Class 4 - Statistics part II

1.3. Comparing more than 2 populations

If sampled data follows a normal distribution, we can take and compare mean values of three (or more populations)... if not we may need to test the median

- **1.3.1 one-way ANOVAs (an ANOVA with only one independent variable)**

(tutorial link)

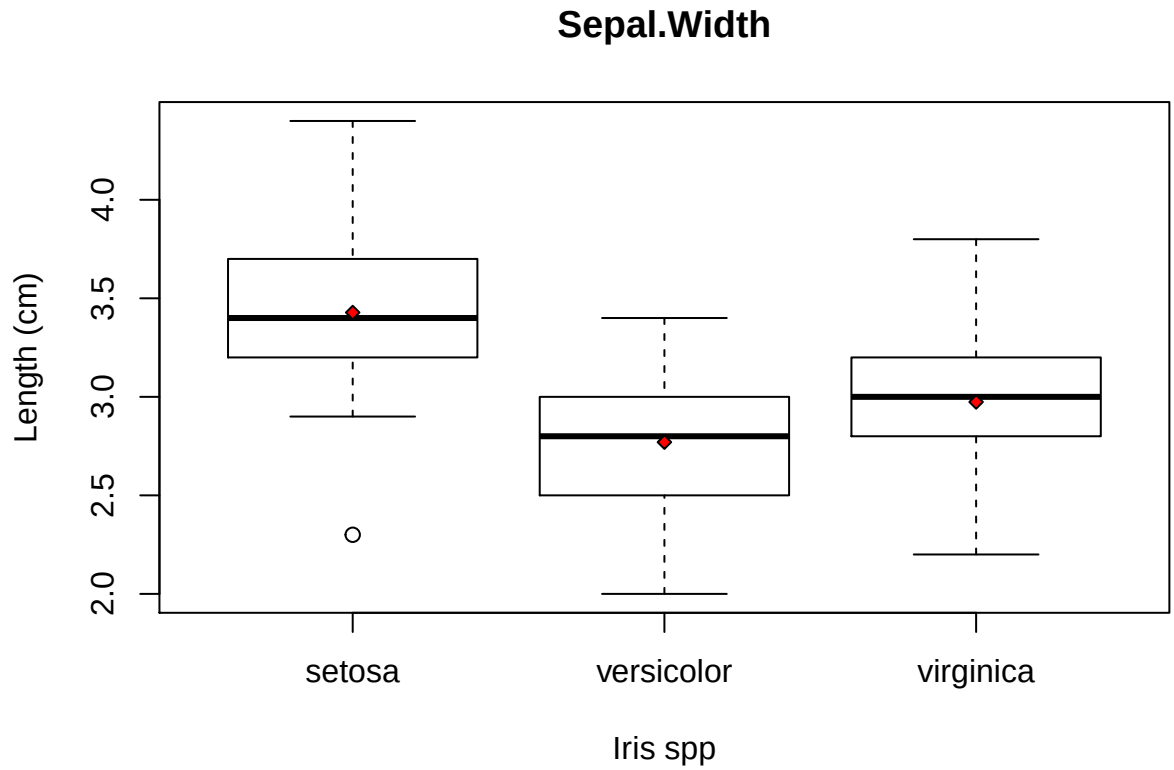
Assumptions: normality of data, samples similar in size (1.5) and variance homogeinety

H0 : the means of the different groups are the same H1: At least one sample mean is not equal to the others.

