PhD. course in Basic Biostatistics – Day 6

Erik Parner, Department of Biostatistics, Aarhus University©

Multiple comparisons: does bran make the man?

Comparing k independent normal samples
- multiple comparison analysis
- multiple linear regression

The CI of an estimated standard deviation

Comparing k independent normal samples
- one-way analysis of variance (F-test, the Bartlett test for variance homogeneity)

The Kruskal-Wallis non-parametric test

Multiple comparisons

An example of analysis of repeated measurements by one-way ANOVA.

Overview

Data to analyse	Type of analysis	Unpaired/Paired	Type	Day
Continuous	One sample mean	Irrelevant	Parametric	Day 1
			Nonparametric	Day 3
	Two sample mean	Non-paired	Parametric	Day 2
			Nonparametric	Day 2
		Paired	Parametric	Day 3
			Nonparametric	Day 3
	Regression	Non-paired	Parametric	Day 5
	Several means	**Non-paired**	**Parametric**	**Day 6**
			Nonparametric	**Day 6**
Binary	One sample mean	Irrelevant	Parametric	Day 4
	Two sample mean	Non-paired	Parametric	Day 4
		Paired	Parametric	Day 4
	Regression	Non-paired	Parametric	Day 7
Time to event	One sample: Cumulative risk	Irrelevant	Nonparametric	Day 8
	Regression: Rate/hazard ratio	Non-paired	Semi-parametric	Day 8

Multiple comparisons: does bran make the man?

Bran is the hard outer layers of cereal grain.

Can eating breakfast cereal determine the sex of your baby?

The original study, "You Are What Your Mother Eats", by Mathews et al (2008) made headlines around the world. Researchers at Exeter and Oxford universities asked 740 pregnant women to record what they ate during pregnancy and just before.

The authors wrote in their conclusion "Over the past 40 years, there have been small, but highly consistent, declines in the proportion of male infants born in industrialized countries… However, population-level changes in the diets of young women may explain the pattern… At the same time, there is good evidence that the prevalence of breakfast skipping is increasing."

Multiple comparisons: does bran make the man?

How did the authors arrive at that conclusion?

"We went on to test whether particular foods were associated with infant sex. Data of the **133 food items** from our food frequency questionnaire were analysed, and we also performed additional analyses using broader food groups. Prior to pregnancy, **breakfast cereal**, but no other item, was strongly associated with infant sex (Wald $\chi 2$ =8.2, p=0.004). Women producing male infants consumed more breakfast cereal than those with female infants. The odds ratio for a male infant was 1.87 (95% CI 1.31, 2.65) for women who consumed at least one bowl of breakfast cereal daily compared with those who ate less than or equal to one bowlful per week. No other foods were significantly associated with infant sex (given the multiplicity of testing, p≤0.01 was considered significant)"

Multiple comparisons: does bran make the man?

So 133 statistical tests and one p-values less than 0.01!

Suppose, there were no association between any of the food items and the sex of the child, **and** all statistical tests were independent, then we would on average expect 133*0.01=1.33 statistical significant findings (type 1 error).

Hence the association between breakfast cereal and the sex of the child may very well be a type 1 error.

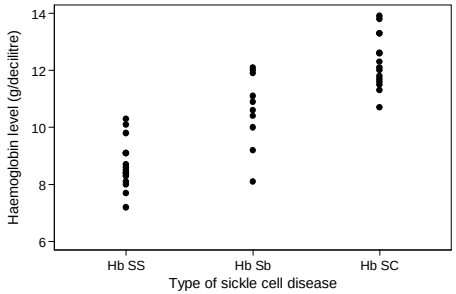
Today we will consider different way of handling many comparison with the focus on normally distributed outcomes.

Comparing k independent normal samples
Example 9.1: Haemoglobin level and sickle cell disease

Question: How does the haemoglobin level differ between persons who suffer from three different sickle cell diseases (Hb SS, Hb Sβ or Hb SC).

