Experimental Design and Statistical Methods

Workshop

NON-PARAMETRIC TESTS

Jesús Piedrafita Arilla

jesus.piedrafita@uab.cat

Departament de Ciència Animal i dels Aliments

UAB

Items

- Non parametric comparisons:
  - Two independent samples: Wilcoxon Sum-Rank Test
  - Two paired samples: Wilcoxon Signed-Rank Test
  - Several independent samples: Kruskal-Wallis test
  - Association/independence in proportions: Chi-square Test
- Basic commands
  - wilcox.test
  - wilcox.test (paired)
  - kuskal.test
  - chisq.test
- Libraries
  - coin (wilcox-test)
  - vcd (assocstats)

Non-parametric procedures

In previous sessions we have analysed parametric tests for comparing means. This session is devoted to present methods to compare location parameters and also proportions through non parametric tests.

	Parametric	**Non parametric**
Two independent samples	t-test	Wilcoxon Sum-Rank test
Paired data	Paired t-test	Wilcoxon Signed-Rank test
Several independent samples	ANOVA	Kruskal-Wallis test
Proportions		χ^2

The theory and examples of this lesson are heavily based on the book from G.A. Walker, 1997.

Comparison of two independent samples. Non parametric Wilcoxon Rank Sum test (WRST) (1)

Analogous of the two sample t-test, **based on ranks of the data**, can be used to compare location parameters (mean or median) of continuous numeric data and ordered categorical data, when the **data are not normally distributed**.

Seroxatene is an anti-depressant that would alleviate back pain, measured in a scale from – 3 to 3.

------ Seroxotene Group -------				-------- Placebo Group ---------			
Pat. No.	Score	Pat. No.	Score	Pat. No.	Score	Pat. No.	Score
2	0	16	-1	1	3	15	0
3	2	17	2	4	-1	18	-1
5	3	20	-3	7	2	19	-3
6	3	21	3	9	3	23	-2
8	-2	22	3	11	-2	25	1
10	1	24	0	13	1	28	0
12	3	26	2				
14	3	27	-1				

Raw data

Response (y)	No. of ties (m)	Ranks	Average Rank	$c_k = m(m^2-1)$
-3	2	1,2	1.5	6
-2	3	3,4,5	4	24
-1	4	6,7,8,9	7.5	60
0	4	10,11,12,13	11.5	60
+1	3	14,15,16	15	24
+2	4	17,18,19,20	18.5	60
+3	8	21,22,23,24, 25,26,27,28	24.5	504
				C=738

Conversion to ranks

⇒information loss

m, number of tied values in group k

WRST (2)

***H_0* of equal means** supported by **similar average ranks** between the two groups, i.e., R_1/n_1 is close to R_2/n_2.

----- Seroxotene Group -----				-------- Placebo Group --------			
Pat. No.	Score Rank	Pat. No.	Score Rank	Pat. No.	Score Rank	Pat. No.	Score Rank
2	11.5	16	7.5	1	24.5	15	11.5
3	18.5	17	18.5	4	7.5	18	7.5
5	24.5	20	1.5	7	18.5	19	1.5
6	24.5	21	24.5	9	24.5	23	4
8	4	22	24.5	11	4	25	15
10	15	24	11.5	13	15	28	11.5
12	24.5	26	18.5				
14	24.5	27	7.5				

We compute:

$$R_1 = 11.5 + 18.5 + 24.5 + \ldots + 7.5 = 261,$$

$$R_2 = 24.5 + 7.5 + 18.5 + \ldots + 11.5 = 145.$$

R_1 is compared to a critical value obtained from a special set of **tables** based on Wilcoxon rank-sum exact probabilities to determine the appropriate rejection region.

The requirement of special tables can be circumvented by using a **normal approximation** for larger samples, in the practice for samples larger than 8.

