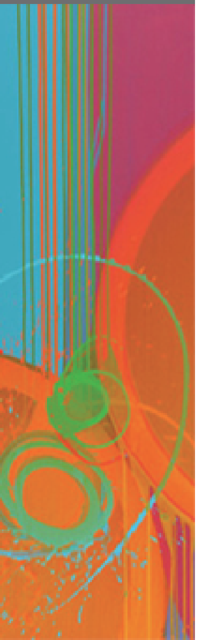
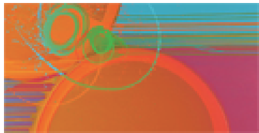
DISCOVERING STATISTICS USING R

Non-parametric Tests



ANDY FIELD

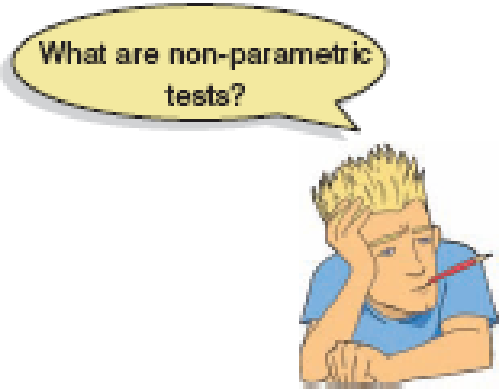


Aims

- When and why do we use non-parametric tests?
  - Wilcoxon rank-sum test (comparing two indep samples)
  - Wilcoxon signed-rank test (comparing paired samples)
  - Kruskal–Wallis test (one-way layout)
  - Jonckheere–Terpstra test (test for ordered alternative)
  - Friedman's ANOVA (clustered one-way layout)
- Ranked based tests
- Interpretation of results
- Reporting results
- Calculating an effect size

When to Use Nonparametric Tests

- Non-parametric tests are used when assumptions of parametric tests are not met.
- Small sample sizes



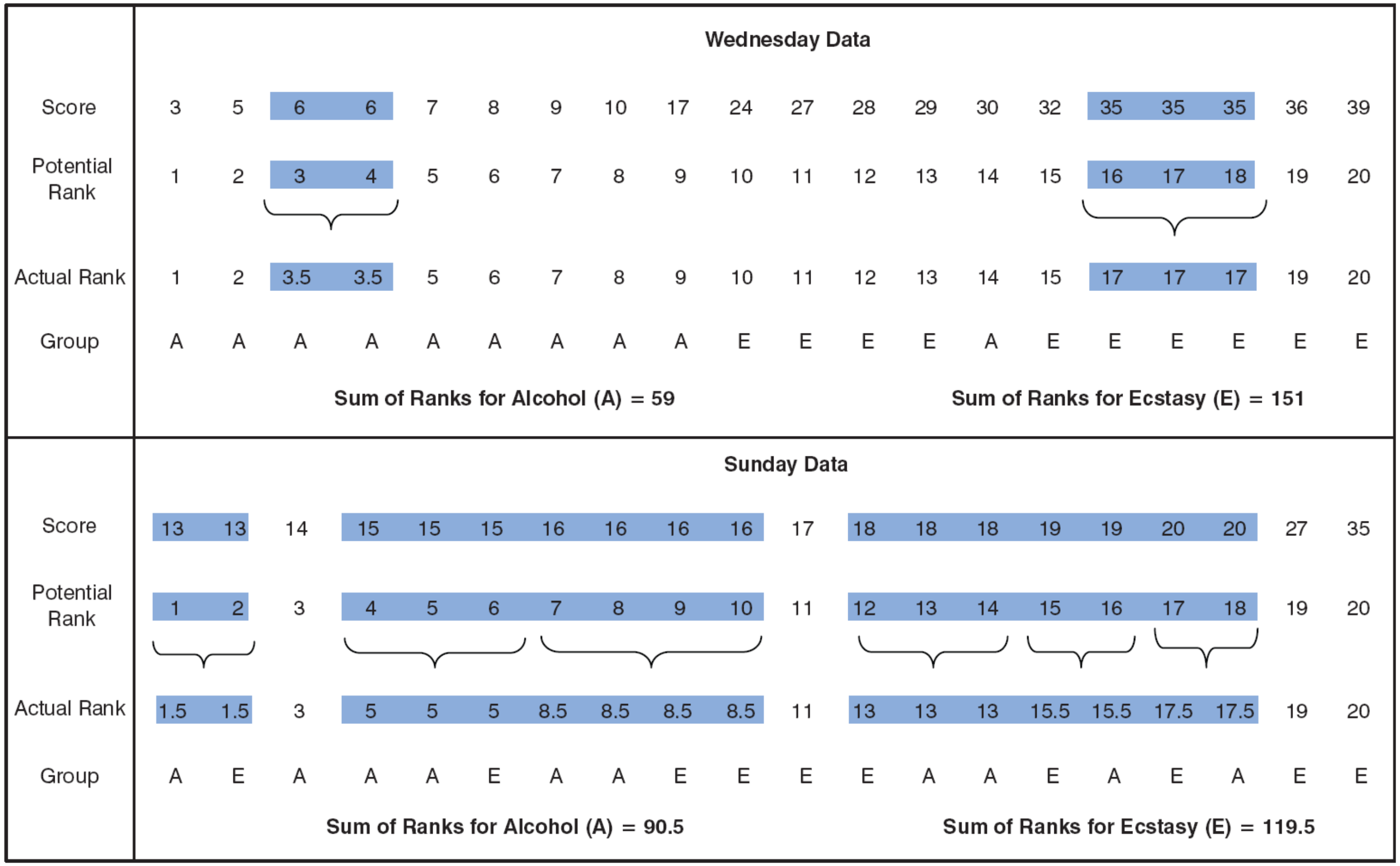
The Wilcoxon rank-sum test

- The non-parametric equivalent of the independent *t-test.*
- *Assumptions:*
  - *The distributions in the two groups must be the same, other than a shift in location; that is, the distributions should have the same variance and skewness*
  - *Independent observations*
- Equivalent to the Mann-Whitney U test

An Example

- A neurologist investigated the depressant effects of certain recreational drugs.
  - Tested 20 clubbers.
  - 10 were given an ecstasy tablet to take on a Saturday night.
  - 10 were allowed to drink only alcohol.
  - Levels of depression were measured using the Beck Depression Inventory (BDI) the day after and midweek.
- Rank the data *ignoring* the group to which a person belonged.
  - A similar number of high and low ranks in each group suggests depression levels do not differ between the groups.
  - A greater number of high ranks in the ecstasy group than the alcohol group suggests the ecstasy group is more depressed than the alcohol group.

