Statistical Science.

Part III. The General Linear Model.

Chapter 10.2 Two Sample t-test

ReCap.	Part I (Chapters 1,2,3,4)
ReCap	Part II (Ch 5, 6, 7)
ReCap	Part III (Ch 9)
10.1	Single Sample t-test
10.2	Two Sample t-test
10.3	One way ANOVA, Fixed Effects
10.4	One way ANOVA, Random Effects

Ch10.xls
Sleep data from Cushny and Peebles
Daphnia ages from Sokal and Rohlf (2012) Table 9.2

on chalk board

ReCap Part I (Chapters 1,2,3,4)
Quantitative reasoning: Example of scallops,
which combined models (what is the relation of scallop density to substrate?)
with statistics (how certain can we be?)

ReCap Part II (Chapters 5,6,7)
Hypothesis testing uses the logic of the null hypothesis to make a decision about an unknown population parameter.
Estimation is concerned with the value of an unknown parameter.

ReCap (Ch 9) The General Linear Model is more flexible and useful than a collection of named tests.
Regression is a special case of the GLM. We have seen an examples with the explanatory variable X fixed, with the explanatory measured with error.

Today:
Two-sample t-test as a special case of the GLM

Wrap-up
ANOVA is a special case of the general linear model..
The explanatory variable consists of categories, which are on a nominal scale.
A *t*-test contrasts two means. It is a special case of one-way (single factor) ANOVA.

GLM. Unpaired *t*-test ANOVA with two categories.
Example. Sleep data.

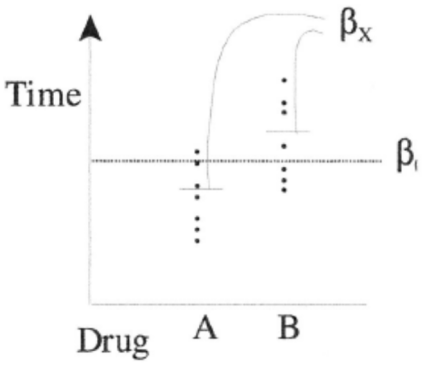
The example will be hours of extra sleep, in two drugs, hyoscyamine (DrugA) and hyoscine-L (DrugB).
Data from
Cushny AR, Peebles AR (1905). The action of optical isomers. II. Hyoscines. J Physiology 32:501-510.

Drugs A and B were administered to 10 patients in a mental hospital. Quoting from the publication "As a general rule a tablet was given on each alternate evening, and the duration of sleep and other features noted and compared with those of the intervening control night on which no hypnotic was given."

```
0.7        1.9
-1.6       0.8
-0.2       1.1
-1.2       0.1
-0.1       -0.1
 3.4       4.4
 3.7       5.5
 0.8       1.6
 0.0       4.6
 2.0       3.4

DrugA DrugB
Cushny.dat
```

1. Construct model

Verbal model: Extra time slept depends on drug.
Graphical model: Comparison of two means.



The verbal and graphical models help us distinguish response from explanatory variables.

The quantity of interest, hours of extra sleep, depends on the explanatory variable, Drug A or B.
Response variable: Hours of extra sleep *T*
Also called the dependent variable.
It has units of hours, it is on a nominal scale.
Explanatory variable: *Drug* (= *A* or *B*)
Drug is a categorical variable. It is on a nominal scale

List variables: role (response/explanatory), name, symbol, units, and type of scale.

	Variable name	Symbol	Units	Scale
Response variable	Hours of extra sleep	*T*	hours	ratio
Explanatory variables	Drug	*Drug*		nominal

State the verbal model using names of quantities and then using symbols
"Hours of extra sleep depend on drug type"
Hours of extra sleep *T* = f(Drug type).

1. Construct model

Formal Model Now write the formal model, which is what the statistical package will use to carry out the analysis. Here it is in generic notation, then in an equivalent notation specific to the data.

$T = \beta_o + \beta_x \cdot x + \varepsilon$ generic notation

β_o is the grand mean

β_x is the difference between β_o and the mean of Drug *A* and of Drug *B*

$T = \beta_o + \beta_{Drug} \cdot Drug + \varepsilon$ equivalent notation

β_o is the intercept, the mean for the first drug (Drug *A*)

β_{Drug} is the contrast (difference) between Drug *A* and *B*

$\beta_0 + \beta_{Drug}$ = mean of Drug *B*

2. Execute analysis.

Data are often displayed by category, as above.

