session-rank-tests

Olga Dethlefsen

Table of contents

Preface

Hypothesis tests that are based on knowledge of the probability distributions (e.g. normal or binomial) that the data follow are known as **parametric tests**. When data do not meet the parametric test assumptions, we can use **non-parametric** tests, also called distribution free tests, that replace the data with their ranks.

Learning outcomes

- know when to use non-parametric tests, their advantages and limitations
- name the main rank methods and their parametric counterparts
- explain how Wilcoxon signed rank test and Wilcoxon rank sum test work in detail
- be able to use R to compute Wilcoxon signed rank test, Wilcoxon rank sum test and Kruskal-Wallis one way analysis of variance

Do you see a mistake or a typo? We would be grateful if you let us know via edu.ml-biostats@nbis.se

This repository contains teaching and learning materials prepared and used during "Introduction to biostatistics and machine learning" course, organized by NBIS, National Bioinformatics Infrastructure Sweden. The course is open for PhD students, postdoctoral researcher and other employees within Swedish universities. The materials are geared towards life scientists wanting to be able to understand and use basic statistical and machine learning methods. More about the course https://nbisweden.github.io/workshop-mlbiostatistics/

1 Introduction

Hypothesis tests that are based on knowledge of the probability distributions (e.g. normal or binomial) that the data follow are known as **parametric tests**. When data do not meet the parametric test assumptions, we can use **non-parametric** tests, also called distribution free tests, that replace the data with their ranks. These tests came before computers enabled resampling to obtain the null distribution as seen before and are still being used in hypothesis testing.

1.1 Pros and cons

Pros

Non-parametric rank based test are useful when:

- we do not know the underlying probability distribution and/or our data does not meet parametric test requirements
- sample size is too small to properly assess the distribution of the data
- transforming our data to meet the parametric test requirements would make interpretation of the results harder

Cons

Some limitations of the non-parametric rank based tests include the facts that:

- they are primary significance tests that often do not provide estimates of the effects of interest
- they lead to waste of information and in consequence they have less power

- when sample size are extremely small (e.g. $n = 3$) rank tests cannot produce small P-values, even when the outcomes in the two groups are very different
- non-parametric tests are less easily extended to situations where we wish to take into account the effect of more than one exposure on the outcome

1.2 Main non-parametric rank tests

- **Wilcoxon signed rank test**
  - compares the sample median against a hypothetical median (equivalent to one sample t-test)
  - or examine the difference between paired observations (equivalent to paired t-test)
- **Wilcoxon rank sum test**
  - examines the difference between two unrelated groups
  - equivalent to two sample t-test
- **Kruskal-Wallis one-way analysis of variance**
  - examines the difference between two or more unrelated groups
  - equivalent to ANOVA
- **Spearman's rank correlation**
  - assess correlation on ranks
  - alternative to Pearson correlation coefficient
- **Kendall's rank correlation**
  - assess correlation on ranks
  - alternative to Pearson correlation coefficient

2 Wilcoxon signed rank test I

for a median

```
# load libraries
library(tidyverse)
library(magrittr)
library(kableExtra)
```

Named after Frank Wilcoxon (1892–1945), Wilcoxon signed rank test was one of the first "non-parametric" methods developed. It can be used to:

i) compare the sample median against a hypothetical median (equivalent to one sample t-test)
ii) examine the difference between paired observations (equivalent to paired t-test). It does make some assumptions about the data, namely:

- the random variable X is continuous
- the probability density function of X is symmetric

Example 2.1 (Wilcoxon signed rank test (for a median))**.** Let's imagine we are a part of team analyzing results of a placebo-controlled clinical trial to test the effectiveness of a sleeping drug. We have collected data on 10 patients when they took a sleeping drug and when they took a placebo.