First let's get a look at the data

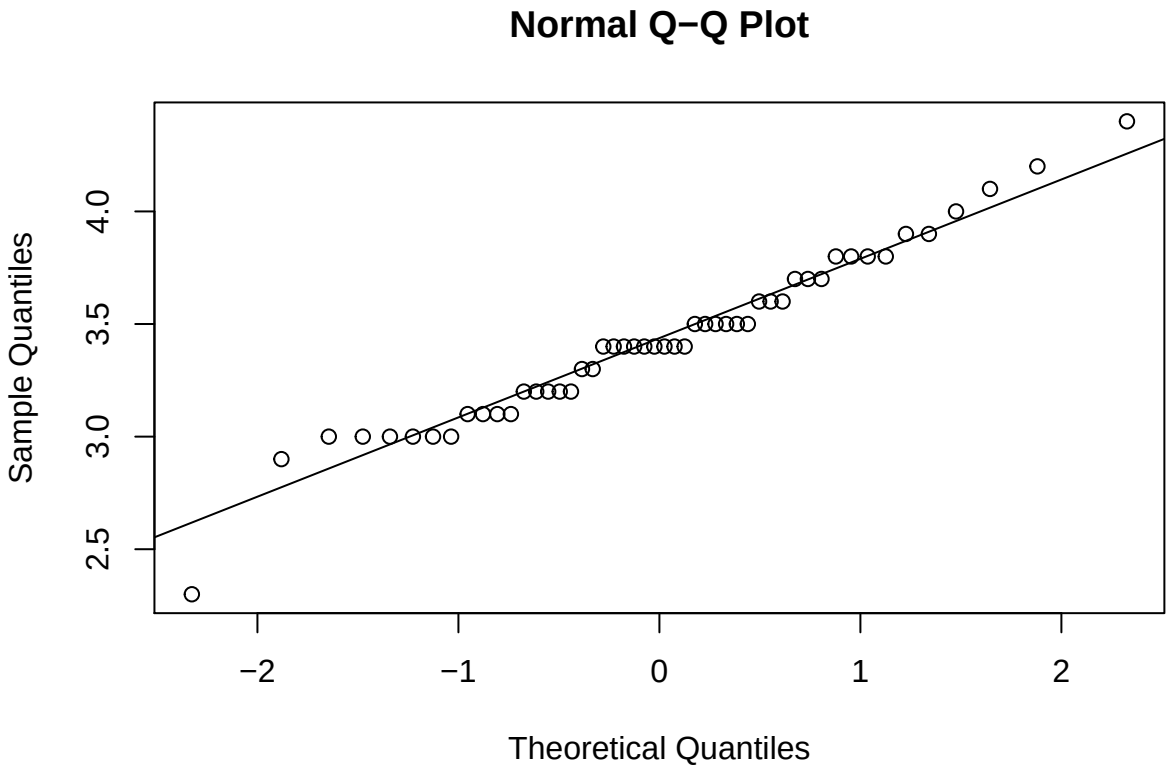
```
iris_df <- iris

mean_sepalw <- as.vector(by(iris_df$Sepal.Width, iris_df$Species, mean))
boxplot(formula = Sepal.Width ~ Species, data = iris_df,
        main = "Sepal.Width",
        xlab = "Iris spp",
        ylab = "Length (cm)")
points(1:3, mean_sepalw, pch = 23, cex = 0.75,
       bg = "red")
```




```
#We haven't tested normality for setosa spp.

qqnorm(iris_df[iris_df$Species == "setosa", "Sepal.Width"])
qqline(iris_df[iris_df$Species == "setosa", "Sepal.Width"])
```




```
# with by() we may apply shapiro.test per species
by(iris_df$Sepal.Width, iris_df$Species, shapiro.test)
```

```
## iris_df$Species: setosa
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.97172, p-value = 0.2715
##
## ------------------------------------------------------------
## iris_df$Species: versicolor
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.97413, p-value = 0.338
##
## ------------------------------------------------------------
## iris_df$Species: virginica
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.96739, p-value = 0.1809
```

- **Test homogeneity of variances**

```
bartlett.test(Sepal.Width~Species, data = iris_df)
```

```
##
```

```
##  Bartlett test of homogeneity of variances
##
## data:  Sepal.Width by Species
## Bartlett's K-squared = 2.0911, df = 2, p-value = 0.3515
```

. . . we acept H0, we can perform ANOVA under all assumptions.