Ranking the Depression Scores for Wednesday and Sunday



Wednesday Data

Score	3	5	6	6	7	8	9	10	17	24	27	28	29	30	32	35	35	35	36	39
Potential Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Actual Rank	1	2	3.5	3.5	5	6	7	8	9	10	11	12	13	14	15	17	17	17	19	20
Group	A	A	A	A	A	A	A	A	A	E	E	E	E	A	E	E	E	E	E	E

Sum of Ranks for Alcohol (A) = 59 **Sum of Ranks for Ecstasy (E) = 151**

Sunday Data

Score	13	13	14	15	15	15	16	16	16	16	17	18	18	18	19	19	20	20	27	35
Potential Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Actual Rank	1.5	1.5	3	5	5	5	8.5	8.5	8.5	8.5	11	13	13	13	15.5	15.5	17.5	17.5	19	20
Group	A	E	A	A	A	E	A	A	E	E	E	E	A	A	E	A	E	A	E	E

Sum of Ranks for Alcohol (A) = 90.5 **Sum of Ranks for Ecstasy (E) = 119.5**

FIGURE 15.3 Ranking the depression scores for Wednesday

wilcox.test {base}

```
> newModel <- wilcox.test(outcome ~ predictor,
data = dataFrame, paired = FALSE)
```

wilcox.test {base}

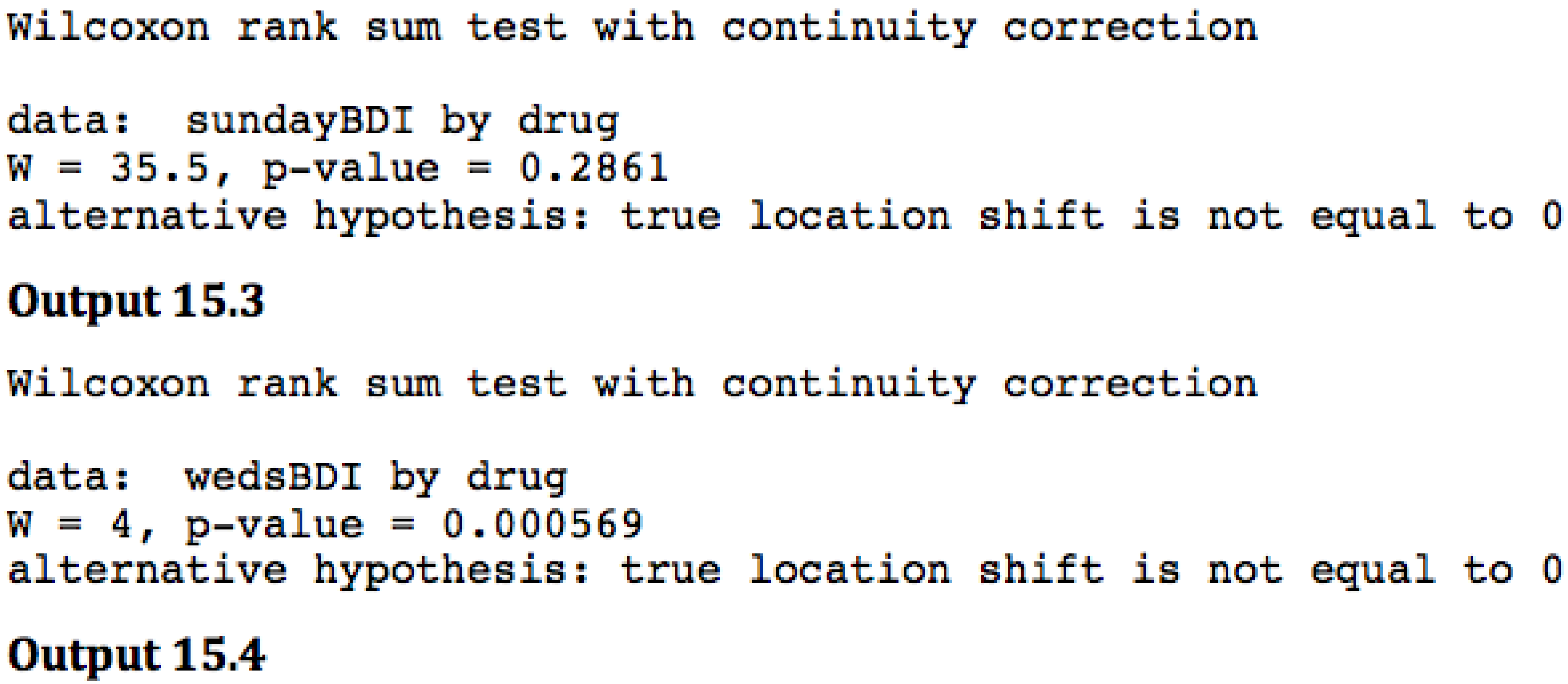
- To compute a basic Wilcoxon test for our Sunday data we could execute:

```
> wilcox.test(sundayBDI ~ drug, data = drugData)
```

- For the Wednesday data:

```
> wilcox.test(wedsBDI ~ drug, data = drugData)
```

Output from the Wilcoxon Rank-Sum Test



```
Wilcoxon rank sum test with continuity correction

data:  sundayBDI by drug
W = 35.5, p-value = 0.2861
alternative hypothesis: true location shift is not equal to 0
```

Output 15.3

```
Wilcoxon rank sum test with continuity correction

data:  wedsBDI by drug
W = 4, p-value = 0.000569
alternative hypothesis: true location shift is not equal to 0
```

Output 15.4

Reporting the Results

- Depression levels in ecstasy users (*median* = 17.50) did not differ significantly from alcohol users (*median* = 16.00) the day after the drugs were taken, *p* = .286. However, by Wednesday, ecstasy users (*median* = 33.50) were significantly more depressed than alcohol users (*median* = 7.50), *p =0.0005*.