See next slide

WRST (3)

For the normal approximation, the expected value of R_1 under H_0 is:

$$\mu_{R_1} = \left(\frac{n_1}{N}\right)\left(\frac{N(N+1)}{2}\right) = \frac{n_1(N+1)}{2} = \ldots = 232$$

$N = n_1 + n_2$

$N(N+1)/2 = R_1 + R_2$ = Sum of the ranks

n_1/N = Proportion of the above sum from group 1

C as before

And the variance of R_1

$$\sigma^2_{R_1} = \frac{n_1 n_2}{12}\left(N + 1 - \frac{C}{N(N-1)}\right) = \ldots = 448.38$$

Test statistic with a 0.5 continuity correction

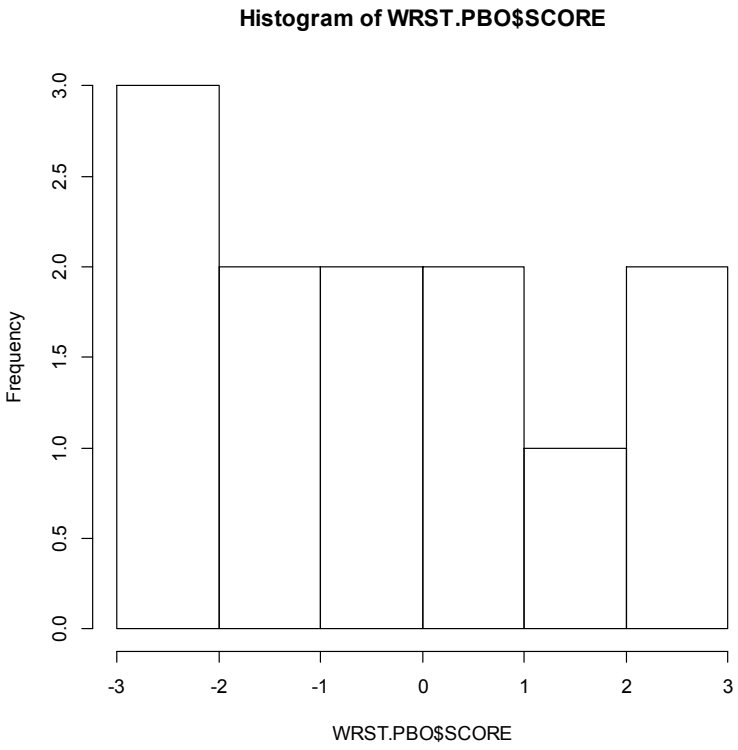
$$z = \frac{\left|R_1 - \mu_{R_1}\right|}{\sigma_{R_1}} = \frac{(261 - 232) - 0.5}{\sqrt{448.38}} = 1.346$$

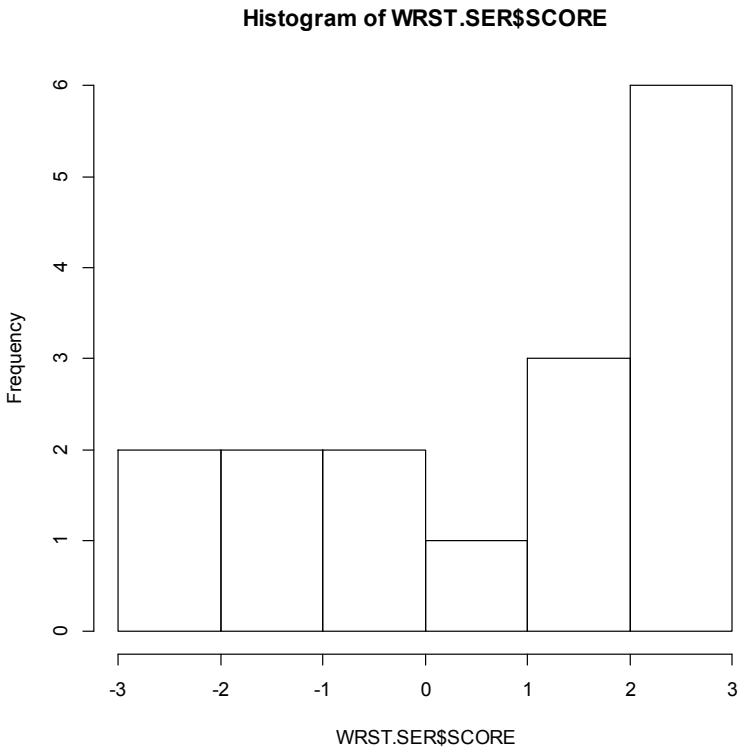
Decision rule: reject H_0 if $|z|>1.96$; so H_0 is not rejected in our example.

