < 0 \\ (3) & H_a & : & \rho \neq 0 \end{array}$$

We may use r_S as a test statistic. The distribution of r_S under H_0 is tabulated in McClave and Sincich.

If Spearman_α is the α tail quantile of the null distribution, we have the following rejection regions:

$$\begin{aligned}
(1) &\quad : \quad \text{Reject } H_0 \text{ if } r_S > \text{Spearman}_\alpha \\
(2) &\quad : \quad \text{Reject } H_0 \text{ if } r_S < -\text{Spearman}_\alpha \\
(3) &\quad : \quad \text{Reject } H_0 \text{ if } |r_S| > \text{Spearman}_{\alpha/2}
\end{aligned}$$

The Role of Randomization and Permutation Tests

Randomization or **Permutation** procedures are useful for computing **exact** null distributions for non-parametric test statistics when sample sizes are small.

We focus first on two sample comparisons; suppose that two data samples $x_1 \ldots, x_{n_1}$ and $y_1 \ldots, y_{n_2}$ (where $n_1 \geq n_2$) have been obtained, and we wish to carry out a comparison of the two populations from which the samples are drawn. The Wilcoxon test statistic, W, is the sum of the ranks for the second sample. The permutation test proceeds as follows:

1. Let $n = n_1 + n_2$. Assuming that there are no ties, the pooled and ranked samples will have ranks

$$1 \quad 2 \quad 3 \quad \dots \quad n$$

2. The test statistic is $W = R_2$, the rank sum for sample two items. For the observed data, W will be the sum of n_2 of the ranks given in the list above.

3. If the null hypothesis

$$H_0 \; : \text{ No difference between population 1 and population 2}$$

   were **true**, then we would expect **no pattern** in the arrangements of the group labels when sorted into ascending order. That is, the sorted data would give rise a **random** assortment of group 1 and group 2 labels.

4. To obtain the exact distribution of W under H_0 (which is what we require for the assessment of statistical significance), we could compute W for all possible permutations of the group labels, and then form the probability distribution of the values of W. We call this the **permutation null distribution**.

5. But W is a rank sum, so we can compute the permutation null distribution simply by tabulating **all possible subsets** of size n_2 of the set of ranks $\{1, 2, 3, \ldots, n\}$.

6. There are

$$\binom{n}{n_2} = \frac{n!}{n_1!\, n_2!} = N$$

say possible subsets of size n_2. For example, for $n = 6$ and $n_2 = 2$, the number of subsets of size n_2 is

$$\binom{8}{2} = \frac{8!}{6!\, 2!} = 28$$

However, the number of subsets increases dramatically as n increases; for $n_1 = n_2 = 10$, so that $n = 20$, the number of subsets of size n_2 is

$$\binom{20}{10} = \frac{20!}{10!\, 10!} = 184756$$

7. The exact rejection region and p-value are computed from the permutation null distribution. Let $W_i, i = 1, \ldots, N$ denote the value of the Wilcoxon statistic for the N possible subsets of the ranks of size n_2. The probability that the test statistic, W, is less than or equal to w is

   $$\Pr[W \leq w] = \frac{\text{Number of } W_i \leq w}{N}$$

   We seek the values of w that give the appropriate rejection region, $\mathcal{R}$, so that

   $$\Pr[W \in \mathcal{R}] = \frac{\text{Number of } W_i \in \mathcal{R}}{N} = \alpha$$

   It may not be possible to find critical values, and define $\mathcal{R}$, so that this probability is **exactly** α as the distribution of W is **discrete**.

Simple Example

Suppose $n_1 = 7$ and $n_2 = 3$. There are

$$\binom{10}{3} = \frac{10!}{7!\,3!} = 120$$

subsets of the ranks $\{1, 2, 3, \ldots, 10\}$ of size 3. The subsets are listed below, together with the rank sums.