Power issues

- Why not always use a non-parametric test?
  - When the assumptions of normality and equal var is met, then highest power is achieved with a parametric test
  - However, the Wilcoxon rank sum test achieves almost equal power as the two-sample t-test
  - 8 data points for the 2 groups combined are minimally required to achieve p-values < 0.05

Comparing Two Related Conditions: the Wilcoxon Signed-Rank Test

- Uses:
  - To compare two sets of scores, when these scores come from the same participants.
  - H_0: difference between the pairs follows a symmetrical distribution around zero
  - H_A: difference between the pairs does not follow a symmetrical distribution around zero
- Assumptions
  - The scale of measurement is at least ordinal.
  - The difference between the two measures has a symmetrical distribution.
  - Other than the obvious pairing, the observations are independent.

Wilcoxon Signed-Rank Test

- Imagine the experimenter in the previous example was interested in the change in depression levels between Sunday and Wednesday for each of the two drugs.
  - We still have to use a non-parametric test because the distributions of scores for both drugs were non-normal on one of the two days.

wilcox.test {stats}

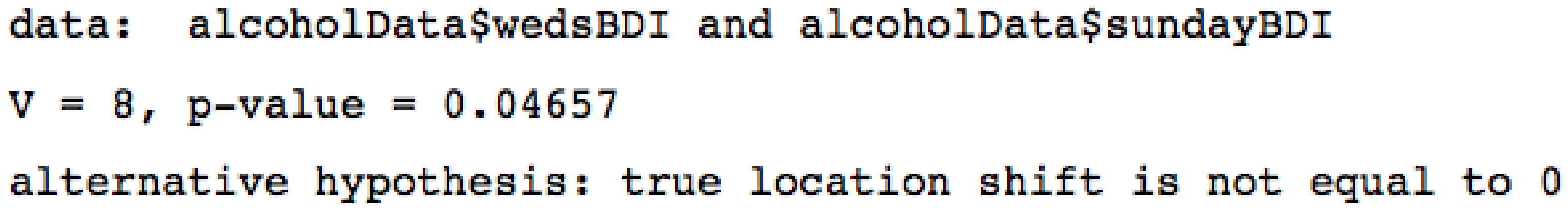
- We want to run our analysis on the alcohol and ecstasy groups separately; therefore, our first job is to split the dataframe into two using the *subset()* function:

```
> alcoholData <- subset(drugData, drug == "Alcohol")
> ecstasyData <- subset(drugData, drug == "Ecstacy")
```

wilcox.test {stats}

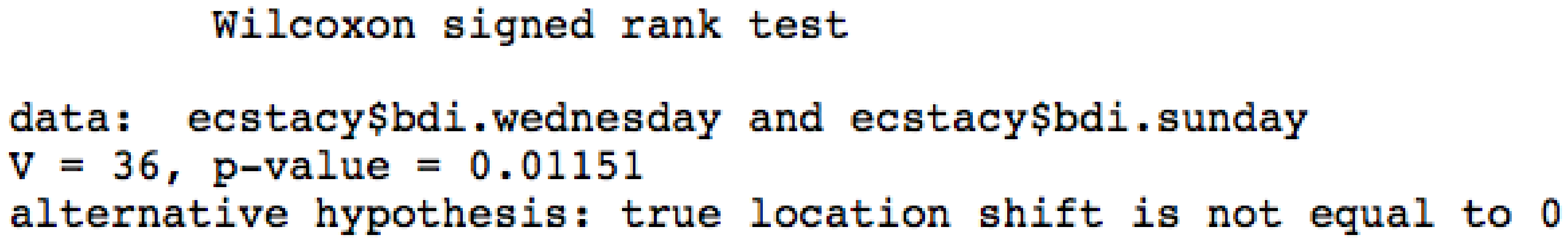
- To run the analysis for the alcohol group execute:
  ```
  wilcox.test(alcoholData$wedsBDI,
  alcoholData$sundayBDI, paired = TRUE,
  correct= FALSE)
  ```
- and for the ecstasy group:
  ```
  wilcox.test(ecstasyData$wedsBDI,
  ecstasyData$sundayBDI, paired = TRUE,
  correct= FALSE)
  ```

Output



```
data:  alcoholData$wedsBDI and alcoholData$sundayBDI

V = 8, p-value = 0.04657

alternative hypothesis: true location shift is not equal to 0
```

Output 15.6



```
        Wilcoxon signed rank test

data:  ecstacy$bdi.wednesday and ecstacy$bdi.sunday
V = 36, p-value = 0.01151
alternative hypothesis: true location shift is not equal to 0
```

Output 15.7

Reporting the results

  - For ecstasy users, depression levels were significantly higher on Wednesday (*median* = 33.50) than on Sunday (*median* = 17.50), *p* = .047. However, for alcohol users the opposite was true: depression levels were significantly lower on Wednesday (*median* = 7.50) than on Sunday (*median* = 16.0), *p* = .012.

- Note:
  - with N = 5, no significant p-value can be obtained. When all 5 differences are in the same direction, p = .0625

Differences between Several Independent Groups: the Kruskal–Wallis test

- The Kruskal–Wallis test (Kruskal & Wallis, 1952) is the non-parametric counterpart of the one-way independent ANOVA.
  - If you have data that have violated an assumption then this test can be a useful way around the problem.
- The theory for the Kruskal–Wallis test is very similar to that of the Wilcoxon rank-sum test:
  - The Kruskal–Wallis test is based on ranked data.
  - The Kruskal–Wallis test compares groups on the mean rank of a variable that is at least ordinal

Kruskal–Wallis

- Assumptions:
  - The distributions in the groups must be the same, other than a shift in location.
  - The distributions should have roughly the same variance. (check boxplots, Brown-Forsythe test)

