ACE2046 Quantitative Techniques
Statistical Computing

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 1, 2012-13

3 Inference for Normal Populations: Analysis of Variance

3.1 Analysis of Variance (ANOVA)

We can use t-tests and confidence intervals with one or two normal samples. Sometimes we have more than two and sometimes more complicated structures. In such cases we use the *analysis of variance* (ANOVA).

Suppose that we have k groups. Here, we test the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

using the F–distribution. The key assumptions in an ANOVA are

- Independence;
- Normality;
- Constant variance.

Again, `Minitab` can be used to check these assumptions. The most common types that you are likely to encounter are:

- Unstructured treatments – **one–way ANOVA**;
- Treatments and blocks – **two–way ANOVA**;

3.1.1 Example: One–way ANOVA

This is the simplest type of ANOVA, and it will be demonstrated using the Cholesterol and diet data.

Problem: We wish to test for differences between eating 0, 3, and 6 portions of grain. To do this we use a 1-way ANOVA table

Analysis

1. Before doing anything, we examine the data using standard graphical methods. *Eg* box-plots.
2. We now set up our Null hypothesis:
   - $H_0 : \mu_0 = \mu_3 = \mu_6$, i.e. eating grain does not affect your cholesterol
3. Now our alternative hypothesis
   - H_A : eating grain does affect your cholesterol

As with the two-sample t-test, there are two ways to do this in Minitab. We can either

1. have the samples in different columns:

   `Stat -> ANOVA -> One-way (Unstacked)`

2. or have all of the cholesterol values in one column and have a separate `factor` column.

   `Stat -> ANOVA -> One-way`

Method 2 was used here. I also selected `Graphs -> Four in one` to produce some *residual plots* as a check of the model assumptions.

```
Source DF SS      MS      F     P
C1     2  0.1884  0.0942  3.92  0.028
Error 37  0.8883  0.0240
Total 39  1.0768
```

Next look at the residual graphs generated to check our assumptions

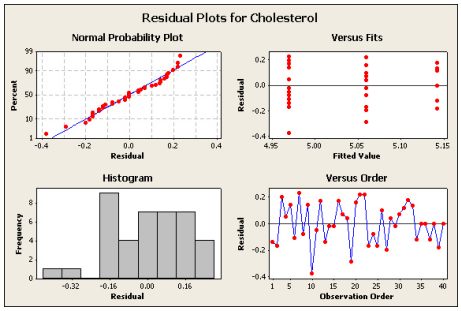

Figure 14: Checking our assumptions

- The Normal plot: points should lie roughly in a straight line
- Versus fit: points should be evenly distributed across treatments
- The Histogram: points should look like a Normal distribution
- Versus Order: points should be a random scatter

Since $p = 0.028 < 0.05$ we reject H_0 and accept H_A, i.e. grain does affect your cholesterol level.

To determine where the differences actually lie, we need to do a multiple comparison test. On the ANOVA menu, click Comparisons -> Tukey's -> OK. Now when you run the ANOVA command you should get additional output.

```
Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of C1

Individual confidence level = 98.04%

C1 = 0 subtracted from:
C1    Lower   Center    Upper   --+---------+---------+---------+-------
3   -0.2392  -0.0827   0.0738         (---------*----------)
6   -0.3250  -0.1726  -0.0202   (---------*----------)
                               --+---------+---------+---------+-------
                               -0.30     -0.15      0.00      0.15


C1 = 3 subtracted from:
C1    Lower   Center   Upper   --+---------+---------+---------+-------
6   -0.2282  -0.0899  0.0484         (--------*--------)
                               --+---------+---------+---------+-------
                               -0.30     -0.15      0.00      0.15
```

- Tukey's test looks for differences between treatments.
- So the row: `3  -0.2392  -0.0827  0.0738` suggests that the difference in mean cholesterol between 0 and 3 grains per day is in the range (-0.2392, 0.0738).
- We look for intervals that **do not** contain zero.
- So we can see that there are differences between 0 and 6 grains per day.

3.1.2 What are residuals?

The technical definition of residuals is:

Observed values - fitted values

What this means, is you compare what you think the value will be according to your model, against what actually happens.

3.2 Two–way ANOVA (randomised blocks design)

What is a Block? A block is usually an unwanted design aspect that you should not ignore. For example, different laboratories, different fields, different person who conducted the experiment.

Problem: The data below are the results of a randomised block design and give the yields (in lbs) of ten strains of carrots grown in four fields. Each field is a "block."