We reorganize the data to model format - - >

1 column with response variable, extra sleep time *T*.

1 column with explanatory variable, *Drug* = A or B

T	Drug
0.7	A
-1.6	A
-0.2	A
-1.2	A
-0.1	A
3.4	A
3.7	A
0.8	A
0.0	A
2.0	A
1.9	B
0.8	B
1.1	B
0.1	B
-0.1	B
4.4	B
5.5	B
1.6	B
4.6	B
3.4	B

Use the formal model to code the analysis in a statistical package.

$$T = \beta_o + \beta_{Drug} \cdot Drug + \epsilon$$

```
MTB> ANOVA 'T' = 'Drug'
MTB> GLM 'T' = 'Drug'
```

If you are using a graphics interface statistical package to run the analysis, be sure to look at the code produced, so that you understand how the model you wrote translates into a model statement in your package.

Run the stat package to obtain fitted (expected) values and residuals from model parameters.

Fitted values Fits = E[*T*] = $\hat{\beta}_o + \hat{\beta}_{Drug} \cdot Drug$

Residuals: Res = *T* – Fits

Here are the parameter estimates

```
MTB > describe 'hrs'; by 'drug'
              drug        N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
 hrs             1       10    0.750    0.350    0.675    1.789    0.566
                 2       10    2.330    1.750    2.237    2.002    0.633
```

Residual and fitted values are calculated from the parameter estimates.

$\hat{\beta}_0 = 0.75$

$\hat{\beta}_{Drug} = 1.58$

$\hat{\beta}_0 + \hat{\beta}_{Drug} = 2.33$

$\bar{T}_A = 0.75$

$\bar{T}_B = 2.33$

2. Execute analysis.

GLM: $T = \beta_o + \beta_{Drug} \cdot Drug + \varepsilon$

```
MTB > GLM 'Time' =          'Drug';
SUBC> fits c3;
SUBC> res c4.
MTB > plot c4*c3
```

GLM routines produce residuals and fits as output.

Here is the model statement in R with code to produce graphical output.

GLM: $T = \beta_o + \beta_{Drug} \times Drug + \varepsilon$

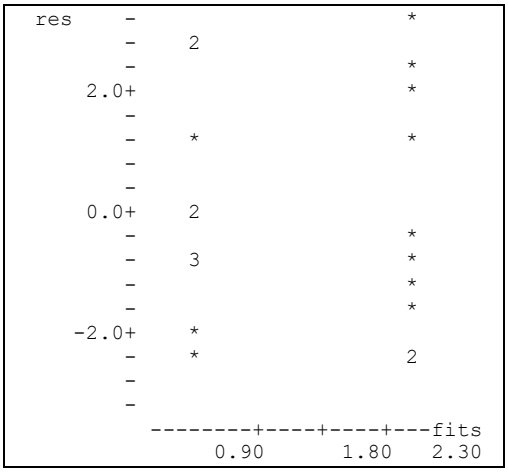
```
> Sleepmodel <- lm(T ~ Drug, data=Cushny)
> plot(Sleepmodel)
```

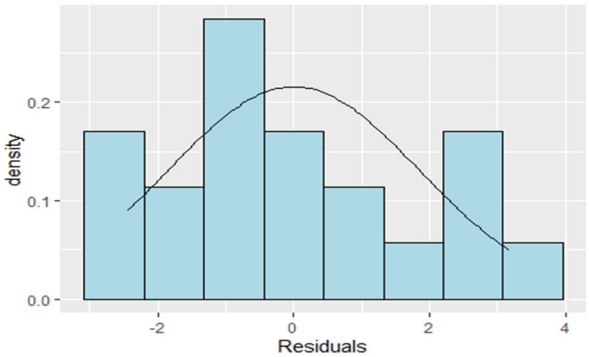
Here is a partial print-out of the data equations showing the residuals. Residuals here are offset by 1 for graphical evaluation of independence of the residuals.