The hours of sleep recorded for each study participant:

```
# input data
data.sleep <- data.frame(id = 1:10,
                         drug = c(6.1, 6.0, 8.2, 7.6, 6.5, 5.4, 6.9, 6.7, 7.4, 5.8),
                         placebo = c(5.2, 7.9, 3.9, 4.7, 5.3, 7.4, 4.2, 6.1, 3.8, 7.3))
```

```
print(data.sleep)
```

```
   id drug placebo
1   1  6.1     5.2
2   2  6.0     7.9
3   3  8.2     3.9
4   4  7.6     4.7
5   5  6.5     5.3
6   6  5.4     7.4
7   7  6.9     4.2
8   8  6.7     6.1
9   9  7.4     3.8
10 10  5.8     7.3
```

Before we investigate the effect of drug, a senior statistician ask us:

"Is the median sleeping time without taking the drug significantly less than the recommended 7 h of sleep?"

2.1 Define the null and alternative hypothesis under study

$H_0 : m = m_0$ the median sleeping time is equal to m_0, $m_0 = 7$ h

$H_1 < m_0$ the median sleeping time is less than m_0, $m_0 = 7$ h

2.2 Calculate the value of the test statistics

1. we subtract the median from each measurement, $X_i - m_0$
2. we find absolute value of the difference, $|X_i - m_0|$
3. we rank the absolute value of the difference

Table 2.1: Demonstrating steps in the calculating W, Wilcoxon signed-rank test statistics on the placebo column: x stands for placebo sleeping hours

id	x	x-m0	abs(x-m0)	R	Z	ZR
1	5.2	-1.8	1.8	6.0	0	0.0
2	7.9	0.9	0.9	3.5	1	3.5
3	3.9	-3.1	3.1	9.0	0	0.0
4	4.7	-2.3	2.3	7.0	0	0.0
5	5.3	-1.7	1.7	5.0	0	0.0
6	7.4	0.4	0.4	2.0	1	2.0
7	4.2	-2.8	2.8	8.0	0	0.0
8	6.1	-0.9	0.9	3.5	0	0.0
9	3.8	-3.2	3.2	10.0	0	0.0
10	7.3	0.3	0.3	1.0	1	1.0

4. we find the value of W, the Wilcoxon signed-rank test statistics as

$$W = \sum_{i=1}^{n} Z_i R_i \tag{2.1}$$

where Z_i is an indicator variable such as:

$$Z_i = \begin{cases} 0 & \text{if } X_i - m_0 < 0 \\ 1 & otherwise \end{cases} \tag{2.2}$$

```
m0 <- 7
data.wilcoxon <- data.sleep %>%
  select(!drug) %>% # remove drug data for now
  rename(x = placebo) %>% # rename placebo column to x so it is easier to type and follow eq
  mutate(`x-m0` = x - 7) %>% # subtract m0
  mutate(`abs(x-m0)` = abs(`x-m0`)) %>% # take absolute value
  mutate(R = rank(`abs(x-m0)`)) %>% # rank
  mutate(Z = ifelse(`x-m0` < 0, 0, 1)) %>% # define indicator variable Z
  mutate(ZR = R*Z) # calculate ranks R times Z

# print the table
data.wilcoxon %>%
  kable() %>% kable_styling(full_width = FALSE)
```

We can now calculate W following equation Equation 2.1 and we get:

```
# sum up the ranks multiplied by Z indicator value
W <- data.wilcoxon$ZR %>% sum()
print(W)
```

```
[1] 6.5
```

2.3 Compare the value to the test statistics to values from known probability distribution

We got $W = 6.5$ and now we need to calculate the P-value associated with W to be able to make decision about rejecting the null hypothesis. We refer to a statistical table "Upper and Lower Percentiles of the Wilcoxon Signed Rank Test, W" that can be found online or here.

For sample size $n = 10$ we can see that probability of observing $W \leq 3$ or $W \geq 52$ is small, 0.005. Probability of observing $W \leq 4$ or $W \geq 51$ is 0.007, still small but slightly larger. While we are getting towards the middle of the distribution the probability of observing W is getting larger and the probability of observing $W \leq 11$ or $W \geq 44$ is 0.053.