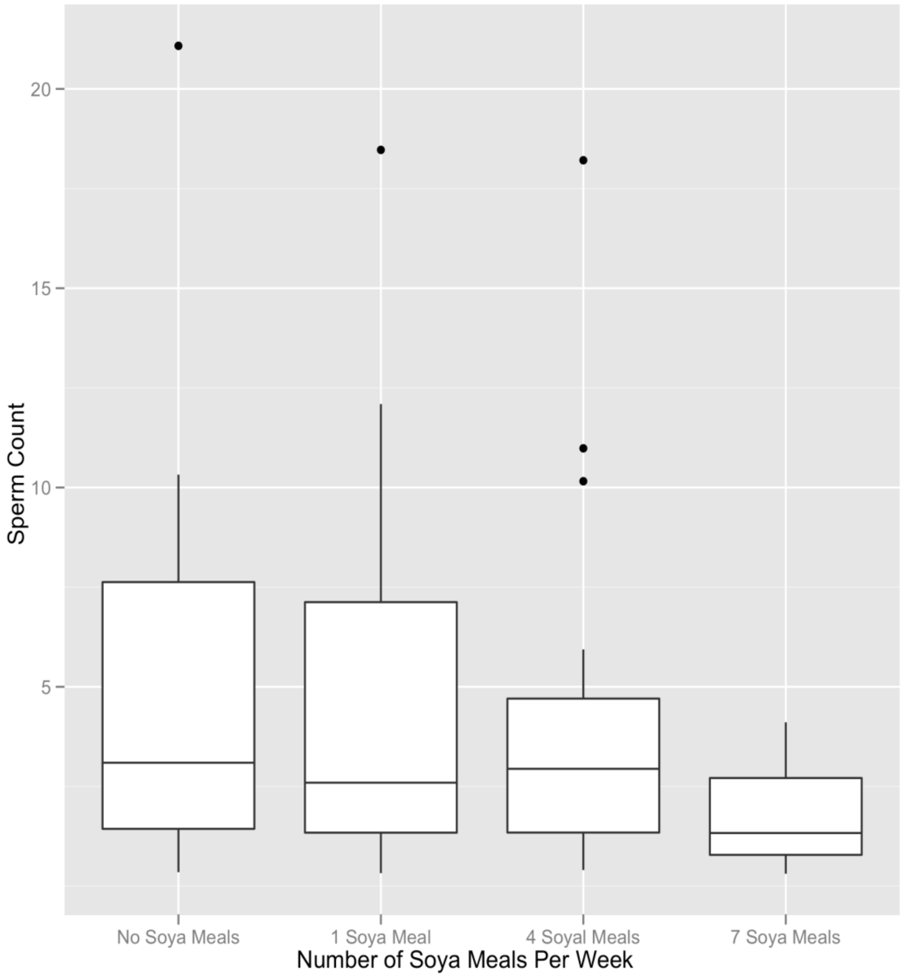
Post-hoc test

- pairwise.wilcox.test()
  - Calculates pairwise comparisons between group levels with corrections for multiple testing
- kruskalmc()
  - comparisons between treatments
    - kruskalmc(resp, catvar,cont=“NULL”)
  - 'one-tailed' and 'two-tailed' comparison treatments versus control.
    - kruskalmc(resp, catvar, cont="one-tailed")
    - kruskalmc(resp, catvar, cont=“two-tailed")

Example

- Does eating soya affect your sperm count?
- Variables
  - Outcome: sperm (millions)
  - IV: Number of soya meals per week
    - No Soya meals
    - 1 Soya meal
    - 4 soya meals
    - 7 soya meals
- Participants
  - 80 males (20 in each group)

Boxplot for the Sperm Counts of Individuals Eating Different Numbers of Soya Meals per Week