WRST - Histograms

```
> WRST.PBO <- WRST[TRT=="PBO",]
> hist(WRST.PBO$SCORE)
```

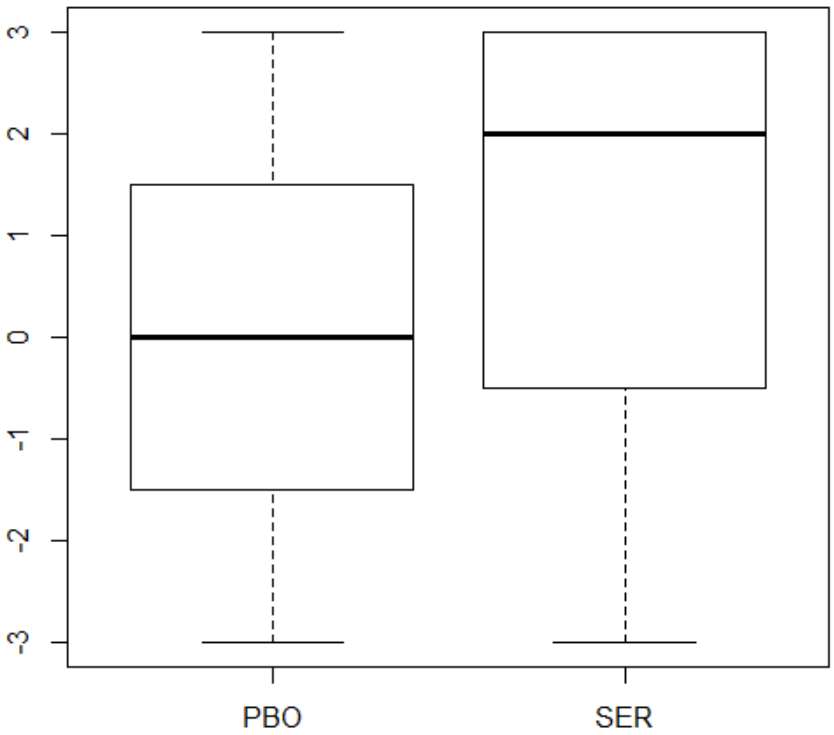
```
> WRST.SER <- WRST[TRT=="SER",]
> hist(WRST.SER$SCORE)
```





WRST - Boxplots and normality tests

```
> boxplot(SCORE~TRT)
```




```
> shapiro.test(WRST.PBO$SCORE)

Shapiro-Wilk normality test

data:  WRST.PBO$SCORE
W = 0.9509, p-value = 0.6501

> shapiro.test(WRST.SER$SCORE)

Shapiro-Wilk normality test

data:  WRST.SER$SCORE
W = 0.8531, p-value = 0.01514
```

Wilcoxon Rank-Sum test in R (1)

```
> wilcox.test(SCORE~TRT)

        Wilcoxon rank sum test with continuity correction

data:  SCORE by TRT
W = 67, p-value = 0.1783
alternative hypothesis: true location shift is not equal to 0

Mensajes de aviso perdidos
In wilcox.test.default(x = c(3L, -1L, 2L, 3L, -2L, 1L, 0L, -1L,  :
  cannot compute exact p-value with ties
```

H_0 is not rejected

The test statistic W is the sum of ranks in the first group minus its theoretical minimum (i.e., it is zero if all the smallest values fall in the first group).