Data: Haemoglobin levels for 41 patients.

Figure 6.1



Notation: Let y_{ij} denote the haemoglobin level for the jth patient in the ith group, e.g. y_{25} is the level for patients no 5 suffering of HB Sβ sickle cell disease.

And let n_i denote the number of patients in the ith group.

Model 0: **Three independent** samples from three normal distributions. Comparable to Day 2.

Let μ_1, μ_2, μ_3 denote the means and $\sigma_1, \sigma_2, \sigma_3$ the standard deviations.

A summary of the data:

Type	n	average	CI low	CI upp	sd	Se	97.5 t-percentile
Hb SS	16	8.713	8.263	9.162	0.8445	0.2111	2.131
HB Sβ	10	10.630	9.711	11.549	1.2841	0.4061	2.262
Hb SC	15	12.300	11.778	12.822	0.9419	0.2432	2.145

Remember the 95% CI's are found as:

$$\overline{y}_i \pm t_{.975}\,(n_i - 1) \cdot sd_i \Big/ \sqrt{n_i}$$

The $n_i - 1$ degrees of freedom reflects that the estimate of the standard deviation is based on n_i **observations minus one estimated mean.**

Comparing k independent normal samples multiple comparison analysis

We are interested in the hypothesis

H_2: $\mu_1 = \mu_2 = \mu_3$

We could use the unpaired t-test to test H by comparing and testing

H_{2A} : $\mu_1 = \mu_2$

H_{2B} : $\mu_1 = \mu_3$

H_{2C} : $\mu_2 = \mu_3$

Result:

Group 1 versus 2: -1.91 (95%CI: -2.78 ; -1.06), p<0.001

Group 1 versus 3: -3.59 (95%CI: -4.24;-2.93), p<0.001

Group 2 versus 3: -1.67 (95%CI: -2.59; -0.75), p=0.001

The **problem** with this analysis is that the overall hypothesis

H_2: $\mu_1 = \mu_2 = \mu_3$

is evaluated with 3 separate analysis each with a Type 1 error of 5%.

Thus the chance that is H_2 is significant if either of H_{2A} , H_{2B} , H_{2C} is significant will occur more often than 5% when there is no difference in the means, i.e. test procedure will have a higher type 1 error than 5%!

This is the main problem of **multiple comparisons (testing)**.

One overall test of the hypothesis H_2 is desired.

Comparing k independent normal samples multiple linear regression

An overall test of no difference in the means can be performed in a multiple linear regression analysis.

Define two dummy variable

Type2 1 if type=2 and 0 otherwise.

Type3 1 if type=3 and 0 otherwise.

The mean of model 0 may then be formulated as

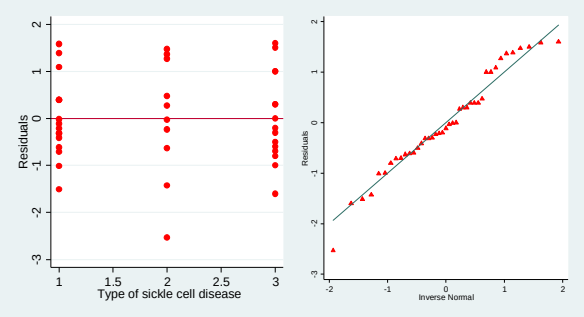
Mean haemoglobin level = $\beta_1 + \beta_2 \cdot$Type2$+ \beta_3 \cdot$Type3

β_1 is the mean haemoglobin level in group 1 (Hb SS)

β_2 is the difference in mean haemoglobin level between group 2 and 1 (Hb Sβ versus Hb SS)

β_3 is the difference in mean haemoglobin level between group 3 and 1 (Hb SC versus Hb SS)

Checking the multiple regression model:



The variation of the residuals seem contant between groups.

Overall, both residual plots looks OK. (A more refined model check appears later.)