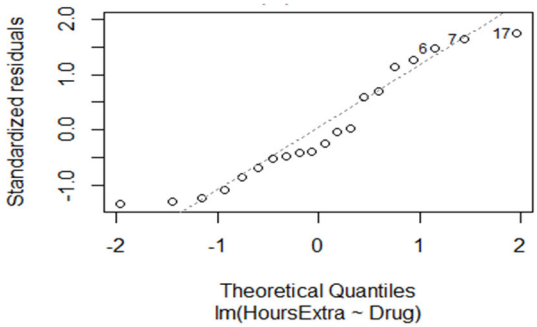
```
MTB > print 'T' 'Drug'  'fits' 'res'

 ROW   T    Drug   fits   res   lag1(res)

   1   0.7    A    0.75  -0.05
   2  -1.6    A    0.75  -2.35   -0.05
   9   0.0    A    0.75  -0.75    0.05
  10   2.0    A    0.75   1.25   -0.75

  11   1.9    B    2.33  -0.43    1.25
  19   4.6    B    2.33   2.27   -0.73
  20   3.4    B    2.33   1.07    2.27
```

3a. Evaluate the structural model.
No slopes (straight lines) used, so no straight line assumption to be checked.

3b. Evaluate the probability model (Normal distribution in this case).
This is especially important when sample size *n* is small (less than 30 or so)
Plot residuals vs fits.



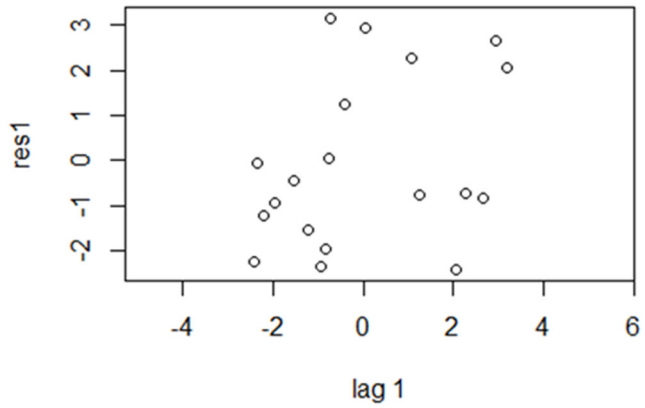




First assumption Homogeneity
The residuals (one stack for each drug) show similarly vertical dispersion around zero.

Second assumption:
Are the errors normally distributed ?
The residuals here are definitely not normal.

3b Evaluate the probability model.



Third assumption. Independent errors?

We do not know the order in which observations were made, and so have little basis for checking this assumption. If we assume the observations were made in the order presented in the publication, we can check the assumption by arranging the residuals in that order and then plotting residuals against their adjacent value. To do this we plot the lagged residuals against the residuals.

There are no upward or downward trends, so we judge the residuals to be independent of each other at lag 1, in the order presented in the publication.

Fourth assumption.

Do the residuals sum to zero? This assumption is usually met because packages use least squares or maximum likelihood to produce unbiased estimate that sum to zero.

Conclusion:

Residuals are homogeneous and independent, but deviate substantially from normal. The data used by Gossett (1908) to introduce the *t*-test did not meet the normality assumption for this test.

Because the residuals depart noticeably from normal we will compare Gossett's results to the results of a randomization test.

4. Partition df and SS according to model.

df_{total} = n-1 df_{Drug} = 2 categories – 1 $d_{fresidual} = df_{total} - df_{Drug}$

SS_{total} = var(T) · df_{total} = 4.072 · 19 = 77.37

$SS_{residual} = \sum\left(T - \overline{T_{DrugA}}\right)^2 + \sum\left(T - \overline{T_{DrugB}}\right)^2$

SSDrug = SS_{total} - $SS_{residual}$ = 12.48 This is the improvement in fit.

GLM:	$T - \beta_o$	=	β_{Drug} Drug	+	ε
Source:	Total	=	Drug		residual
df	20 – 1	=	1	+	18
SS	77.37	=	12.48	+	64.89

4. How good is the evidence?

Calculate likelihood ratio for the overall (omnibus) model.

$LR = (1\text{-}R^2)^{(-n/2)}$ R^2 = 12.48/77.37 $1\text{-}R^2$ = 64.89/77.37

$LR = (64.89/77.37)^{-20/2} = 5.8$

$LR < 20$ There is insufficient evidence of a difference between the two means.

Note that the measure of evidence depends on the normal error assumption, which was not well supported by graphic evaluation.

4. Decide on mode of inference.

The drug is being considered for use with people. So Type I error (risk of false positive) is relevant to reporting the experimental results. We use a frequentist approach to make a decision relevant to Type I error.