Wilcoxon Rank-Sum test in R (2)

```
> library(coin)
> wilcox_test(SCORE~TRT)

  Asymptotic Wilcoxon Mann-Whitney Rank Sum Test

data:  SCORE by TRT (PBO, SER)
Z = -1.3695, p-value = 0.1708
alternative hypothesis: true mu is not equal to 0
```

H_0 is not rejected

```
> 2*pnorm(-1.3695)
[1] 0.170843
```

Comparison of paired data. Wilcoxon Signed-Rank test (WSRT) (1)

Used to compare responses between correlated and **paired data**, **without** requiring the assumption of **normality**.

Given a sample of n non-zero differences (zero's are ignored) we have to compute r_i, the rank of $|y_i|$ (lowest to highest).

$R^{(+)}$ and $R^{(-)}$ represent the sums of the ranks associated with positive and negative values of the y_i's

The test statistic is based on the smaller of $R^{(+)}$ and $R^{(-)}$.

$$S = (R^{(+)} - R^{(-)})/2$$

$$V = [n(n+1)(2n+1)]/24$$

Test statistic (approximated)

$$T = \frac{S\sqrt{n-1}}{\sqrt{nV - S^2}}$$

Decision rule

Reject H_0 if $|T| > t_{\frac{\alpha}{2},n-1}$

WSRT (2)

Let see an example on a new ocular wetting agent in patients of *keratitis sicca*.

Pat No.	Difference	Rank	Pat No.	Difference	Rank
1	7	14.5	13	-2	3
2	7	14.5	14	8	18.5
3	-1	1.5	15	7	14.5
4	-8	18.5	16	6	11
5	8	18.5	17	5	7.5
6	3	4.5	18	-5	7.5
7	-7	14.5	19	6	11
8	-3	4.5	20	0	--
9	-5	7.5	21	8	18.5
10	9	21	22	1	1.5
11	-5	7.5	23	6	11
12	10	22	24	0	--