NOTE: Some authors claim that this test has some weaknesses, particularly in cases where the normality of data is weakly supported. Levene's test is more widely accepted for these cases (function leveneTest() from package "car")

- **Compute one-way ANOVA test**

```
#option2 (same as option1, but using different functions)
aov(Sepal.Width ~ Species, data = iris_df)
```

```
## Call:
##    aov(formula = Sepal.Width ~ Species, data = iris_df)
##
## Terms:
##                   Species Residuals
## Sum of Squares  11.34493  16.96200
## Deg. of Freedom        2       147
##
## Residual standard error: 0.3396877
## Estimated effects may be unbalanced
```

```
# we can save the result in an object and print a summary
res.aov <- aov(Sepal.Width ~ Species, data = iris_df)
summary(res.aov)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  11.35   5.672   49.16 <2e-16 ***
## Residuals   147  16.96   0.115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In these results, the null hypothesis states that the means are equal. Because the p-value is 2.2e-16 , which is less than the significance level of 0.05, you can reject the null hypothesis and conclude that some of the means are different

---

Post-hoc test that may be performed for pairwise comparison of the populations tested in ANOVA include TukeyHSD and pairwise t-test. The main difference between both is that the first considers between and within group variance (determined by ANOVA), while pairwise t-test doesn't.

To control for this, pairwise t-test allows p-value adjustment that control false discovery rate.

- **Tukey multiple pair-wise comparisons**

H0: all means being compared are from the same population

```
# this test may be applied directly to the object resulting from ANOVA

TukeyHSD(res.aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
```

```
##
## Fit: aov(formula = Sepal.Width ~ Species, data = iris_df)
##
## $Species
##                        diff         lwr        upr     p adj
## versicolor-setosa    -0.658 -0.81885528 -0.4971447 0.0000000
## virginica-setosa     -0.454 -0.61485528 -0.2931447 0.0000000
## virginica-versicolor  0.204  0.04314472  0.3648553 0.0087802
```

The Confidence interval of the mean difference don't include the value 0 for any pair, therefore none of the means are equal Confidence intervals that do not contain zero indicate a mean difference that is statistically significant.

- **Pairwise t-tests**

This is an alternatice to TukeyHSD, which allows for p-value correction.

```
# Pairwise t-tests with no assumption of equal variances
pairwise.t.test(iris_df$Sepal.Width, iris_df$Species,
                p.adjust.method = "BH")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  iris_df$Sepal.Width and iris_df$Species
##
##            setosa  versicolor
## versicolor < 2e-16 -
## virginica  6.8e-10 0.0031
##
## P value adjustment method: BH
```

```
pairwise.t.test(iris_df$Sepal.Width, iris_df$Species,
                p.adjust.method = "BY")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  iris_df$Sepal.Width and iris_df$Species
##
##            setosa  versicolor
## versicolor < 2e-16 -
## virginica  1.2e-09 0.0058
##
## P value adjustment method: BY
```

This particular function allows the use of an adjusted p-value.

Problems in multiple comparisons, multiplicity or multiple testing may occur when one considers a set of statistical inferences simultaneously.

The more inferences are made at once in a test, the chances of erroneous inferences increases.

For example: in a RNA-seq study with 10000 genes, 10000 inferences, and with for a significance value of 0.05 (probability of type I error), we expect 500 inferences to significant by chance

"BH" (Benjamini & Hochberg, aka "fdr") and "BY" (Benjamini & Yekutieli) ajustment methods control the false discovery rate, the expected proportion of false discoveries amongst the rejected hypotheses. The false

discovery rate is a less stringent condition than then the ones used in other methods, so these methods are more powerful than the others.

- **Compute one-way ANOVA test relaxing the homogeneity of variance assumption**

```
# Relaxing the homogeneity of variance assumption to allow unequal variances

# ANOVA test with no assumption of equal variances
oneway.test(Sepal.Length ~ Species, data = iris_df)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  Sepal.Length and Species
## F = 138.91, num df = 2.000, denom df = 92.211, p-value < 2.2e-16
```

```
# Pairwise t-tests with no assumption of equal variances
pairwise.t.test(iris_df$Sepal.Width, iris_df$Species,
                p.adjust.method = "BH", pool.sd = FALSE)
```

```
##
##  Pairwise comparisons using t tests with non-pooled SD
##
## data:  iris_df$Sepal.Width and iris_df$Species
##
##            setosa  versicolor
## versicolor 7.5e-15 -
## virginica  6.9e-09 0.0018
##
## P value adjustment method: BH
```

- **Non-parametric multiple comparisons: Kruskal-Wallis test**

H0: is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second sample.