5. State sample and population for frequentist inference.

The population in this case: an infinitely large number of repeats of the same experiment with the same drug. The term β_{Drug} in the model is the long run difference between the two means. We could estimate this parameter for the population by running the experiment repeatedly, then taking the average value of the differences between the two groups. The sample is considered representative of the population of repeats conducted in the same way.

6. For frequentist inference, state H_A / H_o pair, test statistic, its distribution, and tolerance of Type I error.

There is one term in the model. Is this term significant ? (not due to chance).

The research hypothesis is that the drugs differ in effect.	H_A: $\mu_A \neq \mu_B$
The null hypothesis is the drugs do not differ in effect.	H_o: $\mu_A = \mu_B$

Here is an equivalent formulation.

The research hypothesis is that the drugs differ in effect.	H_A: $\beta_{Drug} \neq 0$
The null hypothesis is the drugs do not differ in effect.	H_o: $\beta_{Drug} = 0$

The symbol β_{Drug} has a single value, the difference between the two means.
The hypotheses listed above are equivalent to the following pair of hypotheses

If the means differ, then var($T_A - T_B$) >0	H_A: var($T_A - T_B$) >0
If the means are the same, then var($T_A - T_B$) =0	H_o: var($T_A - T_B$) =0

State test statistic	F ratio or t- statistic. Note: $t^2 = F$.
Distribution of test statistic	F distribution
Tolerance for Type I error	$\alpha = 5\%$

The tolerance for Type I error is called α, which is conventionally set at 5%. This is a compromise between Type I and II error. One can set this at 1% or at 10%, depending on whether one is worried about Type I or II error. Setting α at a low value increases Type II error, the chance of rejecting a true effect.
If many tests are to be made, it is advisable to set tolerance for Type I error at α/n where n = number of tests. This is called a Bonferroni criterion. It takes into account the fact that the Type I error for multiple tests is not the same as for a single test. With α = 5% and 20 tests, you expect one "significant" result even when there is no real effect.

7. ANOVA - Move df and SS to ANOVA table.

Source	df	SS	MS	F	---->	p
Drug	1	12.48				
Res	18	64.89				
Total	19	77.37				

MS stands for the mean squared deviation.
MS = SS / df

$MS_{model} = MS_{Drug} = 12.482$

$MS_{res} = MS_{error} = 3.605$

MS_{tot} = Var(response) = Var(*T*)

MS_{tot} does not appear in the MS column, because $MS_{model} + MS_{res} \neq MS_{tot}$

```
Calculations move from left to right,
  MS from SS and df in ANOVA table
  F from MS
  p from F distribution
```

7. ANOVA - Compute test statistic $F = (MS_{Drug})/(MS_{res}) = (SS_{Drug})/(SS_{total} - SS_{Drug})$

Source	df	SS	MS	F	---->	p
Drug	1	12.48	12.482	3.46		
Res	18	64.89	3.605			
Total	19	77.37				

The *F*-ratio can be thought of as the signal to noise ratio.
How strong is the signal, relative to the noise (error) ?

F is the ratio of the explained variance (due to the entire model,
or due to a factor in the model) to the unexplained variance.

E.g., $F = MS_{model} / MS_{res}$

This can be calculated by hand, if necessary, using MS or SS from computer package.
Forming the correct F-ratio can require considerable skill and experience,
especially with complex experimental designs.
Computer packages sometimes produce incorrect F-ratios. It is a good idea to check with a statistician, if in doubt.

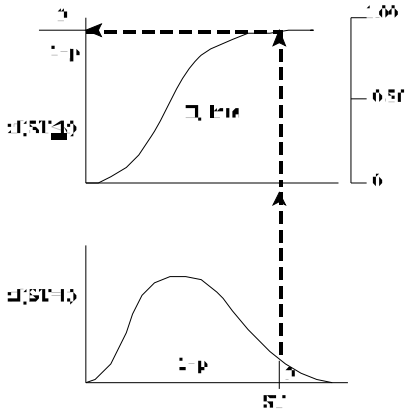
The completed table represents a sequence of computations from left to right. It results in an *F*-ratio, which will be small if the explained variance is small, large if the MS_{Drug} is large

7. ANOVA Table source, df, SS, MS, F-ratio, and p-value.

Here are the ANOVA calculations from a spreadsheet.