Data for the Soya Example with Ranks

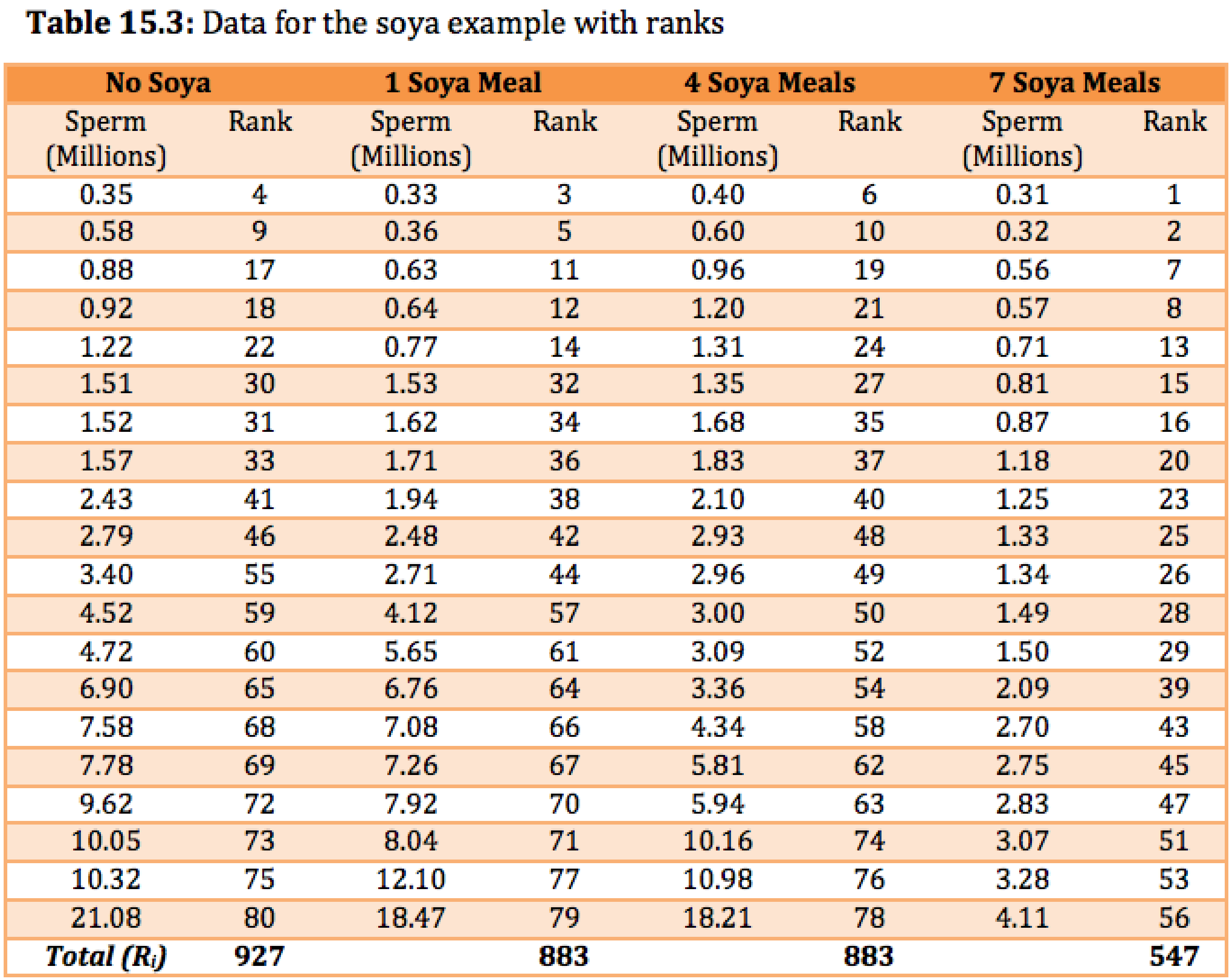


Table 15.3: Data for the soya example with ranks

No Soya		1 Soya Meal		4 Soya Meals		7 Soya Meals	
Sperm (Millions)	Rank	Sperm (Millions)	Rank	Sperm (Millions)	Rank	Sperm (Millions)	Rank
0.35	4	0.33	3	0.40	6	0.31	1
0.58	9	0.36	5	0.60	10	0.32	2
0.88	17	0.63	11	0.96	19	0.56	7
0.92	18	0.64	12	1.20	21	0.57	8
1.22	22	0.77	14	1.31	24	0.71	13
1.51	30	1.53	32	1.35	27	0.81	15
1.52	31	1.62	34	1.68	35	0.87	16
1.57	33	1.71	36	1.83	37	1.18	20
2.43	41	1.94	38	2.10	40	1.25	23
2.79	46	2.48	42	2.93	48	1.33	25
3.40	55	2.71	44	2.96	49	1.34	26
4.52	59	4.12	57	3.00	50	1.49	28
4.72	60	5.65	61	3.09	52	1.50	29
6.90	65	6.76	64	3.36	54	2.09	39
7.58	68	7.08	66	4.34	58	2.70	43
7.78	69	7.26	67	5.81	62	2.75	45
9.62	72	7.92	70	5.94	63	2.83	47
10.05	73	8.04	71	10.16	74	3.07	51
10.32	75	12.10	77	10.98	76	3.28	53
21.08	80	18.47	79	18.21	78	4.11	56
Total (R_i)	**927**		**883**		**883**		**547**

kruskal.test{stats}

- For the current data:

```
kruskal.test(Sperm ~ Soya, data = soyaData)
```

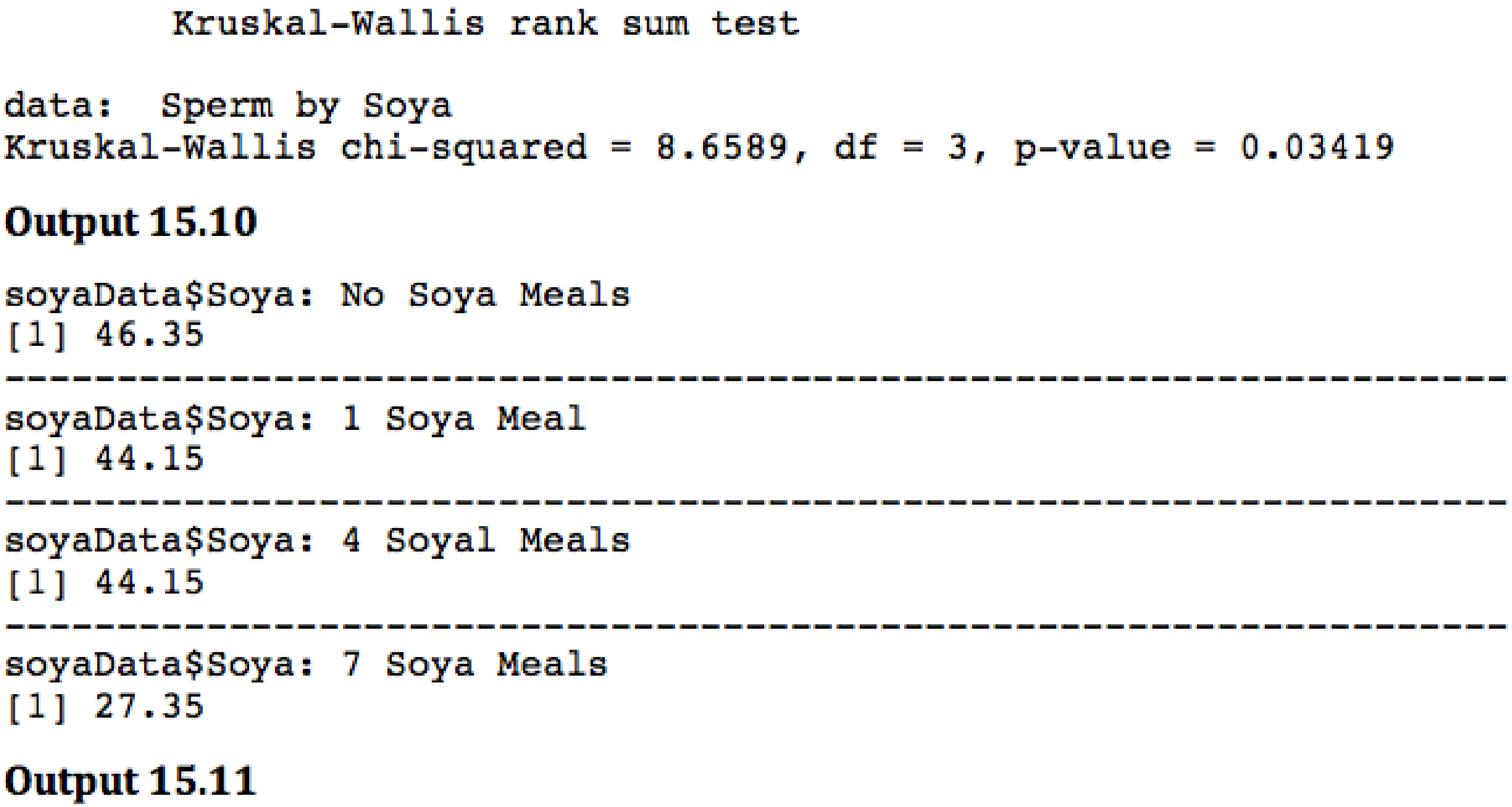
- To interpret the Kruskal–Wallis test, it is useful to obtain the mean rank for each group:

```
soyaData$Ranks<-rank(soyaData$Sperm)
```