In the regression model we may test the hypothesis

H_2: $\mu_1 = \mu_2 = \mu_3$

by testing

H_2: $\beta_2 = \beta_3 = 0$

Stata: Comparing k independent normal samples - multiple linear regression

```
. use haemoglob.dta, clear
(Haemoglobin level and sickle cell disease.)
. regress haemo ib1.type

  Source |       SS     df       MS         Number of obs = 41
---------+------------------------------      F( 2, 38) =   50.00
   Model |  99.8893064   2  49.9446532      Prob > F  =  0.0000
Residual |  37.9585029  38  .998907972      R-squared =  0.7246
---------+------------------------------  Adj R-squared =  0.7101
   Total |  137.847809  40  3.44619523      Root MSE  =  .99945

------------------------------------------------------------------------
  haemo |  Coef.  Std. Err.      t    P>|t|  [95% Conf.Interval]
--------+---------------------------------------------------------------
   type |
 Hb Sb  | 1.9175  .4028927     4.76   0.000   1.101886  2.733114
 Hb SC  | 3.5875  .3592014     9.99   0.000   2.860335  4.314665
        |
  _cons | 8.7125  .2498635    34.87   0.000   8.206678  9.218322
------------------------------------------------------------------------
```


Stata: Comparing k independent normal samples - multiple linear regression

```
. test 2.type 3.type
 ( 1)  2.type = 0
 ( 2)  3.type = 0
       F(  2,    38) =   50.00
            Prob > F =    0.0000

. predict fit if e(sample), xb
. predict res if e(sample), res
. scatter res type
. qnorm res
```

Results:

		Means	CI low	CI upp	P-value
Type					<0.001
	HB SS	8.713	8.207	9.218	
	Hb Sβ - HB SS	1.918	1.102	2.733	
	HB SC - Hb SS	3.588	2.860	4.315	

Stata: Comparing k independent normal samples - multiple linear regression

Note:

The mean values in the first group based on the linear regression analysis:
8.7125 (95% CI: 8.206678-9.218322)

If we had analyzed the first group using a one-sample normal analysis (Day 1) we could derive
8.7125 (95% CI: 8.262502-9.162498)

Both CI are exact.

The first CI is based on a assumption of the same variance in the three groups. The common variance is estimated as weighted estimates as in Day 2.

The precision of an estimated standard deviation - the 95% CI for σ

Recall from Day 1: The precision of such an estimate is given by the degrees of freedom, df, which in general is the **number of observations** minus the number of **unknown parameters** describing the mean.

Finding the 95% CI for σ is a bit complicated, as it involves the upper and lower 2.5 percentile in a chi-squared distribution with df degrees of freedom:

$$\hat{\sigma}\cdot\sqrt{\frac{df}{\chi^2_{df}(0.975)}} \le \sigma \le \hat{\sigma}\cdot\sqrt{\frac{df}{\chi^2_{df}(0.025)}}$$

$$\hat{\sigma}\cdot l(df) \le \sigma \le \hat{\sigma}\cdot u(df)$$

df	$l(df)$	$u(df)$
5	0.624	2.453
10	0.699	1.755
15	0.739	1.548
20	0.765	1.444
25	0.784	1.380
50	0.837	1.243
150	0.899	1.128
200	0.911	1.109

The precision of an estimated standard deviation - the 95% CI for σ

But in Stata you can do these calculation "by hand".
Ex: From the linear regression had df=38 and sd = σ= 0.99945:

```
display 0.99945*sqrt(38/invchi2(38,0.975))
display 0.99945*sqrt(38/invchi2(38,0.025))
```

Giving CI(σ): $0.8<\sigma<1.3$

One-way analysis of variance

A **one-way ANalysis Of VAriance** (one-way ANOVA) will also give **one test/p-value** of the hypothesis:

H_2: $\mu_1 = \mu_2 = \mu_3$

It is another way of deriving the same p-value as in the multiple linear regression model.

