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/

January 9, 2008

Textbook: McClave and Sincich (2006), *Statistics* (10th Edition), Chapters 10-14.

Prerequisites: MATH 203 (or equivalent)

Some statistical computing knowledge useful.

Method of Assessment:

- Assignments
- Mid-Term
- Final

Precise breakdown to be confirmed.

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 $x_1, \dots, x_n$

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

$\overline{x}$ 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

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

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

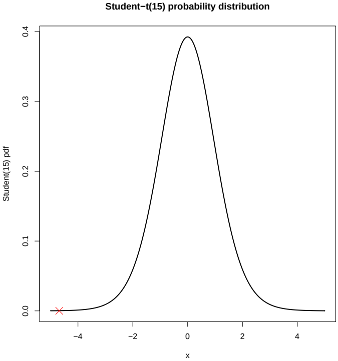
Student-t

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

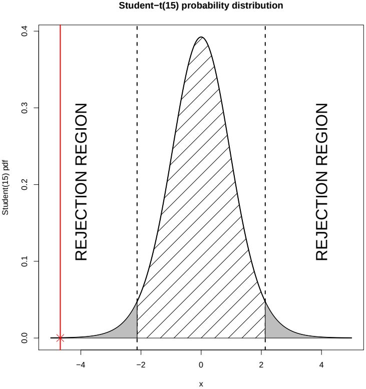
How do we quantify the “statistical significance” ?

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:

$$
\begin{aligned}
H_0 : &\quad \sigma_A = \sigma_B \\
H_a : &\quad \sigma_A \neq \sigma_B
\end{aligned}
$$

Test statistic is

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

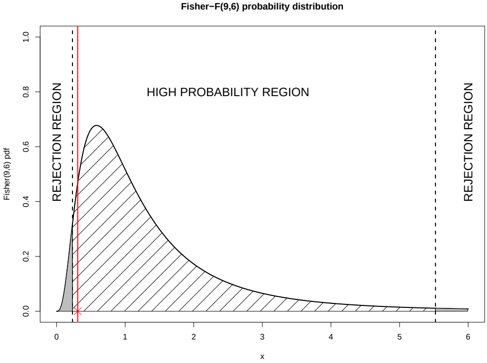
If H_0 is true, F should look like an observation from a

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 $\mathit{Student} - t(15)$ 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 $\longrightarrow$ OBSERVATIONAL STUDY

(b) Doctors $\longrightarrow$ 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

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

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**,

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

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

$$\text{Reject } H_0$$

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

Note: Tables in McClave and Sincich (p 901) 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 $F = 51.851$.

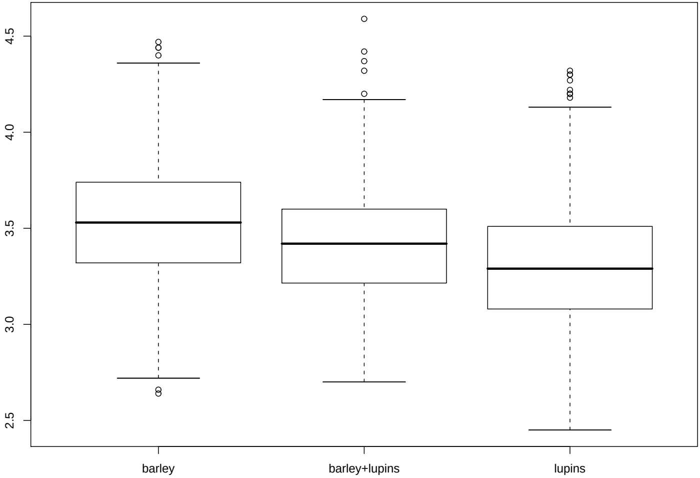
Are the assumptions met ?

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 (p 901)

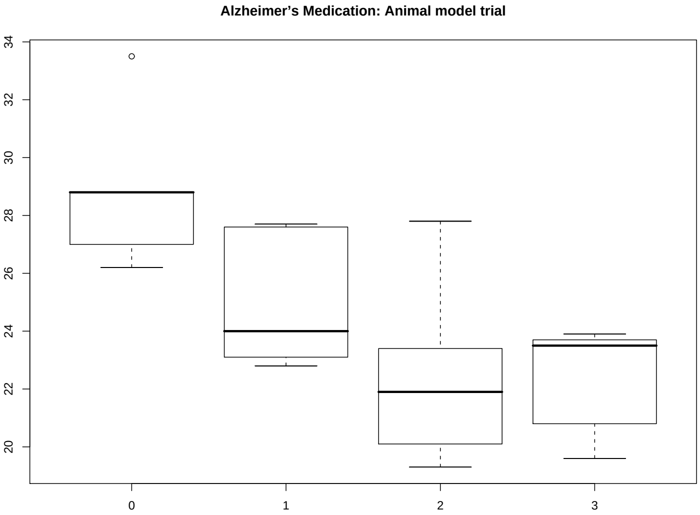
$$F_{0.05}(3, 16) = 3.24$$

and so we

$$\text{Reject } H_0$$

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	$MST = \dfrac{SST}{(k-1)}$	$F = \dfrac{MST}{MSE}$
ERROR	$n-k$	SSE	$MSE = \dfrac{SSE}{(n-k)}$	
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	$F = \dfrac{MST}{MSE}$	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 in McClave and Sincich 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

$$F_1, F_2, \dots, F_N$$

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

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

and the p-value is

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

where

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

is the true test statistic.

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

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

where

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

("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 $\alpha = 0.05$

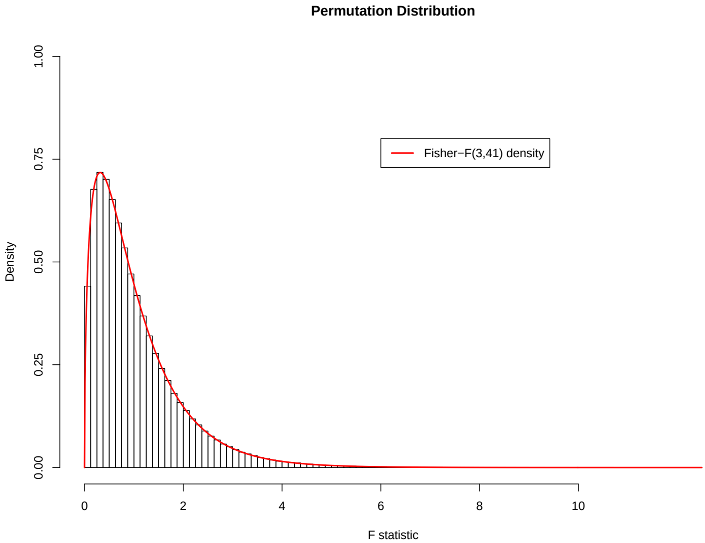
$$\begin{aligned} \text{CRITICAL VALUE} &: \quad C_R = 2.844 \\ p\text{-VALUE} &: \quad p = 0.040 \end{aligned}$$

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.