- This command creates a variable **Ranks** in *soyaData* dataframe that is the ranks for the variable **Sperm**. We can then obtain the mean rank for each group:

```
by(soyaData$Ranks, soyaData$Soya, mean)
```

Output from the Kruskal–Wallis test



```
        Kruskal-Wallis rank sum test

data:  Sperm by Soya
Kruskal-Wallis chi-squared = 8.6589, df = 3, p-value = 0.03419
```

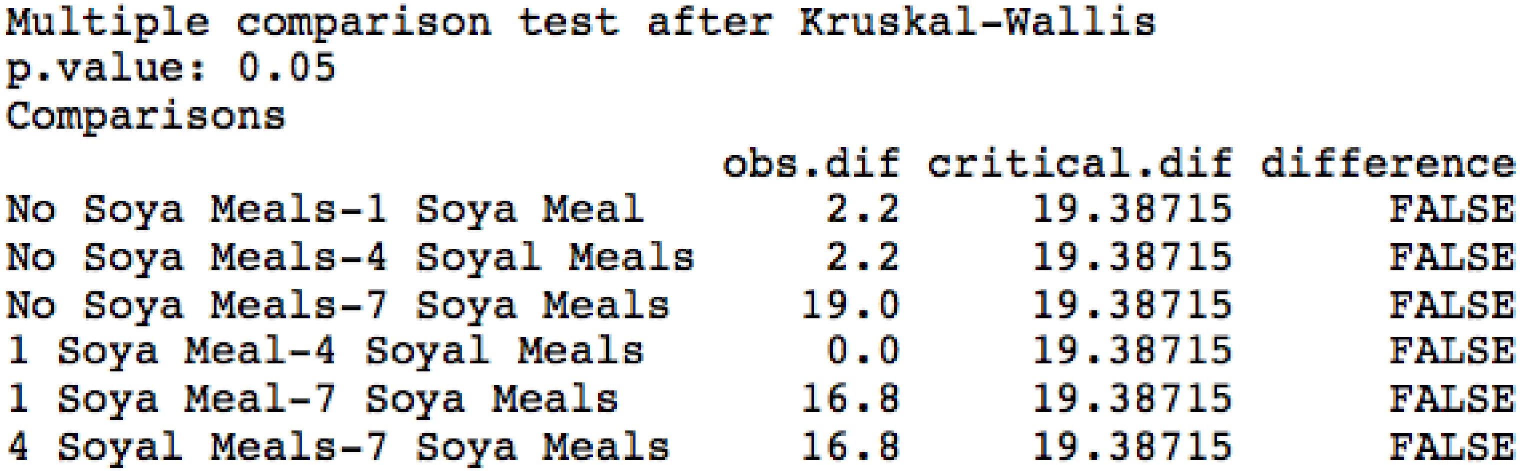
Output 15.10

```
soyaData$Soya: No Soya Meals
[1] 46.35
-------------------------------------------------------------------
soyaData$Soya: 1 Soya Meal
[1] 44.15
-------------------------------------------------------------------
soyaData$Soya: 4 Soyal Meals
[1] 44.15
-------------------------------------------------------------------
soyaData$Soya: 7 Soya Meals
[1] 27.35
```

Output 15.11

Post Hoc Tests for the Kruskal–Wallis Test: kruskalmc{pgirmess}

- kruskalmc(Sperm ~ Soya, data = soyaData)



```
Multiple comparison test after Kruskal-Wallis
p.value: 0.05
Comparisons
                                obs.dif critical.dif difference
No Soya Meals-1 Soya Meal           2.2     19.38715      FALSE
No Soya Meals-4 Soyal Meals         2.2     19.38715      FALSE
No Soya Meals-7 Soya Meals         19.0     19.38715      FALSE
1 Soya Meal-4 Soyal Meals           0.0     19.38715      FALSE
1 Soya Meal-7 Soya Meals           16.8     19.38715      FALSE
4 Soyal Meals-7 Soya Meals         16.8     19.38715      FALSE
```

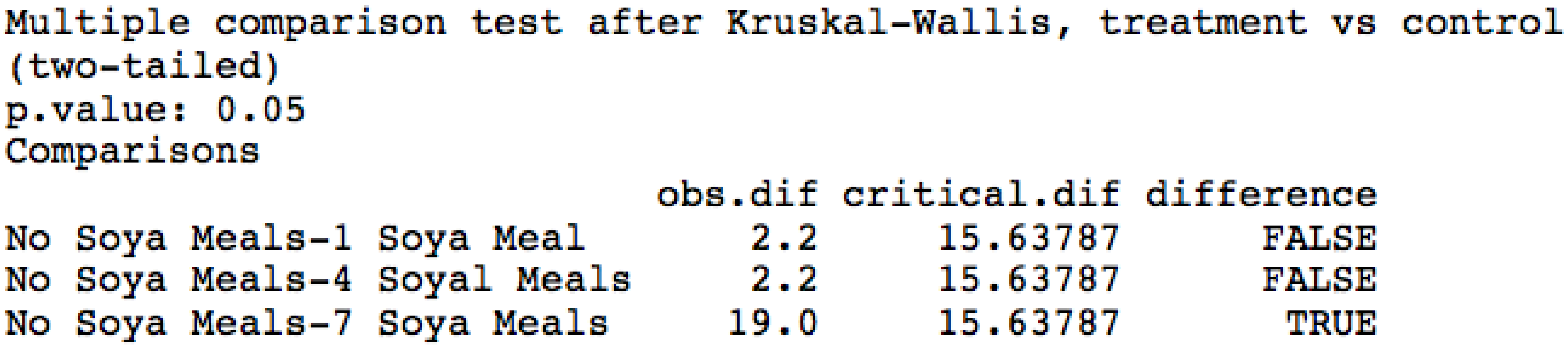
Post Hoc Tests for the Kruskal–Wallis Test

- One of the problems with comparing every group against all others is that we have to be quite strict about accepting a difference as significant otherwise we will inflate the Type I error rate. To reduce this problem we could use more focussed comparisons.
- In this example, we have a control group that had no soya meals. As such, a nice succinct set of comparisons would be to compare each group against the control:
  - Test 1: one soya meal per week compared to no soya meals
  - Test 2: four soya meal per week compared to no soya meals
  - Test 3: seven soya meal per week compared to no soya meals
- Yo compare each group to the no-soya group (using a two-tailed test) we simply execute:

```
kruskalmc(Sperm ~ Soya, data = soyaData, cont = 'two-tailed')
```

(Note: The first factor level is considered the control.)