The P-value associated with observing $W = 6.5$ is just under 0.019. Assuming 5% significance level, we have enough evidence to reject the null hypothesis and conclude that the median is significantly less than 7 hours.

2.4 Obtaining probablity mass function

Where is the statistical table coming from?

Briefly, Wilcoxon described and showed examples how to calculate both the test statistics W for an example data as well as the distribution of W under the null hypothesis Wilcoxon

(1945). We can try to find the distribution of W ourselves for a simple scenario with less, four observation ($n = 4$)

```
# enumerate all rank possibilties (by hand)
r1 <- c(1, -1, 1, 1, 1, -1, -1, -1, 1, 1, 1, -1, -1, -1, 1, -1)
r2 <- c(2, 2, -2, 2, 2, -2, 2, 2, -2, 2, -2, -2, -2, 2, -2, -2)
r3 <- c(3, 3, 3, -3, 3, 3, -3, 3, -3, -3, 3, -3, 3, -3, -3, -3)
r4 <- c(4, 4, 4, 4, -4, 4, 4, -4, 4, -4, -4, 4, -4, -4, -4, -4)

data.w <- rbind(r1, r2, r3, r4)
data.w.ind <- data.w
data.w.ind[data.w < 0] <- 0
r.sum <- apply(data.w.ind, 2, sum)

data.w <- rbind(data.w, r.sum)
rownames(data.w) <- c("id1", "id4", "id3", "id4", "W")
colnames(data.w) <- paste("c", 1:16, sep="")

data.w %>% kable() %>% kable_styling(full_width = TRUE) %>%
  row_spec(5, bold = T, color = "black", background = "#deebf7")
```

```
Warning in latex_new_row_builder(target_row, table_info, bold, italic,
monospace, : Setting full_width = TRUE will turn the table into a tabu
environment where colors are not really easily configable with this package.
Please consider turn off full_width.
```

```
Warning in latex_new_row_builder(target_row, table_info, bold, italic,
monospace, : Setting full_width = TRUE will turn the table into a tabu
environment where colors are not really easily configable with this package.
Please consider turn off full_width.
```

- Given 4 observations, we could get ranks R_i of 1, 2, 3 or 4 only
- Further, depending where the observation would be with respect to m_0, the rank R_i could be positive or negative.
- For example, the first column $c1$ corresponds to all 4 observations having positive ranks, so all $x_i - m_0 > 0$, whereas column $c16$ corresponds to all observations having negative ranks, so $x_i - m_0 < 0$.

Table 2.2: **?(caption)**

(a)

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12	c13	c14	c15	c16
id1	1	-	1	1	1	-	-	-	1	1	1	-	-	-	1	-
		1				1	1	1				1	1	1		1
id4	2	2	-	2	2	-	2	2	-	2	-	-	-	2	-	-
			2			2			2		2	2	2		2	2
id3	3	3	3	-	3	3	-	3	-	-	3	-	3	-	-	-
				3			3		3	3		3		3	3	3
id4	4	4	4	4	-	4	4	-	4	-	-	4	-	-	-	-
					4			4		4	4		4	4	4	4
W	**10**	**9**	**8**	**7**	**6**	**7**	**6**	**5**	**5**	**3**	**4**	**4**	**3**	**2**	**1**	**0**

- As W test statistics is derived by summing up the positive ranks, we can see by listing all the combinations in the table, that $0 \leq W \leq 10$.

