Analysis of Quantitative data
Non-Parametric data

Anne Segonds-Pichon
v2020-12

Babraham Institute

Babraham Bioinformatics

Comparison between 2 groups
Non-Parametric data

Non-parametric test:

Mann-Whitney = Wilcoxon rank test

- Non-parametric equivalent of the *t*-test (and not).
- Not meeting the assumptions for parametric tests is not enough to switch to a non-parametric approach.
  - Like always, data exploration is key.
  - The outcome is a rank or a score with limited amount of possible values: non-parametric approach.

- **How does the Mann-Whitney test work?**

Group 1	Group 2
5	8
7	9
3	6

→

Real values	Ranks
3	1
5	2
6	3
7	4
8	5
9	6

→

	Group 1	Group 2
	2	5
	4	6
	1	3
Sum	7	14

- Statistic of the Mann-Whitney test: **U (W)**
  $\mathbf{U_1}$ = 7-6 = 1 and $\mathbf{U_2}$ = 14-6 = 8
  - Smallest of the 2 Us: U_1 + sample size → **p-value**

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}$$

Where:
- R = sum of ranks
- n = sample size.

- **R:** `wilcox_test()`

Non-parametric test:
Wilcoxon's signed-rank

- Non-parametric equivalent of the paired *t*-test (ish).
- **How does the test work?**

Before	After	Differences
9	3	-6
7	4	-3
10	4	-6
8	5	-3
5	6	1
8	2	-6
7	7	0
9	4	-5
10	5	-5

Abs. Diff.	Ranking	Ranks
0		
1	1	1
3	2	2.5
3	3	2.5
5	4	4.5
5	5	4.5
6	6	7
6	7	7
6	8	7

2+3=5/2=2.5: average rank

	Negative ranks	Positives ranks
		1
	-2.5	
	-2.5	
	-4.5	
	-4.5	
	-7	
	-7	
	-7	
Sum	-35	1

- Statistic of the Wilcoxon's signed-rank test: Sum of signed ranks = **W**
  - Here: W = -35 + 1 = -34
  - Statistic W + sample size → **p-value**
- **R:** `wilcox_test(paired = TRUE)`

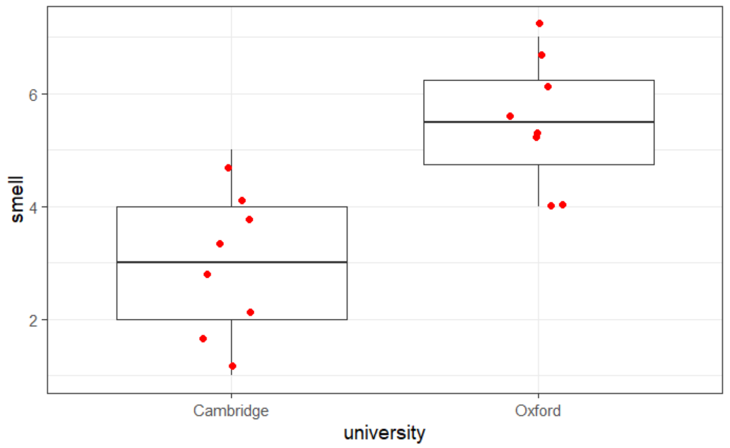
Exercise 13: Independent test

smelly.teeshirt.csv

- Hypothesis: Group body odour is less disgusting when associated with an in-group member versus an out-group member. Two groups of Cambridge University students are presented with one of two smelly, worn t-shirts with university logos.
- **Question**: are Cambridge students more disgusted by worn smelly T-shirts from Oxford or Cambridge? Disgust score: 1 to 7, with 7 the most disgusting
  - Load **smelly.teeshirt.csv**
  - Explore the data with an appropriate combination of 2 graphs
  - Answer the question with a non-parametric approach

Exercise 13: smelly.teeshirt. csv

- **Question**: are Cambridge students more disgusted by worn smelly T-shirts from Oxford or Cambridge?
  - Disgust score: 1 to 7, with 7 the most disgusting




```
read_csv("smelly.teeshirt.csv") -> smelly.teeshirt

smelly.teeshirt %>%
  ggplot(aes(x=university, y=smell))+
  geom_boxplot()+
  geom_jitter(height=0, width=0.1, size=2, colour="red")
```

```
smelly.teeshirt %>%
  wilcox_test(smell~university)
```

.y. <chr>	group1 <chr>	group2 <chr>	n1 <int>	n2 <int>	statistic <dbl>	p <dbl>
smell	Cambridge	Oxford	8	8	5	0.00479

Answer: T-shirts from Oxford are significantly more disgusting than the ones from Cambridge (W=5,p=0.0047).