Ranks			W	Ranks			W	Ranks			W	Ranks			W
1	2	3	6	1	7	8	16	2	7	10	19	4	6	7	17
1	2	4	7	1	7	9	17	2	8	9	19	4	6	8	18
1	2	5	8	1	7	10	18	2	8	10	20	4	6	9	19
1	2	6	9	1	8	9	18	2	9	10	21	4	6	10	20
1	2	7	10	1	8	10	19	3	4	5	12	4	7	8	19
1	2	8	11	1	9	10	20	3	4	6	13	4	7	9	20
1	2	9	12	2	3	4	9	3	4	7	14	4	7	10	21
1	2	10	13	2	3	5	10	3	4	8	15	4	8	9	21
1	3	4	8	2	3	6	11	3	4	9	16	4	8	10	22
1	3	5	9	2	3	7	12	3	4	10	17	4	9	10	23
1	3	6	10	2	3	8	13	3	5	6	14	5	6	7	18
1	3	7	11	2	3	9	14	3	5	7	15	5	6	8	19
1	3	8	12	2	3	10	15	3	5	8	16	5	6	9	20
1	3	9	13	2	4	5	11	3	5	9	17	5	6	10	21
1	3	10	14	2	4	6	12	3	5	10	18	5	7	8	20
1	4	5	10	2	4	7	13	3	6	7	16	5	7	9	21
1	4	6	11	2	4	8	14	3	6	8	17	5	7	10	22
1	4	7	12	2	4	9	15	3	6	9	18	5	8	9	22
1	4	8	13	2	4	10	16	3	6	10	19	5	8	10	23
1	4	9	14	2	5	6	13	3	7	8	18	5	9	10	24
1	4	10	15	2	5	7	14	3	7	9	19	6	7	8	21
1	5	6	12	2	5	8	15	3	7	10	20	6	7	9	22
1	5	7	13	2	5	9	16	3	8	9	20	6	7	10	23
1	5	8	14	2	5	10	17	3	8	10	21	6	8	9	23
1	5	9	15	2	6	7	15	3	9	10	22	6	8	10	24
1	5	10	16	2	6	8	16	4	5	6	15	6	9	10	25
1	6	7	14	2	6	9	17	4	5	7	16	7	8	9	24
1	6	8	15	2	6	10	18	4	5	8	17	7	8	10	25
1	6	9	16	2	7	8	17	4	5	9	18	7	9	10	26
1	6	10	17	2	7	9	18	4	5	10	19	8	9	10	27

There are 22 possible rank sums, $\{6, 7, 8, \ldots, 25, 26, 27\}$; the number of times each is observed is displayed in the table below, with the corresponding probabilities and cumulative probabilities.

W	6	7	8	9	10	11	12	13	14	15	16	17	18	**19**	20	21	22	23	24	25	26	27
Frequency	1	1	2	3	4	5	7	8	9	10	10	10	10	**9**	8	7	5	4	3	2	1	1
Prob.	0.008	0.008	0.017	0.025	0.033	0.042	0.058	0.067	0.075	0.083	0.083	0.083	0.083	**0.075**	0.067	0.058	0.042	0.033	0.025	0.017	0.008	0.008
Cumulative Prob.	0.008	0.017	0.033	0.058	0.092	0.133	0.192	0.258	0.333	0.417	0.500	0.583	0.667	0.742	0.808	0.867	0.908	0.942	0.967	0.983	0.992	1.000

Thus, for example, the probability that $W = 19$ is 0.075, with a frequency of 9 out of 120. From this table, we deduce that

$$\Pr[8 \le W \le 25] = 0.983 - 0.033 = 0.950$$

implying that the two-sided rejection region for $\alpha = 0.05$ is the set $\mathcal{R} = \{6, 7, 26, 27\}$.

Placenta Permeability Data

Example (Placenta Permeability Data)

Measurements of placenta permeability are made on two groups of subjects.

The data and their ranks for are displayed below:

Group	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2
Obs.	0.73	0.80	0.83	1.04	1.38	1.45	1.46	1.64	1.89	1.91	0.74	0.88	0.9	1.15	1.21
Rank	1	3	4	7	10	11	12	13	14	15	2	5	6	8	9

Placenta Permeability Data

Example

Thus the Wilcoxon statistic is

$$W = R_2 = 2 + 5 + 6 + 8 + 9 = 30$$

Now, here $n_1 = 10$ and $n_2 = 5$. There are

$$\binom{15}{5} = \frac{15!}{10!\,5!} = 3003$$

subsets of the ranks $\{1, 2, 3, \ldots, 15\}$ of size 5.