Time	Drug	Fits	Residuals
0.7	0	0.75	-0.05
-1.6	0	0.75	-2.35
-0.2	0	0.75	-0.95
-1.2	0	0.75	-1.95
-0.1	0	0.75	-0.85
3.4	0	0.75	2.65
3.7	0	0.75	2.95
0.8	0	0.75	0.05
0	0	0.75	-0.75
2	0	0.75	1.25
1.9	1	2.33	-0.43
0.8	1	2.33	-1.53
1.1	1	2.33	-1.23
0.1	1	2.33	-2.23
-0.1	1	2.33	-2.43
4.4	1	2.33	2.07
5.5	1	2.33	3.17
1.6	1	2.33	-0.73
4.6	1	2.33	2.27
3.4	1	2.33	1.07
4.072		0.657	3.415 variance
77.368	=	12.482	64.886 SS= 19*variance
19.000		1.000	18.000 df
		12.482	3.605 MS

$SS_{tot} = Var(T)* df_{tot}$
$SS_{fits} = Var(fits) * df_{tot}$
$SS_{res} = Var(res) * df_{tot}$
The same computations can be carried out in any package



7. Calculate p-value for terms in the model.

```
Redraw picture.
Show arrow from F in table to F in graph
```

```
MTB > cdf 3.46;
SUBC> f 1 18.
 0.921   3.46
```

$p = 1 - 0.921 = 0.079$

Source	df	SS	MS	F ---->	p
Drug	1	12.48	12.482	3.46	0.079
Res	18	77.37	3.06		
Total	19	64.89			

8. Recompute p-value if necessary.

When assumptions not met, recompute if: n small (Yes, n = 20)
p near α (Yes, p =0.079)

Because p-value is near α. the decision might change if the p-value recomputed by randomization.

Colquhoun (1971) carried out a randomization test, using
12000 of the 184,756 possible permutations of the data into 2 groups.
The p-value was p = 0.0813 (976/12000)
This is close to p-value from the t-distribution, and it leaves the decision unchanged.
The p-value changed by a factor of 0.0813 / 0.079 = 1.03 (hardly at all)
Ehe substantial violation of normal error assumption had little effect on the estimate of Type I error (the p-value) in this case. This is because the distribution was symmetrical around the mean, despite the deviation in shape.

Note: It is not feasible to construct frequency distribution from all permutations. Instead, we sample from the list of all permutations by sampling at random from the data, compute the F-ratio repeatedly, and construct the frequency distribution of the F-ratio when the null hypothesis is true.

Computer packages produce ANOVA tables with F-ratios and p-values. However, it is important to learn how one quantity is computed from another in this table, in order to understand the table. It is also important to write the model out, before executing the analysis. Writing the model, and the list of explanatory variables, then calculating the degrees of freedom, is useful in making sure the computer executed the analysis you had in mind, rather than something else.

9. Declare and report decision about model terms (compare *p* to α).

I.e. Compare the observed statistic to population of such statistics.

$0.0813 = p > \alpha = 0.05$ So we cannot reject H_o: $Var(\beta_{Drug}) = 0$

Equivalently, we cannot reject H_o: $\beta_{Drug} = 0$

Report decision and conclusion:

Decision: We cannot reject the null.

Conclusion: There is no statistically significant difference in extra time slept.

$F_{1,18} = 3.46$ $p = 0.081$ (randomized)

When we cannot reject the null hypothesis, we then consider Type II error, that of a null hypothesis that is not true.

We ask: What difference could have been detected, given the variance and the sample size? To answer this, we take the observed difference between the means ($\Delta T = 2.33 – 0.75 = 1.58$ hours), then increase this difference until the p-value becomes significant. We start with guess: we increase difference by adding 0.5 hours to each value in the group with the larger mean. This increases the mean to 2.83 hours. It increases the difference to $\Delta T = 2.08$ hours. Then we run the GLM routine to obtain the p-value.

9. Report decision about model terms (continued)

We run the GLM routine repeatedly until we find the difference that results in p = 0.05.

