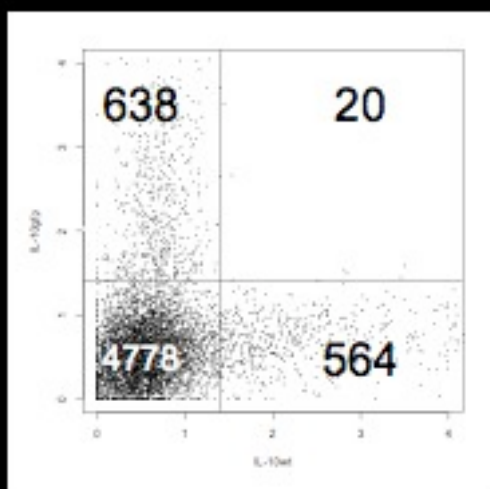
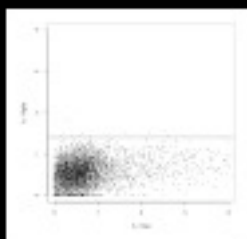
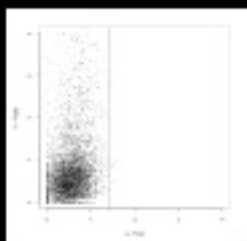


Monday, November 24, 2014

## Dependent and independent variables, Regression and ANOVA

Recapitulating



```
matrix(c(tl, bl, tr, br), 2, 2)
      [,1] [,2]
[1,] 1312 1989
[2,] 1514 1265

fisher.test(matrix(c(tl,bl,tr,br),2,2))
Fisher's Exact Test for Count Data

data: matrix(c(tl, bl, tr, br), 2, 2)
p-value = 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.517577 0.637989
sample estimates:
odds ratio
 0.5742834
```

## ANOVA

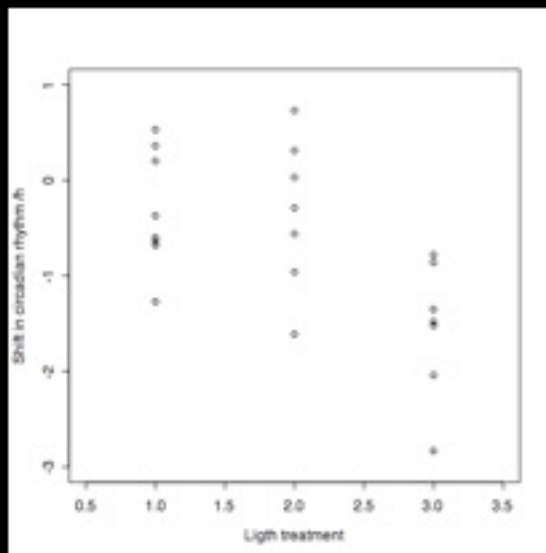
Traveling to a different time zone causes jet lag. Adjustment to the schedule of light resets the internal circadian clock. This change in the internal clock is called "phase shift".

Campbell and Murphy (1998) reported that human circadian rhythm could be reset by exposition of the back of the knee to light. The surprising result was regarded with skepticism by the community.

The following data from Wright and Czeisler (2002) reexamined this question.

Raw data and descriptive statistics of phase shift, in hours, for the circadian rhythm experiment

Treatment	Date /h	mean	s	n
Control	0.53, 0.36, 0.20, -0.37, -0.60, -0.64, -0.68, -12.7	-0.3688	0.6176	8
Knees	0.73, 0.31, 0.03, -0.29, -0.58, -0.96, -1.01	-0.3357	0.7908	7
Eyes	-0.76, -0.86, -1.35, -1.48, -1.52, -2.04, -2.83	-1.5514	0.76063	7



### Statistical test

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

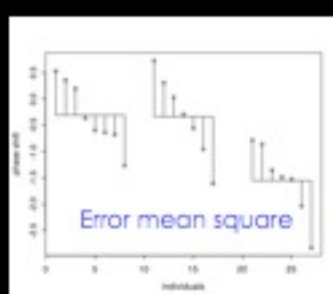
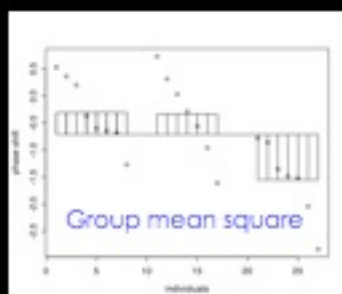
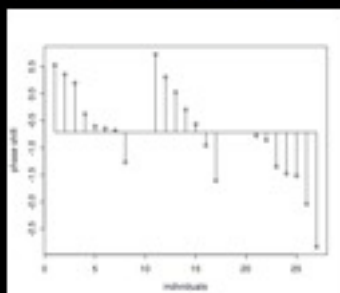
$H_1$ : at least one  $\mu_i$  is different from the others

What is the best statistics model to deal with this?