We can also now write down the probability mass function given the table.

```
# calculate probability mass function
W <- data.w[5,]

df.w <- data.frame(W = W) %>%
  group_by(W) %>%
  summarize(n = n()) %>%
  mutate(per = n / 16)

dist.W <- rbind(W = formatC(df.w$W), `p(W)`=df.w$per)

dist.W %>% t() %>%
  kable(digits = 2) %>%
  kable_styling(full_width = T) %>%
  row_spec(1, )
```

```
# plot pmf
barplot(df.w$per, names.arg = 0:10, ylab = "p(W)", xlab="W")
```

Table 2.3: **?(caption)**

(a)

W	p(W)
0	0.0625
1	0.0625
2	0.0625
3	0.125
4	0.125
5	0.125
6	0.125
7	0.125
8	0.0625
9	0.0625
10	0.0625

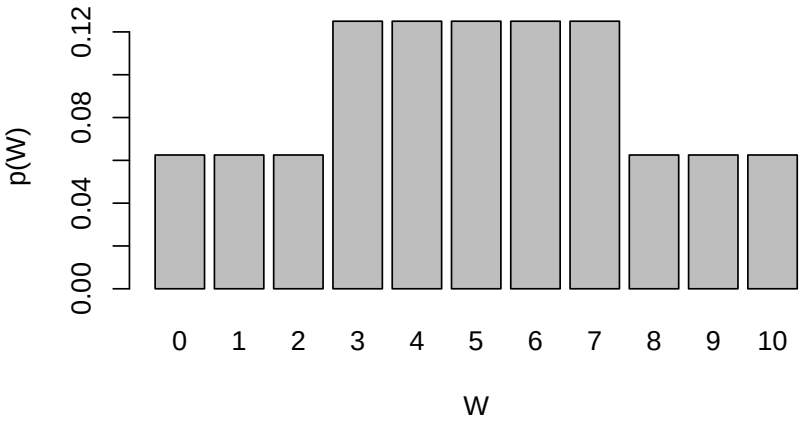


Figure 2.1: Probablity mass function for observations

Now we can use our knowledge from the Probability session on discrete distributions to calculate the probability associated with observing test statistics W given the known probability mass function.

For more examples on how to manually obtain the distribution W under the null hypothesis visit PennState Elbery collage website.

2.5 In R we use `wilcox.test()` function:

```r
# run Wilcoxon signed rank test for a median
wilcox.test(x = data.sleep$placebo,
            y = NULL,
            alternative = "less",
            mu = 7,
            exact = F
            )
```

```
    Wilcoxon signed rank test with continuity correction

data:  data.sleep$placebo
V = 6.5, p-value = 0.01827
alternative hypothesis: true location is less than 7
```

3 Wilcoxon signed rank test II

paired observations

```
# load libraries
library(tidyverse)
library(magrittr)
library(kableExtra)
```

Example 3.1 (Wilcoxon signed rank test (paired observations))**.** Going back to our sleep study, now we are ready to examine whether there is enough evidence to reject a null hypothesis of median of the differences between the paired observations is equal to 0. Being able to reject a null hypothesis would mean that the drug is having an effect.

The data we recorded are shown below:

```
# input data
data.sleep <- data.frame(id = 1:10,
                         drug = c(6.1, 6.0, 8.2, 7.6, 6.5, 5.4, 6.9, 6.7, 7.4, 5.8),
                         placebo = c(5.2, 7.9, 3.9, 4.7, 5.3, 7.4, 4.2, 6.1, 3.8, 7.3))

print(data.sleep)
```

```
   id drug placebo
1   1  6.1     5.2
2   2  6.0     7.9
3   3  8.2     3.9
4   4  7.6     4.7
5   5  6.5     5.3
6   6  5.4     7.4
```

```
7   7  6.9     4.2
8   8  6.7     6.1
9   9  7.4     3.8
10 10  5.8     7.3
```