differences of 0 are ignored

$$R^{(+)} = (14.5+14.5+18.5+4.5+21+22+18.5+14.5+11+7.5+11+18.5+1.5+11) = 188.5,$$

$$R^{(-)} = (1.5+18.5+14.5+4.5+7.5+7.5+3+7.5) = 64.5$$

$$S = (188.5 - 64.5)/2 = 62$$

After correcting for ties

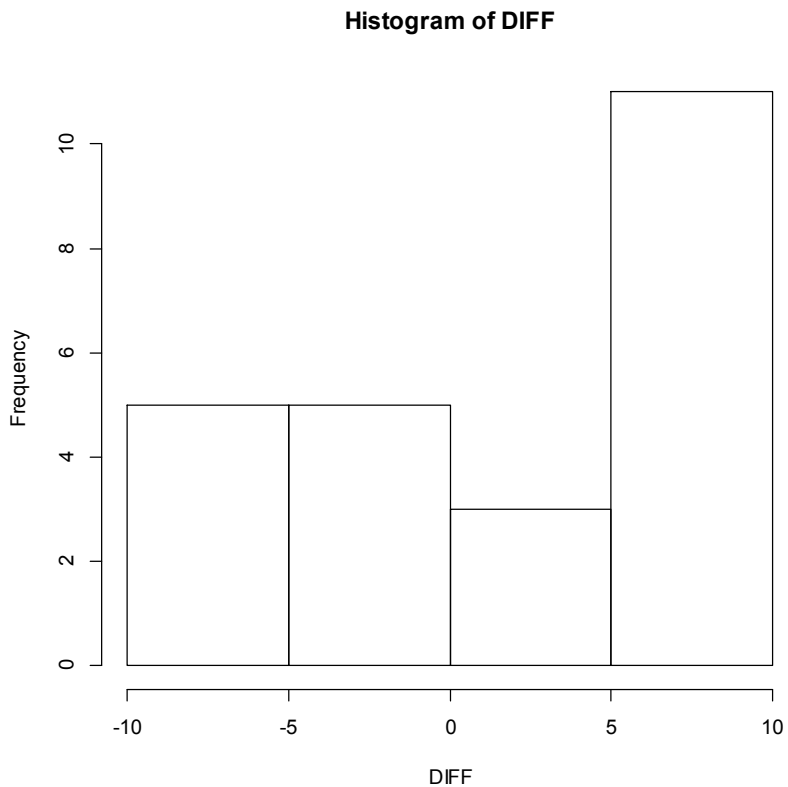
$$V = 944.25$$

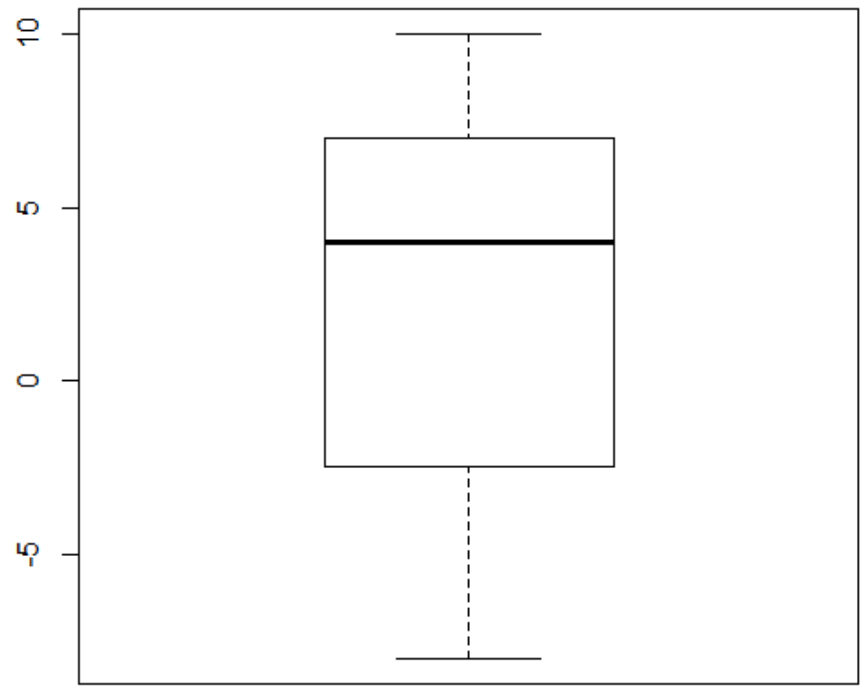
$$T = 2.184 > t_{0.025,21} = 2.080$$

H_0 (no difference) is rejected

WSRT – Histogram and boxplot

```
> DIFF<-(HYPO-OKER)
> hist(DIFF); boxplot(DIFF)
```






```
> shapiro.test(DIFF)
Shapiro-Wilk normality test
data: DIFF
W = 0.9082, p-value = 0.03223
```

Wilcoxon Signed-Rank test in R

```
> wilcox.test(HYPO, OKER, paired=T)

        Wilcoxon signed rank test with continuity correction

data:  HYPO and OKER
V = 188.5, p-value = 0.04535
alternative hypothesis: true location shift is not equal to 0

Mensajes de aviso perdidos
1: In wilcox.test.default(HYPO, OKER, paired = T) :
  cannot compute exact p-value with ties
2: In wilcox.test.default(HYPO, OKER, paired = T) :
  cannot compute exact p-value with zeroes
```

H_0 is rejected

Kruskal-Wallis test – intro -

Non-parametric test **analogue of One-Way ANOVA**. This is an extension of the Wilcoxon Rank-Sum Test, used to compare population location parameters (mean, median, etc.) among two or more groups including independent samples. It is based on the **ranks** of the data.