Multiple pairwise comparisons with t-test?

Bonferroni correction of p-value ?

### The principles of ANOVA



## ANOVA

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$H_1$ : at least one  $\mu_i$  is different from the others

Error mean square of ANOVA is the pooled sample variance, a measure of the variability within the groups

Group mean square of ANOVA represents the variation among individuals belonging to different groups. It should be similar to the error mean square if population means are equal

Error mean square ( $MS_{error}$ ) of ANOVA is the pooled sample variance, a measure of the variability within the groups

$$MS_{error} = \frac{\sum_{i=1}^k s_i^2 (n_i - 1)}{N - k}$$

$N$  = total individuals  
 $n_i$  = number of individuals in group  $i$   
 $s_i$  = standard deviation in each group  
 $k$  = number of groups

Group mean square of ANOVA represents the variation among individuals belong to different groups. It should be similar to the error mean square if population means are equal

$$MS_{groups} = \frac{\sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y})^2}{k - 1}$$

$$\bar{Y} = \frac{\sum_{i=1}^k n_i \bar{Y}_i}{N}$$

$N$  = total individuals  
 $n_i$  = number of individuals in group  $i$   
 $s_i$  = standard deviation in each group  
 $k$  = number of groups  
 $\bar{Y}_i$  = within group mean  
 $\bar{Y}$  = total mean

## ANOVA table

Source of variation	Sum of squares	df	Mean Squares	F-ratio
Groups	$SS_{\text{groups}} = \sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y})^2$	$k - 1$	$\frac{SS_{\text{groups}}}{df_{\text{groups}}}$	$F = \frac{MS_{\text{groups}}}{MS_{\text{error}}}$
Error	$SS_{\text{error}} = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$	$N - k$	$\frac{SS_{\text{error}}}{df_{\text{error}}}$	
Total	$SS_{\text{total}} = SS_{\text{groups}} + SS_{\text{error}}$	$N - 1$		

The ANOVA **statistics** is the variance ratio, **F**, which under the null hypothesis should lie close to 1

$$F = \frac{MS_{\text{groups}}}{MS_{\text{error}}}$$

What is the sampling distribution of **F** ?

**F** is distributed according to the so called **F-distribution** with **k-1** degrees of freedom for the **numerator** and **N-k** degrees of freedom for the **denominator**

As usual, using the appropriate quantiles of the **F-distribution** we can bound **F** according to the null hypothesis

## Variance Explained

**R<sup>2</sup>** is used in ANOVA to summarize the contribution of group differences to the variability in the data

$$SS_{\text{total}} = SS_{\text{groups}} + SS_{\text{error}}$$

$$R^2 = \frac{SS_{\text{groups}}}{SS_{\text{total}}}$$

The test will define whether we can trust the alternative hypothesis that at least one of the groups has a different mean

If you want to detect which group it is you can do:

Planned comparison

Unplanned comparison

Planned comparison

Difference between the means of the two planned groups

$$\bar{Y}_i - \bar{Y}_j$$

The standard error of the difference between the means

$$SE = \sqrt{MS_{error} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}$$

with N-k degrees of freedom

The difference between the means is t-distributed with N-k degrees of freedom

To get a grasp of the power of the planned comparison let's compare it's 95% confidence interval with that of the two sample t-test

Home Work: Read about Kruskal-Wallis test

1. Make pairwise comparisons with Wilcoxon-Mann-Whitney test
2. Make pairwise comparisons with t-test
3. Do ANOVA without using the built-in function of R
4. Do ANOVA with the built-in function of R

## Linear Regression

Regression is a method that predicts the value of a numerical from the value of another

Regression and correlation both describe features of a scatter plot, and measure the relationship between two numerical variables.