$\Delta T = 2.33 - 0.75 = 1.58$	F = 3.46	p = 0.079	p-value from F-distribution
$\Delta T = 2.83 - 0.75 = 2.08$	F = 6.00	p = 0.025	too high, try 0.2 increase
$\Delta T = 2.53 - 0.75 = 1.78$	F = 4.39	p = 0.05	

The minimum detectable difference was 1.78 hours, which is higher than the observed difference by 1.78/1.58 = 1.127. With this sampling effort and variance we could have detected a difference of 1.78 hours. A true difference of 1.77 hours of extra sleep would go undetected by this experiment. A better experiment is needed, one that has a chance of detecting a smaller difference. One way to improve the experiment is to increase the number of trials, which will reduce the error variance.

If we are planning another experiment it is informative to compute the sample size needed to detect a difference, given the variance and contrast between means. To do this we increase sample size until the F-ratio becomes significant at 5%. Because p is already close to α we start with a small increase of 10, from $n = 20$ to $n = 30$.

Try a slightly smaller increase, of 8 (4 per group), from $n = 20$ to $n = 28$

Assuming the same variance and same difference in means, a sample size of 15 per group ($n = 30$) was needed to detect the observed difference. This is a feasible increase. These calculations are readily done in a spreadsheet that recalculates from a change in the total df.

Source	df	SS	MS	F ---->	p
Drug	1	12.48	12.482	4.19	0.051
Res	26	77.37	2.976		
Total	27	64.89			

10. Report and interpret parameters of biological interest.

The estimated effect size was substantial: 2.33 – 0.75 hours = 1.58 hours.
However, the research (alternative) model was only 5.8 time more likely than the null.
There was insufficient evidence for the research model: $LR = 5.8 < 20$
The certainty was low: $p = 0.08 > 0.05$.
To illustrate the t-test, Gossett used an example where the magnitude of the effect size, by itself, would lead to a conclusion of a substantial difference between the two drugs. The effect size was large but the variability did not allow us to reject the null hypothesis at a conventional 5% Type I error rate with an unpaired design. With this sampling effort and variability, we could have detected a difference of 1.78 hours in time of sleep, which is only 13% higher-- 1.78/1.58 = 1.13. An effect smaller than 1.78 hours would go undetected with this sampling effort and variability.
If individual scores on the two drugs are correlated, then use a paired design. This will reduce the residual MS, despite the reduction in degrees of freedom from 18 to 8.

GLM Paired t-test

Does running the previous analysis as a paired t-test (Ch10.1) result in a better analysis? If the hours of extra sleep, relative to control (no drug) are correlated across individuals, then the paired t-test is potentially more statistically powerful. That is, it has greater power to detect a difference.

When we run a correlation of the hours of extra sleep for the two drugs, relative to control, the correlation is strong. The explained variance is substantial, greater than 50%.

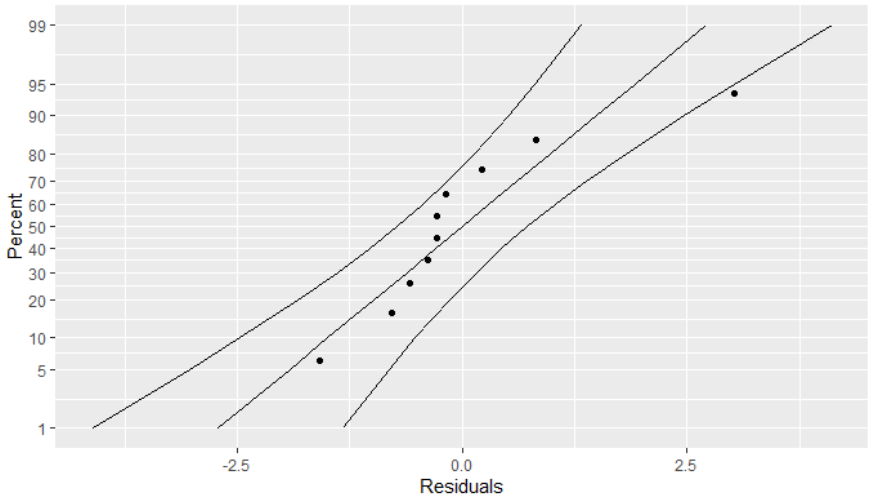
Correlation	0.795
R^2	0.632
1-R^2	0.368
n	10
LR	149

There is good evidence (LR = 150) for correlated responses to the two drugs.