Unlike ANOVA, the assumption of normally distributed responses is not necessary

We will present an example a low dose (0.1%) compared to a high dose (0.2%) of a non-steroidal anti-psoriasis medication, using placebo as a control. The response was measured as degree of psoriatic lesion reduction, rated in an ordinal scale (see next slide).

Kruskal-Wallis test – data -

0.1% Solution		0.2% Solution		Placebo	
Pat No.	Category Code	Pat No.	Category Code	Pat No.	Category Code
1	5	3	5	2	5
6	4	5	8	4	3
9	1	7	2	8	7
12	7	10	8	11	1
15	4	14	7	13	2
19	3	18	4	16	4
20	6	22	5	17	2
23	7	26	4	21	1
27	8	28	6	24	4
32	7	31	4	25	5
				29	4
				30	5

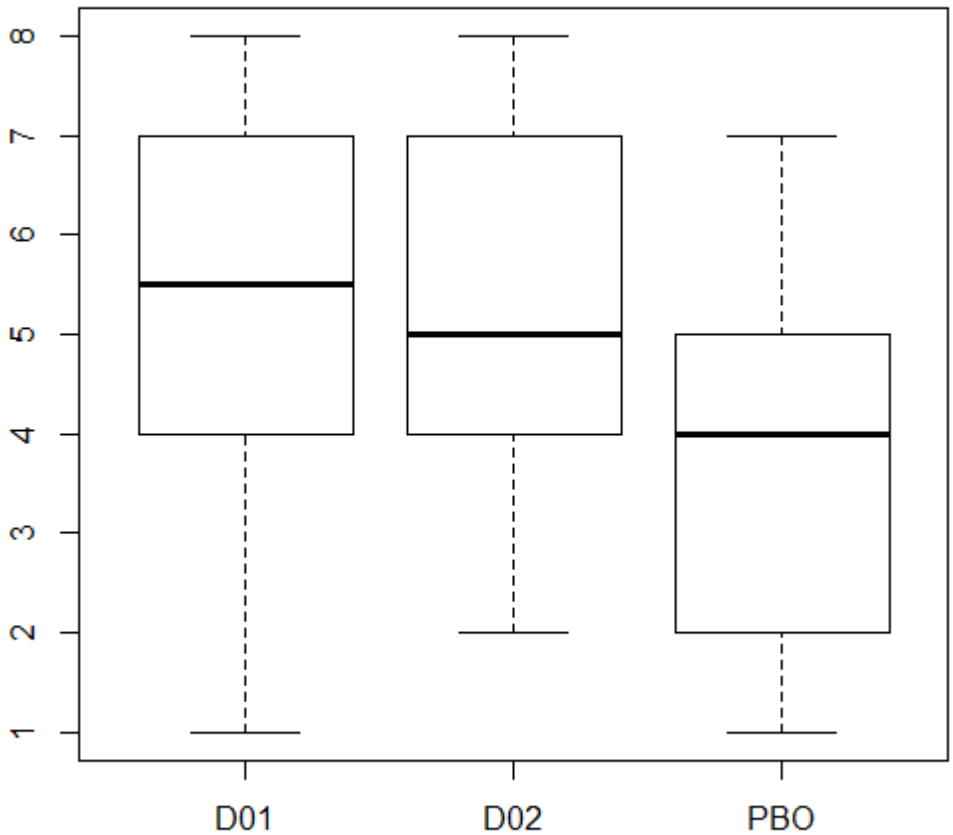
Data are ranked, from lowest to highest over the combined samples

Response Category	Code	---- Frequencies ---- 0.1%	0.2%	Pbo	Ranks	Ave. Rank	Total Fre-quency (m)	$c = m(m^2-1)$
<0%	1	1	0	2	1-3	2	3	24
0%	2	0	1	2	4-6	5	3	24
1-10%	3	1	0	1	7-8	7.5	2	6
11-25%	4	2	3	3	9-16	12.5	8	504
26-50%	5	1	2	3	17-22	19.5	6	210
51-75%	6	1	1	0	23-24	23.5	2	6
76-99%	7	3	1	1	25-29	27	5	120
100%	8	1	2	0	30-32	31	3	24
		10	10	12			32	C = 918

Since the code 1-8 is arbitrary, K-W test is appropriate because the results do not depend on the magnitude of the coded values, but in their ranks.