The idea behind an analysis of variance is to divide the variation in the data, the y's, into **different sources**.

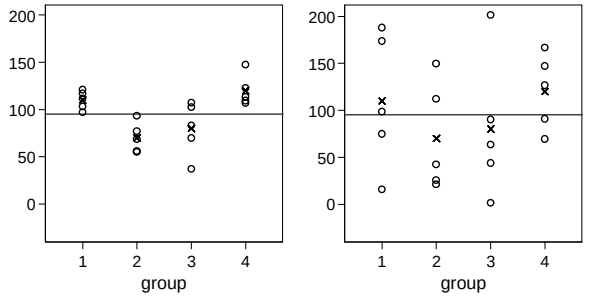
In a one-way ANOVA there are two sources:
within groups and **between** groups

$$\sum_{i,j}\left(y_{ij}-\bar{y}\right)^2 = \sum_{i,j}\left(y_{ij}-\bar{y}_i\right)^2 + \sum_i n_i\cdot\left(\bar{y}_i-\bar{y}\right)^2$$

$$SS_{Total} = SS_W + SS_B$$

total variation=within group variation+between group variation

One-way analysis of variance Two illustrations



SS_B	8,500	8,500
SS_W	5,682	62,574
SS_T	14,182	71,074

A clear difference between the groups on the left plot!

One-way analysis of variance

The hypothesis: H_2: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ is tested by comparing the scaled variations between and within groups by calculating:

$$F = \frac{SS_B/(k-1)}{SS_W/(n-k)} = \frac{\sum_{i=1}^{k} n_i \cdot (\bar{y}_i - \bar{y})^2 \Big/ (k-1)}{\hat{\sigma}_W^2}$$

If the group averages differ much, then you will get a large value of F, i.e. a large F value is critical for the hypothesis.

The p-value is found in an F-distrib. with (k-1) and (n-k) df's.

In the two examples on the previous slide we get:

$$F_{Left} = \frac{8{,}500/(4-1)}{5{,}682/(20-4)} = 7.98 \qquad F_{Right} = \frac{8{,}500/(4-1)}{62{,}574/(20-4)} = 0.72$$

$$p_{Left}=0.18\% \qquad p_{Right}=55\%$$

Stata: one way ANOVA (and Bartlett test)

```
. oneway haemo type

                    Analysis of Variance
    Source          SS       df      MS          F     Prob > F
------------------------------------------------------------------------
Between groups   99.8893064    2   49.9446532   50.00     0.0000
 Within groups   37.9585029   38   .998907972
------------------------------------------------------------------------
    Total        137.847809   40   3.44619523

Bartlett's test for equal variances:  chi2(2) =   2.1251
Prob>chi2 = 0.346
```

The test for equal variances will be discussed in a moment.

Stata: one way ANOVA

number of groups -1

Number of obs - number of groups

$\frac{SS}{df}$

```
Analysis of Variance
    Source              SS         df      MS          F     Prob > F
------------------------------------------------------------------------
Between groups      99.8893064      2   49.9446532   50.00     0.0000
 Within groups      37.9585029     38   .9989080
------------------------------------------------------------------------
    Total           137.847809     40   3.44619523
```

$\hat{\sigma}_W^2$

number of obs -1

The F-test for the hypothesis of no difference in the means. Here we reject the hypothesis!

Testing for identical standard deviations = Test of variance homogeneity

On Day 2 you learned how to test the hypothesis that two standard deviations were identical, by an F-test.

Here we want to test that **several** standard deviations are identical: $H_1: \quad \sigma_1 = \sigma_2 = \sigma_3$

Several tests can do this. The best is

Bartlett's test for equal variances

This test is not covered by Kirkwood & Sterne and we will leave out the technical details.

Another valid test is **Levene's** for equal variances.