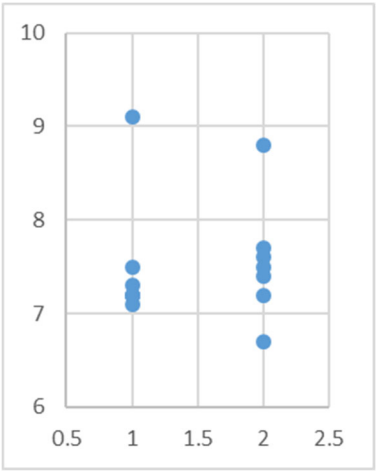
```
0.7         1.9
-1.6        0.8
-0.2        1.1
-1.2        0.1
-0.1        -0.1
 3.4        4.4
 3.7        5.5
 0.8        1.6
 0.0        4.6
 2.0        3.4

DrugA     DrugB
Cushny.dat
```

Using Ch10.1 as a template, run an analysis of the data for Drug A vs Drug B as a paired comparison.

Here is the Normal error probability plot for the 10 residuals from a paired comparison.



Example. Daphnia ages



Data from Box 9.5 p 220 Sokal and Rohlf 1995

Does time to maturity differ in two genetic crosses in the water flea *Daphnia* ?

1. Construct model

Verbal model: age depends on strain.

Graphical model.

Draw means at 7.56 (Strain 1) and at 7.51 (strain 2)

Response variable

A = age (in days) at beginning of reproduction in *Daphnia longispina* in two genetic crosses I and II (ratio type of scale)

Explanatory variable.

St = I or II (nominal scale)

n = 14 observations, 7 in each of groups I and II

Formal Model $A = \beta_o + \beta_{St} \cdot St + \varepsilon$

β_o is the intercept (mean of reference group, Strain I).

β_{St} is the contrast (dfference) between the two groups

$\beta_o + \beta_{St} \cdot St$ = mean of second group Strain II

2. Execute analysis. Place data in model format:

Column with response variable, Age A.

Column with explanatory variable, Strain = I or II

Code model statement in statistical package according to the GLM

$A = \beta_o + \beta_{St} \cdot St + \varepsilon$

```
MTB> ANOVA 'A' = 'Strain'
MTB> GLM 'A' = 'Strain';
SUBC> fits c3;
SUBC> res c4.
MTB > plot c4*c3
```

Code for R

```
> Daphniamodel <- lm(A ~ St, data=Daphnia)
> plot(Daphniamodel)
```

parameters reported by GLM routine

$\hat{\beta}_0 =$ 7.5571 = Strain 1 mean, the intercept

$\hat{\beta}_{St} = 0.0428$ = Contrast, the difference between the two means.

3a. Evaluate structural model

No slopes (straight lines) used, so no need to check for bowls/arches.

3b. Evaluate error model.

Homogeneous?

The two stacks of residuals in this plots are of similar spread, so we conclude the residuals are homogeneous.

```
MTB > plot 'res' 'fits'

          -
     0.80+                                                          *
          -
res       -  *
          -  3                                                      *
          -  *                                                      *
     0.00+  *                                                      2
          -                                                        *
          -
          -
          -
    -0.80+
          -
          -
          -                                                        *
          -
    -1.60+  *
            --------+---------+---------+---------+---------+--------fits
               7.5200    7.5280    7.5360    7.5440    7.5520
```

Normal?

No, residuals area skewed toward negative residuals by outliers.

```
MTB > hist 'res'

Histogram of res   N = 14

Midpoint   Count
    -1.6       1  *
    -1.2       1  *
    -0.8       0
    -0.4       0
     0.0       5  *****
     0.4       6  ******
     0.8       1  *
```

Independent? Plot residuals at lag 1.

No evident up or down trend.

Residuals judged to be independent.

```
MTB > plot 'res' 'lagres'
          -
     0.80+                                    *
          -
 res      -                                      *
          -     *                              *             *
          -                                 *       *
     0.00+             *                    *         *
          -                                      *
          -
          -
          -
    -0.80+
          -
          -
          -                                    *
          -
    -1.60+                                              *
          ------+---------+---------+---------+---------+---------+lagres
            -1.50     -1.00     -0.50      0.00      0.50      1.00
```

3b. Evaluate probability model

Conclusion. Residuals are homogeneous and independent, but deviate from normal .

4, Partition df and variance according to model.

```
ANOVA table headings on chalk board, upper right.
GLM just to the left.
Headings under model, then move to ANOVA table
```

Calculate $df_{total} = n - 1 = 14-1 = 13$ Partition df according to model

Calculate SS_{tot} from Var(A), the variance of response variable.