Kruskal-Wallis test – Normality -

```
> boxplot(SCORE~DOSE)
```



The student is invited to contrast whether or not the distributions deviate from normality

Kruskal-Wallis test – table of ranks and calculus -

0.1% Solution		0.2% Solution		Placebo	
Pat No.	Cate-gory Rank	Pat No.	Cate-gory Rank	Pat No.	Cate-gory Rank
1	19.5	3	19.5	2	19.5
6	12.5	5	31	4	7.5
9	2	7	5	8	27
12	27	10	31	11	2
15	12.5	14	27	13	5
19	7.5	18	12.5	16	12.5
20	23.5	22	19.5	17	5
23	27	26	12.5	21	2
27	31	28	23.5	24	12.5
32	27	31	12.5	25	19.5
				29	12.5
				30	19.5
$R_1 = 189.5$ ($n_1=10$)		$R_2 = 194.0$ ($n_2=10$)		$R_3 = 144.5$ ($n_3=12$)	

$$R_i = \sum_{j=1}^{n_i} r_{ij}; \quad \overline{R}_i = R_i / n_i; \quad \overline{R} = (N+1)/2; \quad N = \sum_i n_i$$

When H_0 is true (means do not differ), the average rank for each group should be close to the overall average rank. The Kruskal-Wallis test is

$$h^* = \frac{12}{N(N+1)}\left[\sum_{i=1}^{k} \frac{R_i^2}{n_i}\right] - 3(N+1)$$

$$= \frac{12}{32(33)}\left[\frac{189.5^2}{10} + \frac{194.0^2}{10} + \frac{144.5^2}{12}\right] - 3(33)$$

$$= 4.348$$

and correcting for ties,

$$h = \frac{4.348}{\left[1 - \frac{918}{32(32^2-1)}\right]} = \frac{4.348}{0.972} = 4.473$$

H_0 is not rejected, because $h < \chi^2_{2\,(0.05)}$ (=5.991)

Kruskal-Wallis test – R program and results -

```
> kruskal.test(SCORE~DOSE)

    Kruskal-Wallis rank sum test

data:  SCORE by DOSE
Kruskal-Wallis chi-squared = 4.4737,
df = 2, p-value = 0.1068
```

h corrected for ties

Not significant

When the K-W test is significant, pair-wise comparisons can be carried out with the Wilcoxon Rank Sum Test for each pair of groups.

Some comments on non-parametric comparisons

1. The *t*-test is a more powerful test in detecting true differences when data are normally distributed. Since normality occurs often in nature, the *t*-test is the method of choice for a wide range of applications.
2. The WRST does assume the two population distributions have the same shape and differ only by a possible shift in location. Thus we assume the same dispersion, which is analogous to the variance homogeneity required in two sample *t*-tests.
3. The Mann-Whitney U-test is mathematically equivalent to the WRST.
4. The WSRT does require the assumption of symmetrical underlying distribution. When the data are highly skewed or non symmetrical, the Sign test can be used.
5. One-sample *t*-test can be used for larger samples (n>30) regardless the distribution of data. But, the *t*-test should be used only if the mean is the appropriate measure of central tendency for the population being studied.
6. When you perform a WRST to compare each pair of means after a K-W test, the problem of overall error rate alteration must be considered for larger values of k (number of comparisons).

χ^2 test – association / independence -

The general layout is a r-by-c **contingence table**, with a total of r×c cells. A contingency table express the possibility that something happens or not.

The **null hypothesis** is that of **random distributions** among the levels of the two factors, or more generally, independence of row and column factors.