Correlation treats both variables equally, whereas regression predicts the value of one variable based on the other variable.

Correlation measures the strength of association between the two variables, whereas regression measure how steeply the response variable changes, on average, with the change in the explanatory variable.

## Linear regression

The equation of the regression line:

$$Y = a + bX$$

The formula has two coefficients:

The intercept  $a$  is the value of  $Y$  when  $X$  is zero

The slope  $b$  is the rate of change in  $Y$  per unit of  $X$

## Linear regression

Slope

$$b = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

(Compare with expression for correlation)

Intercept

$$a = \bar{Y} - b\bar{X}$$

## Linear regression

Assumptions

At each value of  $X$  there is a population of possible  $Y$ -values whose mean lies on the true regression line

At each value of  $X$ , the distribution of possible  $Y$ -value is normal

The variance of  $Y$ -values is the same at all values of  $X$

At each value of  $X$ , the  $Y$ -measurements represent a random sample from the possible  $Y$ -values



## Linear regression

Predicted value

$$\hat{Y} = a + bX$$

The predicted value of a regression estimates the mean value of Y for all individual measurements that have given X.

Residuals

$$Y_i - \hat{Y}_i$$

$$MS_{\text{residual}} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n - 2}$$

## Linear regression

In the sample we have

$$Y = a + bX$$

In the population we have the true values

$$y = \alpha + \beta x$$

How can we infer the true values of the parameters  $\alpha$  and  $\beta$  from the measured ones  $a$  and  $b$  ?

Can we predict the true value of y as a function of x ?

## Linear regression

Interval of confidence for  $\beta$

When the assumptions of the linear regression are met then the sampling distribution of b is a normal distribution with mean  $\beta$  and standard error estimated from the sample by:

$$SE_b = \sqrt{\frac{MS_{\text{residual}}}{\sum (X_i - \bar{X})^2}}$$

With this we can provide confidence intervals for the true value of  $\beta$

$$b - t_{\alpha(2)df} SE_b < \beta < b + t_{\alpha(2)df} SE_b$$

## Linear regression

Testing hypothesis for slope  $\beta$

$$t = \frac{b - b_0}{SE_b}$$

t-distributed with  $df=n-2$

Since we are comparing mean values we could also use ANOVA

## Linear regression

ANOVA method for testing zero slope

Source of variation	Sum of squares	df	Mean Squares	F-ratio
Regression	$SS_{\text{regression}} = \sum (\hat{Y}_i - \bar{Y})^2$	1	$\frac{SS_{\text{regression}}}{df_{\text{regression}}}$	$F = \frac{MS_{\text{regression}}}{MS_{\text{residual}}}$
Residual	$SS_{\text{residual}} = \sum (Y_i - \hat{Y}_i)^2$	$n - 2$	$\frac{SS_{\text{residual}}}{df_{\text{residual}}}$	
Total	$SS_{\text{total}} = \sum (Y_i - \bar{Y})^2$	$n - 1$		

## Linear regression

Variance Explained

$$R^2 = \frac{SS_{\text{regression}}}{SS_{\text{total}}}$$

A measure of the quality of the fitting

## Linear regression

Confidence bands for the predicted mean  
measure the precision of the mean Y value for each value of X

Prediction interval of any one measure of Y  
measure the precision of the estimate of any individual which as a  
value of X

The second is broader denoting the greater uncertainty

## Linear regression

Practical example 1:

"Is there a limit to human life expectancy? [\[zip\]](#)"

Practical example 2:

"Is a specific treatment of autoimmune diabetes effective? [\[zip\]](#)"

## Regression to the mean

Appears when two variables have a correlation less than 1

Individuals that are far from the mean for one of the measurements  
will, on average, lie closer to the mean for the other measurement

## Assumptions

At each value of  $X$  there is a population of possible  $Y$ -values whose mean lies on the true regression line

At each value of  $X$ , the distribution of possible  $Y$ -value is normal

The variance of  $Y$ -values is the same at all values of  $X$

At each value of  $X$ , the  $Y$ -measurements represent a random sample from the possible  $Y$ -values

## General Linear Models

A general linear model can have more than one explanatory variable

The explanatory variables can be also categorical

Typically with deal with them with ANOVA  
(remember feeding a linear model to anova in R)