What do you think of the design??

Exercise 14: Dependent test

botulinum.long.csv

A group of 9 disabled children with muscle spasticity (or extreme muscle tightness limiting movement) in their right upper limb underwent a course of injections with botulinum toxin to reduce spasticity levels.
A neurologist (blinded) assessed levels of spasticity pre- and post-treatment for all 9 children using a 10-point ordinal scale.
Higher ratings indicated higher levels of spasticity.

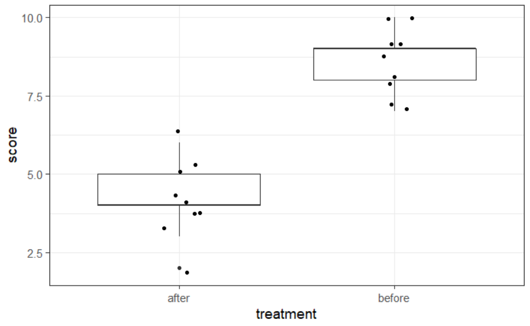
- **Question**: do botulinum toxin injections reduce muscle spasticity levels?
  - Score: 1 to 10, with 10 the highest spasticity

  - Load **botulinum.long.csv**
  - Plot the data
  - Answer the question with a non-parametric approach
  - Work out and plot the difference (after – before)

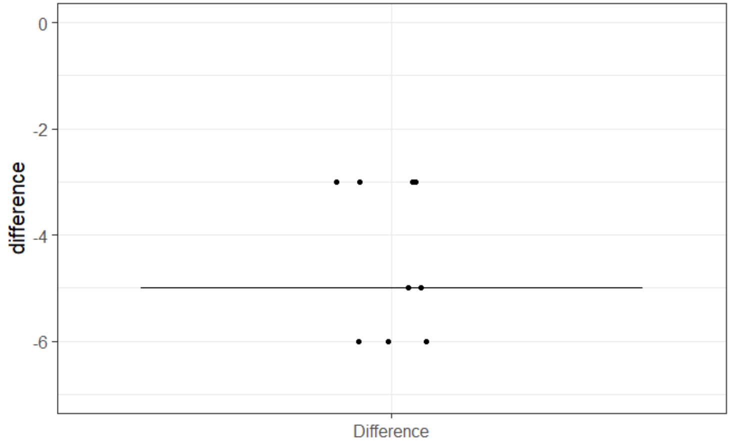
Exercise 14: Dependent test - botulinum.csv


```
read_csv("botulinum.long.csv") -> botulinum
```

```
botulinum %>%
  ggplot(aes(x=treatment, y=score))+
  geom_boxplot()+
  geom_jitter(height=0, width=0.1)
```




```
botulinum %>%
 mutate(difference = scores - scores[treatment == 'before']) %>%
    ggplot(aes("Difference", difference))+
        geom_jitter(height = 0, width=0.1)+
        stat_summary(geom="errorbar",
                 fun=median,fun.min=median,fun.max=median)+
        ylim(-7, 0)+
        xlab(NULL)
```




```
botulinum.long %>%
  wilcox_test(score~treatment, paired = TRUE)
```

.y. <chr>	group1 <chr>	group2 <chr>	n1 <int>	n2 <int>	statistic <dbl>	p <dbl>
score	after	before	9	9	0	0.00826

Answer: There was a significant difference pre- and post- treatment in ratings of muscle spasticity (p=0.008). *Note: T=V*

Comparison between more than 2 groups

One factor

Non-Parametric data

Kruskal-Wallis and Friedman tests

- Non-parametric equivalents of the One-Way ANOVA
  - Also based on ranks
  - **Kruskal-Wallis**: independent measures
  - **Friedman**: repeated measures

- Statistic associated with **Kruskal-Wallis** is **H**
- Statistic associated with **Friedman** is **Q** or **T1** or **FM**
- The statistics have a Chi^2 distribution
  - Kruskal-Wallis H = Friedman statistic = One-Way ANOVA F

- Post-hoc test associated with Kruskal-Wallis and Friedman: **Dunn's test**
  - The Dunn's test works pretty much like the Mann-Whitney test.

Kruskal-Wallis test: Example

- Creatine, a supplement popular among body builders
- Three groups: No creatine; Once a day; and Twice a day.
- **Question**: does the average weight gain depend on the creatine group to which people were assigned?