3.1 Define the null and alternative hypothesis under the study

H_0 : the median difference in the population equals to zero

H_1 : the median difference in the population does not equals to zero

3.2 Test statistics: calculate difference and rank it

We start by calculating the difference between hours of sleep for each study participant. We exclude any difference that is equal to 0. The rest of values we rank in ascending order, ignoring the sign. As a result the smallest difference value, here 0.6 is ranked 1.

```
# calculate pair difference and rank it
df.wilcox.signed <- data.sleep %>%
  mutate(diff = drug - placebo) %>%
  mutate(rank = rank(abs(diff))) %>%
  print()
```

```
   id drug placebo diff rank
1   1  6.1     5.2  0.9    2
2   2  6.0     7.9 -1.9    5
3   3  8.2     3.9  4.3   10
4   4  7.6     4.7  2.9    8
5   5  6.5     5.3  1.2    3
6   6  5.4     7.4 -2.0    6
7   7  6.9     4.2  2.7    7
8   8  6.7     6.1  0.6    1
```

```
9   9  7.4      3.8  3.6    9
10 10  5.8      7.3 -1.5    4
```

3.3 Test statistics: sum up the ranks of the negative differences and of positive differences and denote these sums by T_- and T_+ respectively

We get $T_- = 40$ and $T_+ = 15$

Why? If there were no differences in effectiveness between the sleeping drug and the placebo then the sums T_- and T_+ would be similar. If there were a difference then one sum would be much smaller and one sum would be much larger than expected.

```r
# sum up the ranks of the positive and negative differences
data.tsum <- df.wilcox.signed %>%
  mutate(sign = ifelse(diff < 0, "+", "-")) %>%
  group_by(sign) %>%
  summarize(T = sum(rank)) %>%
  print()
```

```
# A tibble: 2 x 2
  sign      T
  <chr> <dbl>
1 -        40
2 +        15
```

3.4 Test statistics: denote the smaller sum by T and interpret the *P*-value

The Wilcoxon signed rank test is based on assessing whether T, the smaller of T_- and T_+, is smaller than would be expected by chance, under the null hypothesis that the median of the paired differences is zero.

The hypothesis is that T is equal to the sum of the ranks divided by 2, so that the smaller T the more evidence there is against the null hypothesis.

Having our T value we can check what is the probability of observing the value of T under the null hypothesis, by checking the statistical table of "Critical values for the Wilcoxon matched pairs signed rank test" found online or here.

In our example, $T = 15$ and the sample size is $n = 10$, where n is the number of non-zero differences (we had none). According to the table, the 5% percentage point is at 8. Since $T = 15 > 8$ our $P - value > 0.05$ and we do not have enough evidence to reject the null hypothesis. There is no evidence of the sleeping drug working.

3.5 In R we use `wilcox.test()` function adjusting `paired` argument:

```
# run Wilcoxon signed rank test for paired observations
wilcox.test(x = data.sleep$placebo,
            y = data.sleep$drug,
            alternative = "two.sided",
            mu = 0,
            paired = TRUE,
            exact = F)
```

```
    Wilcoxon signed rank test with continuity correction

data:  data.sleep$placebo and data.sleep$drug
V = 15, p-value = 0.2213
alternative hypothesis: true location shift is not equal to 0
```

4 Wilcoxon rank sum test

two unrelated groups

Wilcoxon rank sum test is used to assess whether an outcome variable differs between two exposure groups, so it equivalent to the non-parametric two sample t test. It examines whether **the median difference between two groups is equal to zero**. Let's follow an example to get a better idea how it works.

Example 4.1 (Wilcoxon rank sum test)**.** We have weighted new born babies born to 5 non-smokers and 6 smokers. The measurements, with weight in kg, are shown below. Let's see if there is enough evidence to reject the null hypothesis of the median difference between the groups being equal to zero.