The χ^2 test is based on the tests statistic

$$\chi^2 = \sum_{i=1}^{kg} \frac{(O_i - E_i)^2}{E_i}$$

where O_i and E_i represent the observed and expected values, respectively, in the i-th cell.

The expected value is computed under the assumption that the null hypothesis is true, found by multiplying the marginal frequencies by the total sample size, N.

The test statistic is compared with the critical chi-square value with (r-1)×(c-1) degrees of freedom to obtain the decision rule

Caution must be used when cell sizes are small.

χ^2 - Contingency table – data -

Data used come from a study relating eye and hair colours in men of Baden country.

	HBLACK	HBLOND	HBROWN	HRED	Total
EBLUE	**189**	**1768**	**807**	**47**	2811 (0.4134)
EBROWN	**288**	**115**	**438**	**16**	857 (0.1260)
EGREEN	**746**	**946**	**1387**	**53**	3132 (0.4606)
Total	1223 (0.1799)	2829 (0.4160)	2632 (0.3871)	116 (0.0171)	

Marginal frequencies expressed in absolute -1223- or relative terms: 1223 / 6800 = 0.1799

χ^2 - Contingency table – table in R -

First we define the elements of a matrix by the c() command. The order is the elements of each row from left to right, starting with the first row and so on. Then we define that the matrix has 3 rows and that the elements are ordered within row.

```
> BADEN.MEN <- matrix(c(189,1768,807,47,288,115
+ ,438,16,746,946,1387,53),nrow=3, byrow=T)
```

After that we define the names of the column and row headers.

```
> colnames(BADEN.MEN) <- c("HBLACK","HBLOND","HBROWN","HRED")
> rownames(BADEN.MEN) <- c("EBLUE","EBROWN","EGREEN")
```

χ^2 test – results in R (1) -

```
> chisq.test(BADEN.MEN)$observed

       HBLACK HBLOND HBROWN HRED
EBLUE     189   1768    807   47
EBROWN    288    115    438   16
EGREEN    746    946   1387   53

> chisq.test(BADEN.MEN)$expected

         HBLACK    HBLOND    HBROWN     HRED
EBLUE  505.5666 1169.4587 1088.0224 47.95235
EBROWN 154.1340  356.5372  331.7094 14.61941
EGREEN 563.2994 1303.0041 1212.2682 53.42824
```

The **expected value** is calculated as

116/6800*3132/6800 * 6800 = 53.428

χ^2 test – results in R (2) -

```
> O <- chisq.test(BADEN.MEN)$observed
> E <- chisq.test(BADEN.MEN)$expected
> (O-E)
          HBLACK    HBLOND    HBROWN       HRED
EBLUE  -316.5666  598.5413 -281.0224 -0.9523529
EBROWN  133.8660 -241.5372  106.2906  1.3805882
EGREEN  182.7006 -357.0041  174.7318 -0.4282353

> chisq.test(BADEN.MEN)

   Pearson's Chi-squared test

data:  BADEN.MEN
X-squared = 1073.508, df = 6, p-value < 2.2e-16
```

The colour of the eyes and the colour of the hair are **not independent**, i.e., not distributed at random among each other.

```
> 1-pchisq(1073.508,6)
[1] 0
```

χ^2 test – results in R (3) -

```
> assocstats(BADEN.MEN)

                     X^2 df P(> X^2)
Likelihood Ratio 1137.6  6        0
Pearson          1073.5  6        0

Phi-Coefficient   : 0.397
Contingency Coeff.: 0.369
Cramer's V        : 0.281
```

H_0 of random distribution of colours of eyes and hair is rejected

Measurements of the degree of association

The measurements of the degree of association are derived from the χ^2 estimate. These measurements are similar in concept to a correlation coefficient between variables.

For 2×2 designs (1 degree of freedom) and cells with less than 5 observations, use the Fisher Exact Test (write `fisher.test`).

References

Walker G.A. 1997. *Common Statistical Methods for Medical Research.* SAS Institute, Cary, NC.