Note, these tests will not focus on a special pattern in the deviations from the hypothesis. Hence you should yourself look out for the **most common deviation,** where the standard deviations **increase with the means**.

Checking the model behind the one-way ANOVA

Assumptions:

1. Independence **between** groups and independent observations **within** each group.
2. Normal distribution **within** each group.
3. The same standard deviation in all the groups.

As before 1. is checked by going through the design.

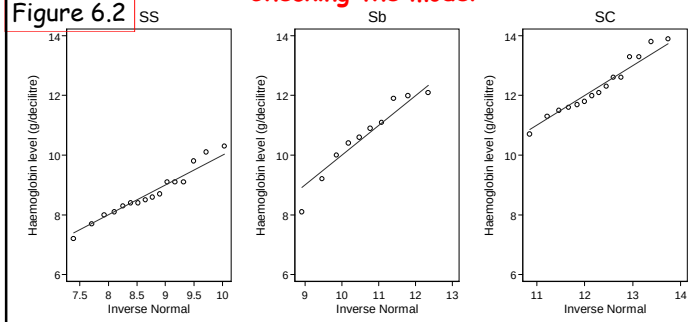
And 2. by QQ-plots within each group. Here you should look out for the same type of deviations from normality in all groups.

The last assumption can be checked by Bartlett's test.

If the data consist of **many small groups**, then normality is best checked by a **QQ-plot of the residuals**.

Haemoglobin level and sickle cell disease
Checking the model

Figure 6.2



Normality ok.

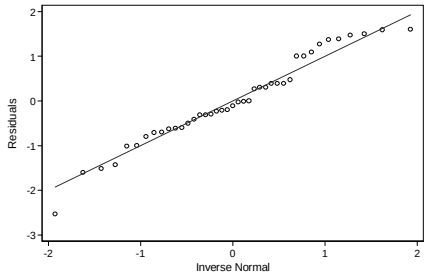
Bartlett's test from Stata: chi2(2) = 2.1251 **p= 34.6%**
We can accept the hypothesis of identical standard deviations

Haemoglobin level and sickle cell disease
Checking the model

Each of the QQ-plots on previous slide, was based on a relatively few observations and it is a good idea to supply these plots with a **QQ plot of the residuals under the model:**

$$r_{ij} = y_{ij} - \overline{y}_i$$

Figure 6.3



This also look ok.

Haemoglobin level and sickle cell disease - formulations

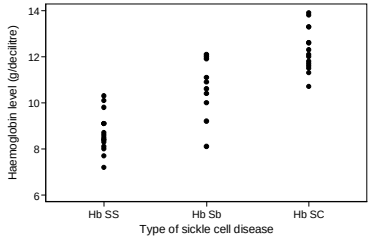
Methods
The three groups were analyzed by a one-way analysis of variance. Normality was checked by QQ-plots and the of assumption variance homogeneity by Bartlett's test. Means and differences between means are given with 95% confidence intervals.

Results
The mean haemoglobin levels are shown in a table on the next slide. There was a highly significant difference between the groups ($p<0.001$). Patients suffering from type SS had mean haemoglobin levels 1.9(1.1;2.7) and 3.6(;2.9;4.3) lower than patients suffering from type Sβ and SC, respectively. The mean difference between the two latter was 1.7(.8;2.5).

Conclusion
??????????????

Haemoglobin level and sickle cell disease - formulations

	Estimate	CI low	CI upp
Means			
Hb SS	8.7	8.2	9.2
HB Sβ	10.6	10.0	11.3
Hb SC	12.3	11.8	12.8
Differences			
Hb Sβ - HB SS	1.9	1.1	2.7
Hb SC - HB SS	3.6	2.9	4.3
HB SC - Hb Sβ	1.7	0.8	2.5



An overview of the models

Model 0: Independent observations and

$$y_{ij} = \mu_i + E_{ij} \qquad E_{ij} \sim N\left(0, \sigma_i^2\right)$$

H_1 : $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_W$ Bartlett´s test

Model 1: Independent observations and

$$y_{ij} = \mu_i + E_{ij} \qquad E_{ij} \sim N\left(0, \sigma_W^2\right)$$

H_2 : $\mu_1 = \mu_2 = \mu_3 = \mu$ Oneway ANOVA : F-test

A Non-parametric comparison of several groups The Kruskal-Wallis test

In the two sample setting we could test the assumption of no systematic difference between the two groups by a Wilcoxon-Mann-Whitney rank sum test.