Kruskal-Wallis

Example: creatine.csv

Actual values

No	Once	Twice
63	0	2239
-261	-652	171
-153	4724	40
-13	-2	1395
965	0	
	-86	

Ranks

No	Once	Twice
10	7.5	14
2	1	11
3	15	9
5	6	13
12	7.5	
	4	
32	**41**	**47**

$$H = \left[\frac{12}{n(n+1)} \sum_{j=1}^{c} \frac{T_j^2}{n_j}\right] - 3(n+1)$$

$$\mathbf{H} = \left[\frac{12}{15(15+1)}\left(\frac{32^2}{5} + \frac{41^2}{6} + \frac{47^2}{4}\right)\right] - 3(15+1) = \mathbf{3.868}$$

Where:

•n = sum of sample sizes for all samples,

•c = number of samples,

•T_j = sum of ranks in the j^{th} sample,

•n_j = size of the j^{th} sample.

Friedman test: Example

- An auction house is putting three violins, A, B, and C, up for bidding. Ten violinists are blindfolded are asked to rate the instruments and each player plays the violins in a randomly determined sequence (BCA, ACB, etc.).

- After each violin is played, the violinist rates the instrument on a 10-point scale of overall excellence (1=lowest, 10=highest).

- **Question**: which violin is the best according to the 10 violinists?

Friedman test

Example: violin.csv

Actual values

Violinists	Violin A	Violin B	Violin C
1	9	7	6
2	9.5	6.5	8
3	5	7	4
4	7.5	7.5	6
5	9.5	5	7
6	7.5	8	6.5
7	8	6	6
8	7	6.5	4
9	8.5	7	6.5
10	6	7	3

Ranks

Violinists	Violin A	Violin B	Violin C
1	3	2	1
2	3	1	2
3	2	3	1
4	2.5	2.5	1
5	3	1	2
6	2	3	1
7	3	1.5	1.5
8	3	2	1
9	3	2	1
10	2	3	1
Sum	77.5	67.5	57

$$\text{Q or T1 or FM} = \frac{n(k-1)\left[\sum_{i=1}^{k}\frac{R_i^2}{n} - C_F\right]}{\sum r_{ij}^2 - C_F}$$

$$C_F = \left(\frac{1}{4}\right)nk(k+1)^2$$

Where:

- n = sum of sample sizes for all samples,
- k = number of samples,
- R_j = sum of ranks in the j^{th} sample,
- r_{ij} = rank i of the j^{th} sample.

Kruskal-Wallis and Friedman tests

Exercise 15: creatine.csv

- **Question**: does the average weight gain depend on the creatine group to which people were assigned?

`kruskal_test(y~x)` produces omnibus part of the analysis

`dunn_test(y~x)` produces pairwise comparisons results

Exercise 16: violin.csv

- **Question**: which violin is the best according to the 10 violinists?

```
friedman_test(y~x|id)
wilcox_test(y~x, paired = TRUE, p.adjust.method = "bonferroni")
```

Have a go !

Exercise 15: creatine.csv

- Creatine, a supplement popular among body builders
- Three groups: No creatine; Once a day; and Twice a day.

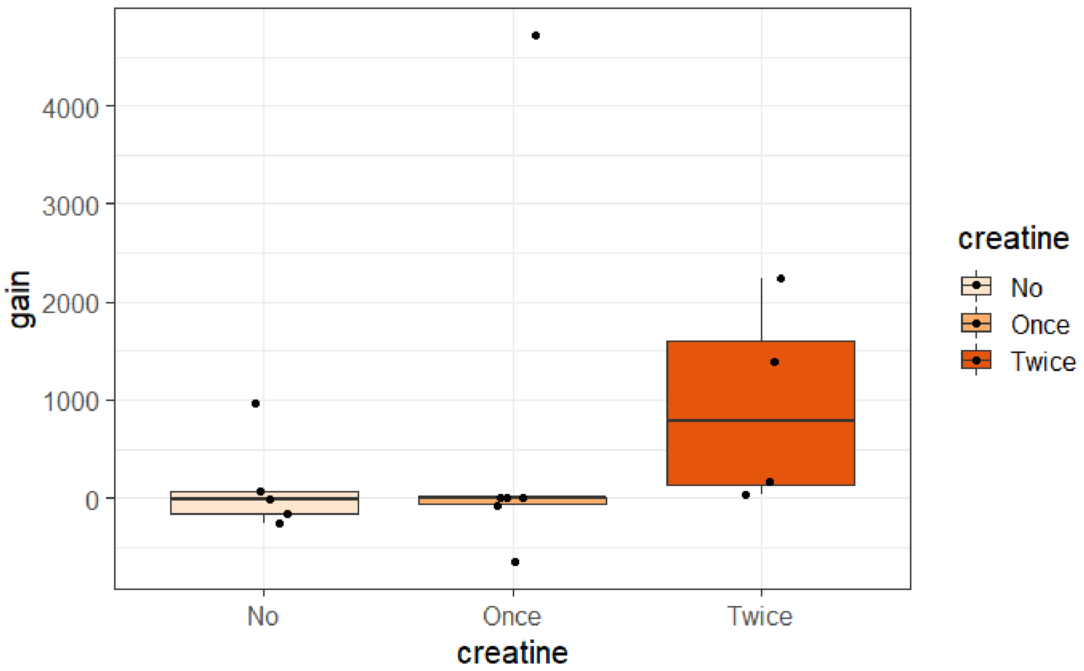
- **Question**: does the average weight gain depend on the creatine group to which people were assigned?

```
read_csv("creatine.csv") -> creatine

creatine %>%
ggplot(aes(creatine, gain, fill=creatine))+
  geom_boxplot(outlier.shape = NA)+
  geom_jitter(height = 0, width=0.1)+
  scale_fill_brewer(palette="Oranges")

creatine %>%
  group_by(creatine) %>%
    summarise(sd=sd(gain))
```