In the permutation null distribution, the possible values of W are $\{15, 16, \ldots, 64, 65\}$; the probabilities are given below.

Placenta Permeability Data

Example

W	15	16	17	18	19	20	21	22	23	24	25	26	27
Frequency	1	1	2	3	5	7	10	13	18	23	30	36	45
Prob.	0.000	0.000	0.001	0.001	0.002	0.002	0.003	0.004	0.006	0.008	0.010	0.012	0.015
Cumulative Prob.	0.000	0.001	0.001	0.002	0.004	0.006	0.010	0.014	0.020	0.028	0.038	0.050	0.065

W	28	29	30	31	32	33	34	35	36	37	38	39	40
Frequency	53	63	72	83	92	103	111	121	127	134	137	141	141
Prob.	0.018	0.021	0.024	0.028	0.031	0.034	0.037	0.040	0.042	0.045	0.046	0.047	0.047
Cumulative Prob.	0.082	0.103	0.127	0.155	0.185	0.220	0.257	0.297	0.339	0.384	0.430	0.477	0.523

W	41	42	43	44	45	46	47	48	49	50	51	52	53
Frequency	141	137	134	127	121	111	103	92	83	72	63	53	45
Prob.	0.047	0.046	0.045	0.042	0.040	0.037	0.034	0.031	0.028	0.024	0.021	0.018	0.015
Cumulative Prob.	0.570	0.616	0.661	0.703	0.743	0.780	0.815	0.845	0.873	0.897	0.918	0.935	0.950

W	54	55	56	57	58	59	60	61	62	63	64	65	
Frequency	36	30	23	18	13	10	7	5	3	2	1	1	
Prob.	0.012	0.010	0.008	0.006	0.004	0.003	0.002	0.002	0.001	0.001	0.000	0.000	
Cumulative Prob.	0.962	0.972	0.980	0.986	0.990	0.994	0.996	0.998	0.999	0.999	1.000	1.000	

Placenta Permeability Data

Example

By inspection of the table, we see that

$$\Pr[25 \le W \le 55] = 0.972 - 0.038 = 0.934$$

and

$$\Pr[24 \le W \le 56] = 0.980 - 0.028 = 0.952$$

Placenta Permeability Data

Example

Thus for a symmetric two-sided interval which contains at most probability 0.95, we take the interval

$$\{25, 26, \ldots, 54, 55\}$$

and hence define the rejection region

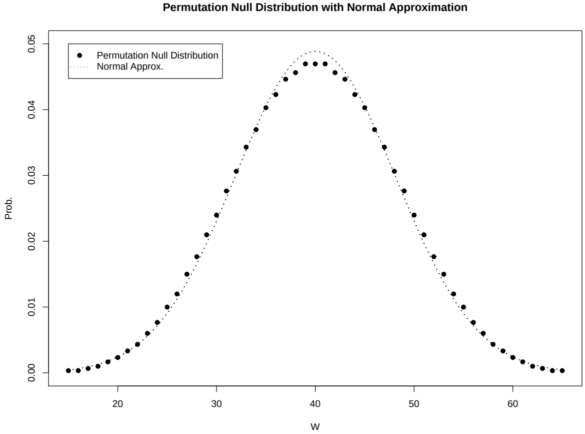
$$\mathcal{R} = \{16, 17, \ldots, 23, 24, 56, 57, \ldots, 64, 65\}$$

Note that this choice of rejection region ensures that there is at least probability 0.025 in each tail.

Thus in this example we **do not reject** the hypothesis of equal medians.

Normal Approximation

The permutation null distribution of W is displayed below.



The normal approximation is given by

$$W \mathrel{\dot\sim} \text{Normal}\left(\frac{n_2(n_1+n_2+1)}{2}, \frac{n_1 n_2(n_1+n_2+1)}{12}\right)$$