The rank based non-parametric test comparing k groups is the **Kruskall–Wallis** test, which is a **scaled version of**

$$\sum_{i=1}^{k} n_i \cdot \left(\overline{R}_i - \overline{R}\right)^2$$

$\overline{R}_i$ is the average rank in group i and

$\overline{R}$ the overall average rank.

Large values are of course critical.

Some comments to the one-way ANOVA and the Kruskal-Wallis test

If we only have **two groups**, then the one-way ANOVA F-test is **equivalent** to the unpaired t-test.

If we only have **two groups**, then Kruskal-Wallis test is **equivalent** to the Wilcoxon-Mann-Whitney test.

If there exists an ordering of the groups it will **not be noted** by either of the two methods.

E.g. the test will have low power to detect, if the groups consist of an in**creasing dose** of a drug.

If this is the case one should turn to **regression models**.

Multiple comparison

The multiple regression model and the one-way ANOVA F-test (or the Kruskall-Wallis-test) supplies you with **one** p-value for the hypothesis of no difference (in means) between the groups.

That is, the multiple regression model and the one-way ANOVA is **only relevant** in the situation where non of the pairwise comparisons are **a priori of special interest**.

In this situation both analyses **prevents a fishing expedition** among the $\frac{1}{2}k(k-1)$ pair wise comparisons.

Remember, even if there is **no difference** between the k groups, then for each pairwise t-test there is a **5% chance** of getting a 'statistically significant result'.

Multiple comparison

If the overall comparison the multiple regression model or the one-way ANOVA is significant then the pairwise comparison is allowed to discover the difference.

In general, if a priori some comparisons are of special interest (e.g. treatments versus control), **then you should focus on this**!!! Se an example of hierarchical testing on the next slide.

The multiple comparison/testing problems has had a lot of focus in the medical/statistical literature. It has given rise to **many procedures/algorithms** that will assure that the risk of type 1 error is 5%.

The most common are the Bonferroni and hierarchical testing.

Multiple comparison

Bonferroni: K = number of tests.
Overall significance level α.
Each test is tested on level: $\alpha^* = \alpha/K$
Example.
Pairwise comparison of three groups: 3 tests.
$\alpha^* = 0.05/3 = 0.017$

Hierarchical testing: Let H_1 and H_2 be two hypotheses of interest, for example a high and low doses of a drug compared to control. H_1 is considered the most important and H_2 is of interest once H_1 has been rejected. The following hierarchical test procedure will main a overall level of 5%:

H_1 is always tested on level α, but H_2 is only tested level α if H_1 is statistical significant.

Multiple comparison

Comments

The use of one-way ANOVA before all pairwise comparisons controls the overall significance level (**familywise error rate**, FWER) in a **weak sense**: the type 1 error is 5% when all pairwise hypothesis are true, i.e. when all means are equal.

A procedure controls the FWER in the **strong sense** if the FWER control at level 5% is guaranteed for any configuration of true and non-true null hypotheses (including the global null hypothesis).

The Bonferoni controls the FWER in the strong sense.

Hierarchical controls the FWER in the strong sense provided the hypothesis are ordered with respect seriousness.

Repeated measures

Multiple comparison issues arise often if the data consists of repeated measure of the same individual.

Repeated measure of continuous data are often analyzed using analysis of variance methods.

In some cases an analysis of a summary measure will be sufficient. Here we will consider one such application.