creatine <chr>	sd <dbl>
No	488.5317
Once	2005.1585
Twice	1047.8519



Exercise 15: creatine.csv

```
creatine %>%
kruskal_test(gain~creatine)
```

.y. <chr>	n <int>	statistic <dbl>	df <int>	p <dbl>	method <chr>
gain	15	3.86774	2	0.145	Kruskal-Wallis

```
creatine %>%
dunn_test(gain~creatine)
```

Not needed here

	.y. <chr>	group1 <chr>	group2 <chr>	n1 <int>	n2 <int>	statistic <dbl>	p <dbl>	p.adj <dbl>	p.adj.signif <chr>
1	gain	No	Once	5	6	0.160162	0.87275346	0.8727535	ns
2	gain	No	Twice	5	4	1.784928	0.07427304	0.2228191	ns
3	gain	Once	Twice	6	4	1.704706	0.08824926	0.2228191	ns

Answer: this study did not demonstrate any effect from creatine (χ^2 = 3.87, p = 0.14).

Exercise 16: violin.csv

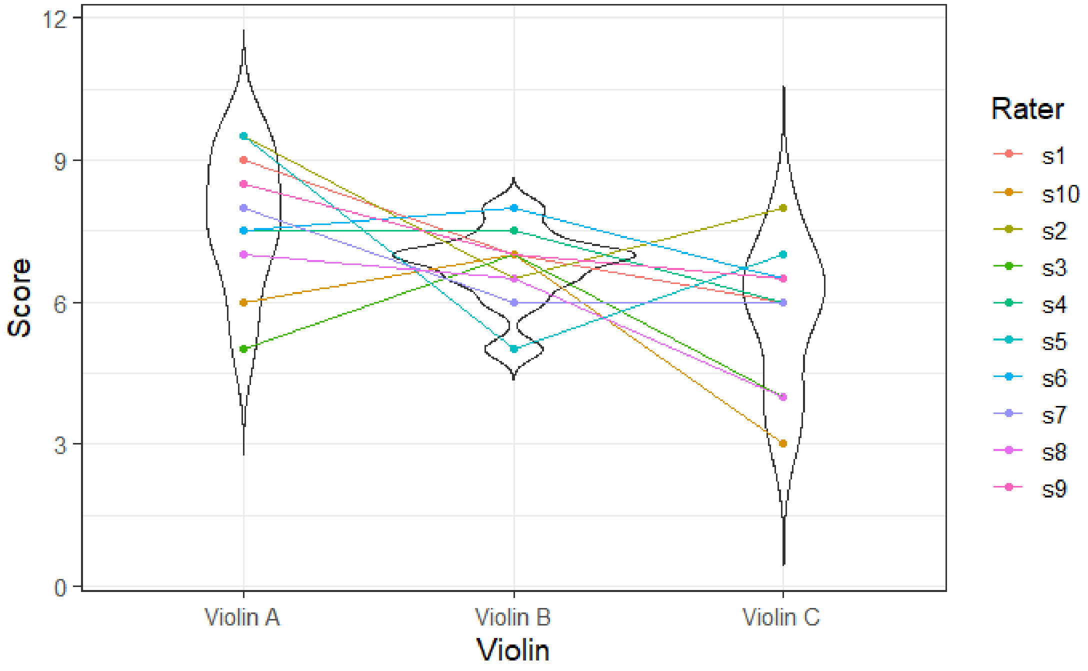
- 3 violins, each tested by 10 violinists.
- **Question**: which violin is the best according to the 10 violinists?

```
read_csv("violin.csv") -> violin

violin %>%
  ggplot(aes(Violin, Score)) +
  geom_violin(trim=FALSE)+
  geom_point(aes(colour=Rater)) +
  geom_line(aes(group = Rater, colour=Rater))
```

```
violin %>%
  group_by(Violin) %>%
    summarise(median = median(Score), sd=sd(Score))
```