Output



```
Multiple comparison test after Kruskal-Wallis, treatment vs control
(two-tailed)
p.value: 0.05
Comparisons
                            obs.dif critical.dif difference
No Soya Meals-1 Soya Meal       2.2     15.63787      FALSE
No Soya Meals-4 Soyal Meals     2.2     15.63787      FALSE
No Soya Meals-7 Soya Meals     19.0     15.63787       TRUE
```

Parametric tests on ranks

- You can convert the original values to ranks and analyze the rank-values using a t-test or ANOVA. This method does not give identical results to using the Wilcoxon tests or Kruskal-Wallis, but the results are quite similar, and the method is an acceptable alternative if you don't have software that will do the rank tests

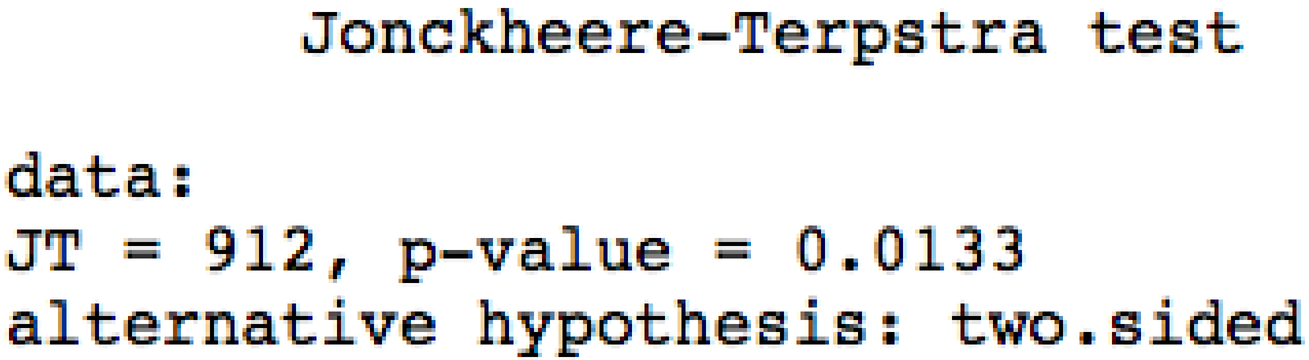
Testing for Trends: the Jonckheere–Terpstra Test

- This statistic tests for an ordered pattern to the medians of the groups you're comparing.
- Essentially it does the same thing as the Kruskal–Wallis test but it incorporates information about whether the order of the groups is meaningful.
  - Use this test when you expect the groups you're comparing to produce a meaningful order of medians.
  - In the current example we expect that the more soya a person eats, the more their sperm count will go down.

jonckheere.test{clinfun}

- We can conduct a Jonckheere test by executing:

```
jonckheere.test(soyaData$Sperm,
as.numeric(soyaData$Soya))
```



```
                Jonckheere-Terpstra test

data:
JT = 912, p-value = 0.0133
alternative hypothesis: two.sided
```

Rank test for blocked one-way lay-outs: Friedman

- Example
  - Does the Atkins diet work?
  - Variables
    - Outcome: weight (kg)
    - IV: time since beginning the diet
      - Baseline
      - 1 month
      - 2 months
- Participants
  - 10 women

Table 15.5: Data for the diet example with ranks

	Weight				Weight		
	Start	**Month 1**	**Month 2**		**Start (Ranks)**	**Month 1 (Ranks)**	**Month 2 (Ranks)**
Person 1	63.75	65.38	81.34		1	2	3
Person 2	62.98	66.24	69.31		1	2	3
Person 3	65.98	67.70	77.89		1	2	3
Person 4	107.27	102.72	91.33		3	2	1
Person 5	66.58	69.45	72.87		1	2	3
Person 6	120.46	119.96	114.26		3	2	1
Person 7	62.01	66.09	68.01		1	2	3
Person 8	71.87	73.62	55.43		2	3	1
Person 9	83.01	75.81	71.63		3	2	1
Person 10	76.62	67.66	68.60		3	1	2
				R_i	19	20	21

Theory of Friedman's ANOVA

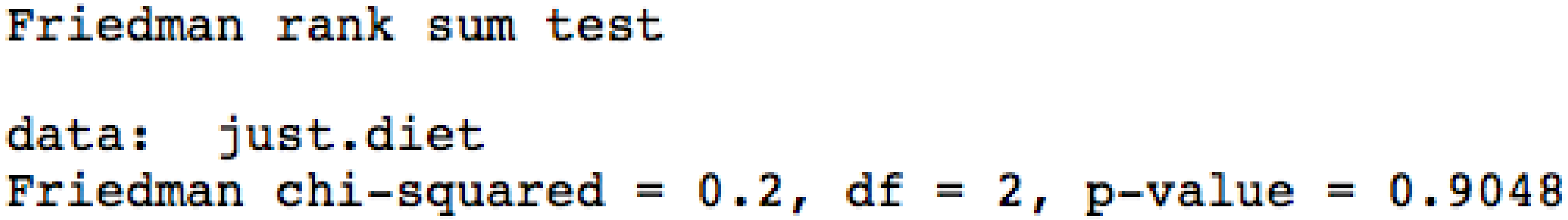
- The theory for Friedman's ANOVA is much the same as the other tests:
  - it is based on ranked data.
  - Assumes that the location shift model holds

friedman.test{stats}

- To run the Friedman test we simply input the name of our dataframe, but within the *as.matrix()* function, which converts it to a matrix. In this example, we would execute:

```
friedman.test(as.matrix(dietData))
```