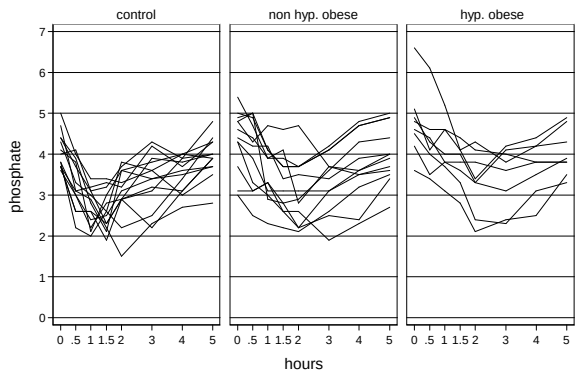
Plasma phosphate levels in three groups of subjects
Measurements over time

Aim
Evaluate the association between hyperglycemia and relative hyperinsulinemia.

Design
The plasma inorganic phosphate level was measured 0, ½, 1, 1½, 2, 3, 4 and 5 hours after a standard-dose oral glucose challenge in 13 controls, 12 non-hyperinsulinemic obese and 8 hyperinsulinemic obese persons.

Comments
This is a **repeated measurements design** with 8 measurements on each subject.
A full analysis of such a design is outside the scope of the course, but it can to some extend be handled by what we know.

Plasma phosphate levels in three groups of subjects



Always plot the data !

Plasma phosphate levels in three groups of subjects

One (and often the best) way to analyze curves like these is to summarize the curve **for each subject** into a single observation – **a summary measure**.

The choice of the summary measure **should** of course be based on the **prior knowledge** and **focus of the study**.

The summary measure should be decided **before looking at the data**.

There are of course **many options**: the average phosphate level, minimum, the maximum, the area under the curve (=the time averaged mean), the slope after fitting a line, etc.

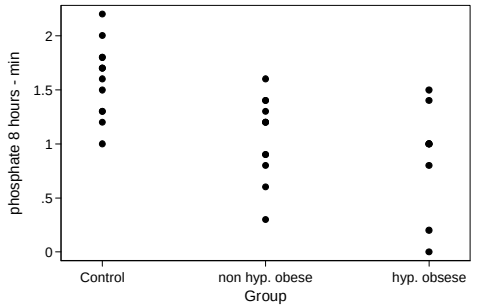
Sometimes you might calculate **two or three summary** measures for each subject.

Here we look at **increase** = level at 5 hours – minimum level

Plasma phosphate levels in three groups of subjects

increase = level at 5 hours – minimum level

In the three groups:



These data can now be analyzed as three independent normal samples.

Plasma phosphate levels in three groups of subjects

	Estimate	CI low	CI upp
Means			
Controls(n=13)	1.60	1.37	1.83
Non-hyp. obese(n=12)	1.07	0.83	1.30
Hyp. obese(n=8)	0.86	0.57	1.15
Differences			
Hyp. obese - controls	-0.74	-1.11	-0.37
Hyp. - non-hyp obese	-0.20	-0.58	0.17
Non-hyp. obese - controls	-0.53	-0.86	-0.21

one-way ANOVA F(2,30)=9.89 p=0.0005

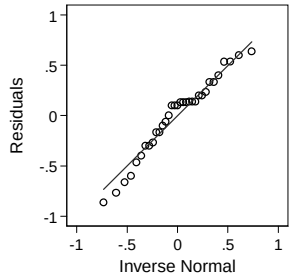
Conclusion

???

Plasma phosphate levels in three groups of subjects checking the model

group	Summary of phosphate 8 hours - min Mean	Std. Dev.	Freq.
control	1.6	.33416564	13
non hyp.	1.0666667	.37497474	12
hyp. obe	.8625	.52627396	8

Bartlett's test: chi2(2) = 1.9618 p = 0.375



No serious differences between the standard deviations.
QQ-plot of the residuals looks ok.