		Block				Treatment
		1	2	3	4	means
Treatment (strain of carrot)	A	30.0	28.0	28.3	31.7	29.5
	B	33.5	29.6	31.9	29.8	31.2
	C	32.8	29.0	31.2	25.8	29.7
	D	22.2	25.2	24.7	22.3	23.6
	E	30.2	33.1	30.1	27.8	30.3
	F	30.5	30.0	28.5	31.8	30.2
	G	32.6	31.7	30.1	30.8	31.3
	H	30.0	29.7	27.6	29.5	29.2
	I	36.7	35.7	35.6	32.4	35.1
	J	28.5	35.0	32.0	28.1	30.9
	Block means	30.7	30.7	30.0	29.0	30.1

In this analysis, not only have we got different treatments, but each treatment is carried out in different fields. We are trying to determine:

- Are the carrot strains different
- Do the fields affect the carrot yields

Analysis

1. We begin by plotting the data to get an overview of what is happening.
2. We form our hypotheses.
   - Our Null hypothesis is:

     H_0 : There are no differences between the means for different strains of carrot.

   - Our alternative hypothesis is:

     H_A : There are differences between the means for different carrot strains.
3. Using Stat -> ANOVA -> Two-way. The *response* is the yield. The *column factor* is a column indicating the treatment. The *row factor* is the column indicting the field. This gives

```
Two-way ANOVA: Yield (lbs) versus Carrot strain, block
Analysis of Variance for Yield}

Source      DF   SS       MS      F      P
Carrot str  9    287.68   31.96   8.34   0.000
block       3    19.40    6.47    1.69   0.193
Error       27   103.54   3.83
Total       39   410.62
```

4. Next we check the residuals graph, and make sure our data are normal.

3.3 Two-way ANOVA with two factors of interest

Problem: In some cases we are interested in both factors rather than one factor being a *nuisance factor* such as blocks.

Solution: We proceed exactly the same as above. The term **Block** just highlights the fact that we are not really interested in that effect.

3.4 Factorial experiments

It is possible to design and analyse much more complicated experiments with many factors and where we allow for the effect of *interactions* between factors. Experiments involving a number of factors are known as *factorial experiments*.

Practical 3

Instructions

1. You have an individual reference number. See the list on the last page of this handout.

   The reference numbers have changed since last week.

   You should use the data assigned to your reference number. Please write your reference number on your report as well as your name.

2. Please also mark on your report my name (Dr Malcolm Farrow) and "School of Mathematics & Statistics".

3. Answer both questions.

4. This assignment is to be submitted via NESS, no later than 12.00 noon on Wednesday 24th October.

5. Write each solution in the form of a (brief) report. This should have an introduction to the problem, a description of the analysis and a clear statement of conclusions, illustrated, where appropriate, with graphs and tables.

6. Graphs should be properly labelled with appropriate axis labels etc.

7. In your report explain (briefly) how you obtained your results using Minitab so that, if anything has gone wrong, I might be able to see where it is.

Questions

1. Data are provided on butterfat percentages in milk from random samples of mature (≥ 5 years) and 2-year-old cows of five breeds.

   The data file `butterdat.txt` can be downloaded from the module Web page (via Blackboard or otherwise) or it can be read directly into `Minitab` using

   ```
   File -> Other files -> Import special text
   ```

   entering `c9-c40` in the box marked `Store data in column(s)` and entering

   ```
   http://www.mas.ncl.ac.uk/~nmf16/teaching/ace2046/butterdat.txt
   ```

   for the file name.

   The breeds are in column `c10`. Name this column `Breed` or some other suitable name. The breeds are coded as follows.

   1 Ayrshire
   2 Canadian
   3 Guernsey
   4 Holstein-Fresian
   5 Jersey

   Column `c9` contains 1 for a mature cow and 2 for a 2-year-old cow. Name this column `Age` or some other suitable name.

   Your butterfat percentages are in the column corresponding to your reference number. For example, if your reference number is 20 then your data are in column `c20`. Name your column `Butterfat` (or some other suitable name).

(a) Check that it is reasonable to assume equal variances. Try using a boxplot. Use

`Graph -> Boxplot -> One Y With groups`

Use your `Butterfat` column for the `Graph variable` and put both `Breed` and `Age` in the `categorical variable for grouping` box.

Do you think the spreads look reasonably equal?

(b) We have 10 groups (one for each combination of age and breed). This makes it a bit difficult to use a normal probability plot to check for normality. However we can apply the check to the residuals later.

(c) Use a two-way analysis of variance to test the null hypotheses that the mean butterfat percentage does not depend on the breed, for a given age, and that it does not depend on the age for a given breed.

Use

`Stat -> ANOVA -> Two-way`

- The `Response` is your `Butterfat` column.
- The `Row Factor` is `Age` .
- The `Column Factor` is `Breed` .
- In both cases tick `Display means`.
- Tick `Fit additive model`.
- Select

  `Graphs -> Four in one`

  to produce residual plots to check the assumptions.

Your ANOVA table will give two tests, one for the null hypothesis of equality between breeds and one for the null hypothesis of equality between ages.