Output from Friedman’s ANOVA



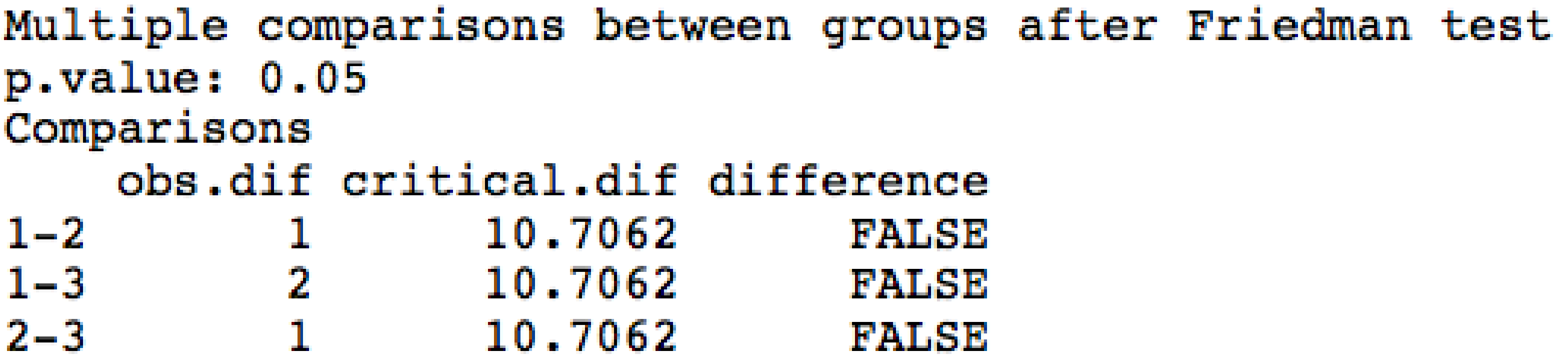
```
Friedman rank sum test

data:  just.diet
Friedman chi-squared = 0.2, df = 2, p-value = 0.9048
```

Output 15.16

Post Hoc Tests for Friedman's ANOVA {pgirmess}

- For the current data we would execute:

friedmanmc(as.matrix(dietData))



```
Multiple comparisons between groups after Friedman test
p.value: 0.05
Comparisons
    obs.dif critical.dif difference
1-2       1      10.7062      FALSE
1-3       2      10.7062      FALSE
2-3       1      10.7062      FALSE
```

To Sum Up …

- When data violate the assumptions of parametric tests we can sometimes find a non-parametric equivalent
  - Usually based on analysing the ranked data
- Wilcoxon rank-sum test
  - Compares two independent groups of scores
- Wilcoxon signed-rank test
  - Compares two dependent groups of scores
- Kruskal–Wallis test
  - Compares more than two *independent* groups of scores
- Friedman's test
  - Compares more than two *dependent* groups of scores

DEMO NONPARAMETRIC TESTS

Open the program Ch15_Nonparametric.R

Exercises

- Task 1
  - qPCR analysis was performed on untreated Arabidopsis samples and Arabidopsis samples treated with methyl-jasmonate. The data **1471-2105-7-85-S9.txt** contains Δ ΔCt values of 1 gene of interest. Verify whether there is differential expression with an appropiate test. The data comes from Yuan *et al*. 2006, BMC Bioinformatics.

Exercises

- Task 2
  - There's been much speculation over the years about the influence of subluminal messages on records. To name a few cases, both Ozzy Osbourne and Judas Priest have been accused of putting backward masked messages on their albums that subliminally influence poor unsuspecting teenagers into doing things like blowing their heads off with shotguns. A psychologist was interested in whether backward messages really did have an effect. He took the master tapes of Britney Spears 'Baby one More Time' and created a second version that had the masked message 'deliver your soul to the dark lord' repeated in the chorus. He took this version, and the original, and played one version (randomly) to a group of 32 people. He took the same group six months later and played them whatever version they hadn't heard the time before. So each person heard both the original, and the version with the masked message, but at different points in time. The psychologist measured the number of goats that were sacrificed in the week after listening to each version. It was hypothesized that the backward message would lead to more goats being sacrificed. The data are in the file **DarkLord.dat**. Analyse with an appropriate test.

Exercises

- Task 3
  - A psychologist was interested in the effects of television programs on domestic life. She hypothesized that through ‘learning by watching’, certain programs might actually encourage people to behave like the characters within them. This in turn could affect the viewer’s own relationships (depending on whether the program depicted harmonious or dysfunctional relationships). She took episodes of three popular TV shows and showed them to 54 couples, after which the couple were left alone in the room for an hour. The experimenter measured the number of times the couple argued. Each couple viewed all three of the programs at different points in time (a week apart) and the order in which the programs were viewed was counterbalanced over couples. The TV programs selected were *Eastenders* (which typically portrays the lives of extremely miserable, argumentative, London folk who like nothing more than to beat each other up, lie to each other, sleep with each other’s wives and generally show no evidence of any consideration to their fellow humans!). *Friends* (which portrays a group of unrealistically considerate and nice people who love each other oh so very much), and a *National Geographic* program about whales (this was supposed to act as a control). The data are in the file **Eastenders.dat**. Conduct an appropriate analysis.

Exercises

- Task 4
  - suppose weights of poplar trees are different based on treatments (none treatment, fertilizer, irrigation, or fertilizer and irrigation). Each weight observation is independent and random, and each sample size is 5. But the weight observations are not normally distributed. The research question is to test whether the poplar tree weights are different under the four treatments. Conduct an appropriate test.
  - Data: **poplar.csv**
    - http://www.stat.purdue.edu/~tqin/system101/method/method_kruskal_wallis_sas.htm

Exercises

- Task 5
  - A winery wanted to find out whether people preferred red, white or rosé wines. They invited 12 people to taste one red, one white and one rose' wine with the order of tasting chosen at random and a suitable interval between tastings. Each person was asked to evaluate each wine with a score from 1 to 10 (maximum appreciation)
  - Data **winery.csv**
    - http://www.real-statistics.com/anova-repeated-measures/friedman-test/