```
kruskal.test(Sepal.Width ~ Species, data = iris_df)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Sepal.Width by Species
## Kruskal-Wallis chi-squared = 63.571, df = 2, p-value = 1.569e-14
```

1.4. Chi-square tests for homogeneity and independence

- **Chi-square for independency**

Consider two categorical variables that can be atributed to one population.

H0: the two variable are independent H1: variables are dependent

e.g. Study of drosophila fruit flies to analyse if survival (dead/aline) to a specific treatment is independent of sex (male/female)

```
# create a matrix with count values

contTable_flies <- as.matrix(data.frame(alive = c(133, 260),
                                        dead = c(90,93), row.names = c("male", "female")))
```

```r
flies_chsq <- chisq.test(contTable_flies)

flies_chsq
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contTable_flies
## X-squared = 11.742, df = 1, p-value = 0.000611
```

```r
#if variables were independent, this would be the expected frequencies
flies_chsq$expected
```

```
##          alive      dead
## male   152.151  70.84896
## female 240.849 112.15104
```

```r
# checking the residuals (r) highlights the cells the contribute more to the dependency
# r = (observed - expected) / sqrt(expected)

# Cells with the highest absolute standardized residuals contribute the most
# to the total Chi-square score

flies_chsq$residuals
```

```
##            alive      dead
## male   -1.552583  2.275232
## female  1.234014 -1.808384
```

The p-value is lower than 0.05 so we can reject H0 and conclude that the survival rate and sex are dependent. The treatment affects flies survival in a sex-dependent manner (death and male can be associated)

- **Chi-square for Homogeneity**

Considering two or more samples obtained from 1 or more populations. Is the distribution of the population homogeneous among the samples?

H0: Population is homogeneous across samples H1: At least one sample has higher obervations

Example:

Iris species setosa, virginica and versicolor were collected from two fields with different pHs (in a pre-defined area). Is the distribution of the species homogeneous?

```r
contTable_spp <- as.matrix(data.frame(field1 = c(168,140, 160), field3 = c(134,156,140),
                                      row.names = c("setosa", "virginica", "versicolor")))

iris_chsq <- chisq.test(contTable_spp)

iris_chsq
```

```
##
##  Pearson's Chi-squared test
##
## data:  contTable_spp
## X-squared = 4.4259, df = 2, p-value = 0.1094
```

```r
iris_chsq$expected
```

```
##              field1   field3
```

```
## setosa     157.3898 144.6102
## virginica  154.2628 141.7372
## versicolor 156.3474 143.6526
```

```
iris_chsq$residuals
```

```
##                field1     field3
## setosa      0.8457406 -0.8823195
## virginica  -1.1483502  1.1980172
## versicolor  0.2921138 -0.3047480
```

We can conclude that distribution of the three species is homogeneous across the two fields.

interesting visualizations of chi-square test results in this (link)[http://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r]

Activity 3

a) Hypothesis tests for 3 or more populations:

Load feeders.RData that you saved in Class2/ folder and repeat the selection indicated in the following box

```
load("../Class2/feeders.RData")

feeders_select <- feeders[feeders$SITE %in% c("Glide","Rita") & feeders$DurGROWTH == "7", ]
```

- a.1) What is the size (n) of the pH and OD_avg samples for each treatment?
- a.2) Create a boxplot for pH and another for OD_avg in function of TREATMENT, and plot the corresponding mean values as blue dots
- a.3) Run the the Shapiro-wilk test and the Bartlett test for OD_avg for all the three TREATMENT variables.
- a.4) Run the adequate ANOVA test to investigate if there are any differences between TREATMENTS for OD_avg
- a.5) Run the following command. What does function table() do to the data provided ?
- a.6) Assume that the OBSERVER were suppose to be homogeneously distributed in their task to take the samples in all the three types of cages (TREATMENTS). Was this true? Which would be the expected frequencies for the most even distribution?