$SS_{tot} = \text{Var}(A) \cdot df_{tot} = 0.42247 \cdot 13 = 5.49214$

Use statistical package to partition SS_{tot} according to model

GLM:	$A - \beta_o$	=	$\beta_{St} \cdot St$	+	ε
Source:	Total	=	Strain +	residual	
df	13	=	12	+	1
SS	5.492	=	0.00641	+	5.4857

4. How good is the evidence for a difference?

Calculate the likelihood ratio for model.

$LR = (5.48571/5.492)^{-14/2} = 1.008$

$LR < 10$ There is no evidence of a difference.

5. Decide on mode of inference. Is hypothesis testing appropriate?

In the absence of any evidence for a difference, evidentialist inference is appropriate. However, the likelihood ratio was calculated assuming a normal error, which was not warranted by examination of the histogram of the residuals. We could recompute the likelihood using randomization in place of assuming a normal error (Owen text).

10. Report and interpret parameters of biological interest.

$\overline{A}_I = 7.5571$ days stdev = 1.319 days $n = 7$

$\overline{A}_{II} = 7.5143$ days stdev = 1.976 days $n = 7$

The two means differ by only 6 parts in a 1000 (7.5571 – 7.5143)/7.5357 = 0.006

The parameter of biological interest is the average time to maturity, regardless of strain, which is $\overline{A} = 7.5357$days stdev = 1.997 days $n = 14$

There is no evidence of any difference in time to maturity between the two strains.

Was the lack of evidence due to poor execution? To address this we look at the minimum difference that could have been detected, given the variance and sample size.

To do this we keep increasing the difference between two groups until the difference reaches a threshold, such as a Type I error of 5%. In practice we add an offset to one group, compute the t-statistic and p-value, increase the offset, compute the t-statistic and p-value again, and continue until the p-value falls below the significance level (5%).

10. Report and interpret parameters (continued)

This computation can be done in a spreadsheet.

Age	Strain	Fits	Residuals	Strain 0	Offset
8.002	0	8.3163	-0.31429	7.2	0.802
7.902	0	8.3163	-0.41429	7.1	0.802
9.902	0	8.3163	1.585714	9.1	0.802
8.002	0	8.3163	-0.31429	7.2	0.802
8.102	0	8.3163	-0.21429	7.3	0.802
8.002	0	8.3163	-0.31429	7.2	0.802
8.302	0	8.3163	-0.01429	7.5	0.802
8.8	1	7.5571	1.242857		
7.5	1	7.5571	-0.05714		
7.7	1	7.5571	0.142857		
7.6	1	7.5571	0.042857		
7.4	1	7.5571	-0.15714	Strain 0	8.3163
6.7	1	7.5571	-0.85714	Strain 1	7.5571
7.2	1	7.5571	-0.35714	Diff	0.7591

0.577		0.155	0.422	variance	
7.503	=	2.017	5.486	SS=	13*variance
13.000		1.000	12.000	df	
		2.017	0.457	MS	
			4.412281	F	
			0.050037	p	

The two strains would have to differ by 0.76 days to reach the 5% Type I error threshold.
I.e. the strains would have to differ by (0.7591/7.5571) = 10% to be significant.
The analysis was capable of detecting a 10% difference in age.
The absence of evidence for a difference cannot be attributed to a poorly executed study.

Extra

The t-test is a special case of a one-way ANOVA. The *F*-ratio, by definition, is t^2

Here is the ANOVA table calculation of the *F*-ratio

Source	df	SS	MS	F
strain	1	0.00641	0.00641	0.014
error	12	5.48571	0.45714	

Here is the calculation of the F- ratio from the formula for the *t*-statistic.

$$t = \frac{\left(\bar{X}_I - \bar{X}_{II}\right) - \left(\mu_I - \mu_{II}\right)}{\sqrt{\frac{1}{n}\left(s_I^2 + s_{II}^2\right)}} \qquad t = \frac{\left(7.5571 - 7.5143\right) - \left(0 - 0\right)}{\sqrt{\frac{1}{7}(0.50476 + 0.40952)}} = \frac{0.4286}{\sqrt{\frac{0.9143}{7}}} = \frac{0.4286}{0.3614} = 0.1186$$

$$t^2 = 0.1186^2 = 0.014$$