(d) Comment on the residual plots. Is there any suggestion that the assumptions were not valid?

(e) Report your conclusions clearly.

2. Data are provided on weight gains in rats fed on four different diets distinguished by amount of protein (low or high) and by source of protein (beef or cereal).

The data file `weightdat.txt` can be downloaded from the module Web page (via Blackboard or otherwise) or it can be read directly into `Minitab` using

`File -> Other files -> Import special text`

entering `c9-c40` in the box marked `Store data in column(s)` and entering

`http://www.mas.ncl.ac.uk/~nmf16/teaching/ace2046/weightdat.txt`

for the file name.

The sources are in column `c9` and the amounts are in column `c10` . However, to improve the presentation of the results we will recode these as follows.

Select

`Data -> Code -> Numeric to text`

- In the box `Code data from columns` enter `c9` .
- In the box `Store data in columns` enter `c7` .
- Under `Original values` enter 1 in the first row and 2 in the second row.
- Under `New` enter `Beef` in the first row and `Cereal` in the second row.
- Click `OK` .

This recodes the 1s and 2s for Source to `Beef` and `Cereal` respectively. Give column `c7` the name `Source` .

Use a similar procedure to recode the amounts.

Select

```
Data -> Code -> Numeric to text
```

- In the box `Code data from columns` enter `c10` .
- In the box `Store data in columns` enter `c8` .
- Under `Original values` enter 1 in the first row and 2 in the second row.
- Under `New` enter `Low` in the first row and `High` in the second row.
- Click `OK` .

This recodes the 1s and 2s for Amount to `Low` and `High` respectively. Give column `c8` the name `Amount` .

Your weight gains are in the column corresponding to your reference number. For example, if your reference number is 20 then your data are in column `c20`. Name your column `Weight gain` (or some other suitable name).

(a) Check that it is reasonable to assume equal variances. Try using a boxplot. Use

```
Graph -> Boxplot -> One Y With groups
```

Use your `Weight gain` column for the `Graph variable` and put both `Source` and `Amount` in the `categorical variable for grouping` box.

Do you think the spreads look reasonably equal?

(b) Use normal probability plots to check for normality.

- Select
  `Graph -> Probability Plot -> Multiple`
- In the `Graph variables` box enter `Weight gain` .
- Select
  `Multiple graphs -> By variables`
- Put both `Source` and `Amount` in the box named
  `By variables with groups in separate panels` .

(c) Use a two-way analysis of variance to test the null hypotheses that the mean weight gain does not depend on the amount of protein, for a given source, and that it does not depend on the source of protein for a given amount.

Use

```
Stat -> ANOVA -> Two-way
```

- The `Response` is your `Weight gain` column.
- The `Row Factor` is `Source` .
- The `Column Factor` is `Amount` .
- In both cases tick `Display means`.
- Tick `Fit additive model`.
- Select
  `Graphs -> Four in one`
  to produce residual plots to check the assumptions.

Your ANOVA table will give two tests, one for the null hypothesis of equality between sources and one for the null hypothesis of equality between amounts.

(d) Comment on the residual plots. Is there any suggestion that the assumptions were not valid?

(e) Report your conclusions clearly.

(f) Try repeating the analysis but this time do not tick `Fit additive model` . This allows the fitting of an "Interaction" term. This means that the effect of moving from "Low" to "High" amount might depend on which source is used. You need not include all of the results of this in your report but comment on any differences which you notice compared with the previous "additive" analysis.

Student Reference Numbers

In the practicals for this week and later weeks, in at least some of the exercises, each student will have different data. In order to identify which data each student should have, each student is given a reference number according to the table below. Please make a note of your reference number because we will use the same numbers in later weeks. **Please note that the numbers have changed since last week.**

Agapiou	Kristy Marie	11
Au	Hoi Ching	12
Begin	Abigail Ruth	13
Breininger	Stella Panagio	14
Cepelis	Aivaras	15
Choo	Siew Li	16
Clarke	Timothy Oliver	17
Crawley	Michael	18
Darwin	Darwin	19
Drayton	Chloe Roseanne	20
Faleti	Hannah Ifedolapo	21
Golding	Carolina De Arouca	22
Grant	Georgia Kate	23
Hui	Angeline Kosim	24
Jenkins	Grace Evangeline	25
Linn	Sophia Campbell	26
Lo	Kin Wa	27
MacKinlay	Brittney Lauren	28
Obertelli	Georgia May	29
Rathod	Neel Harish	30
Rowland	Maisie Katharine	31
Saberioon	Ghazaleh	32
Song	Anqi	33
Tam	Wai Tsun Rachel	34
Trikerioti Chatziioannou	Danai	35
Wright	Jessica Laura Kate	36
Zawarska	Santana Anna	37