Violin <chr>	median <dbl>	sd <dbl>
Violin A	7.75	1.4766704
Violin B	7.00	0.8249579
Violin C	6.00	1.5491933



Exercise 16: violin.csv

```
violin %>%
friedman_test(Score ~ Violin|Rater)
```

	.y. <chr>	n <int>	statistic <dbl>	df <dbl>	p <dbl>	method <chr>
1	Score	10	10.47368	2	0.005317021	Friedman test

```
violin %>%
  wilcox_test(Score ~ Violin, paired = TRUE, p.adjust.method = "bonferroni")
```

	.y. <chr>	group1 <chr>	group2 <chr>	n1 <int>	n2 <int>	statistic <dbl>	p <dbl>	p.adj <dbl>	p.adj.signif <chr>
1	Score	Violin A	Violin B	10	10	34.5	0.171	0.513	ns
2	Score	Violin A	Violin C	10	10	55.0	0.006	0.017	*
3	Score	Violin B	Violin C	10	10	35.0	0.154	0.462	ns

Answer: Violin A seems to be the best one.

Non-Parametric:
Spearman Correlation Coefficient

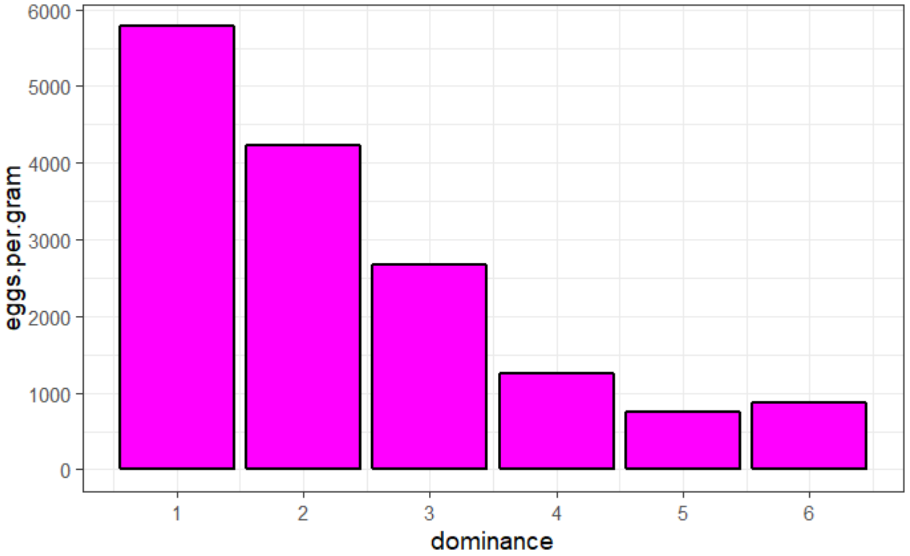
- Only really useful for ranks (either one or both variables)
- **ρ (rho)** is the equivalent of r and calculated in a similar way

```
read_csv("dominance.csv") -> dominance
```

- **Example**: **dominance.csv**
- Six male colobus monkeys ranked for dominance
- Question: is social dominance associated with parasitism?
  - Eggs of *Trichirus* nematode per gram of monkey faeces

monkey	rank	eggs.per.gram
Erroll	1	5777
Milo	2	4225
Fraiser	3	2674
Fergus	4	1249
Kabul	5	749
Hope	6	870

```
dominance %>%
ggplot(aes(rank, eggs.per.gram))+
  geom_col(fill="Magenta", colour="black", size=1)+
  scale_x_continuous(breaks=seq(1:6))+
  scale_y_continuous(breaks = seq(0, 6000, 1000))
```



Non-Parametric:
Spearman Correlation Coefficient

- **Example**: **dominance.csv**

```
dominance %>%
  cor_test(rank,eggs.per.gram, method = "spearman")
```

var1 <chr>	var2 <chr>	cor <dbl>	statistic <dbl>	p <dbl>	method <chr>
rank	eggs.per.gram	-0.94	68	0.0167	Spearman

- **Answer**: the relationship between dominance and parasitism is significant (ρ =-0.94, p=0.017) with high ranking males harbouring a heavier burden.