The data are shown below:

```
# input data
bw.nonsmokers <- c(3.99, 3.89, 3.6, 3.73, 3.31)
bw.smokers <- c(3.18, 2.74, 2.9, 3.27, 3.15, 2.42)

# group labels
grp.nonsmokers <- rep("No", 1, length(bw.nonsmokers))
grp.smokers <- rep("Yes", 1, length(bw.smokers))

# no. of observations per group
n.nonsmokers <- length(bw.nonsmokers)
n.smokers <- length(bw.smokers)

# put data into one data frame
data.babies <- data.frame(id = 1:(n.nonsmokers + n.smokers),
                          weight = c(bw.nonsmokers, bw.smokers),
                          smoking = c(grp.nonsmokers, grp.smokers))
```

```
# print data
data.babies %>%
  print()
```

```
   id weight smoking
1   1   3.99      No
2   2   3.89      No
3   3   3.60      No
4   4   3.73      No
5   5   3.31      No
6   6   3.18     Yes
7   7   2.74     Yes
8   8   2.90     Yes
9   9   3.27     Yes
10 10   3.15     Yes
11 11   2.42     Yes
```

4.1 Define the null and alternative hypothesis under study

H_0 : the difference between the medians of the two groups equals to zero

H_1 : the difference between the medians of the two groups does not equals to zero

4.2 Test statistics: rank the values

We rank the values of the weights from both groups together in *ascending* order of magnitude. If any of the values are equal, we average their ranks.

```
# rank weight variable in ascending order
df.wilcoxon.rank.sum <- data.babies %>%
  mutate(rank = rank(weight)) %>%
  print()
```

```
   id weight smoking rank
1   1   3.99      No   11
2   2   3.89      No   10
3   3   3.60      No    8
4   4   3.73      No    9
5   5   3.31      No    7
6   6   3.18     Yes    5
7   7   2.74     Yes    2
8   8   2.90     Yes    3
9   9   3.27     Yes    6
10 10   3.15     Yes    4
11 11   2.42     Yes    1
```

4.3 Test statistics: sum up the ranks in the smaller group

We add up the ranks in the group with the smaller sample size. If both groups have equal number of measurements just pick one group. Here, the smaller group are the no smokers, and the rank sum up to $T = 45$

```r
# sum up ranks for the smaller group
data.sumrank <- df.wilcoxon.rank.sum %>%
  group_by(smoking) %>%
  summarize(T = sum(rank)) %>%
  filter(smoking == "No") %>%
  pull(T) %>%
  print()
```

```
[1] 45
```

4.4 Test statistics: find & interpret the *P*-value

We compare the T value with the values in "Critical range for the Wilcoxon rank sum test" found online or here. The range

shown for $P = 0.05$ is from 18 to 42 for sample size 5 and 6 respectively. T value below 18 or above 42 corresponds to $P - value < 0.05$. In our case $T = 45$ so above 42, hence we have enough evidence to reject the null hypothesis that the median birth weight of children born to smokers is the same as the median birth weight of children born to non-smokers.

4.5 In R

In R we compute the test with `kruskal.test()` function changing `paired` parameter to False.

```
# compute Wilcoxon rank sum test
wilcox.test(data.babies$weight ~ data.babies$smoking,
            exact = T,
            paired = F)
```

```
    Wilcoxon rank sum exact test

data:  data.babies$weight by data.babies$smoking
W = 30, p-value = 0.004329
alternative hypothesis: true location shift is not equal to 0
```

4.6 Note on confidence intervals

To get the confidence intervals we could set `conf.int = T`:

```
# compute Wilcoxon rank sum test incl. CI
wilcox.test(data.babies$weight ~ data.babies$smoking,
            exact = F,
            paired = F,
            conf.int = T)
```

```
    Wilcoxon rank sum test with continuity correction
```

```
data:  data.babies$weight by data.babies$smoking
W = 30, p-value = 0.008113
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 0.3300051 1.2499709
sample estimates:
difference in location
             0.7224397
```

or we could obtain CI via bootstrapping as we have seen earlier today:

```r
# calculate bootstrapping CI
n <- 1000 # number of bootstrapped samples
v.mdiff <- c() # vector to hold difference in means for each iteration

for (i in 1:n){
  s.nonsmokers <- sample(bw.nonsmokers, replace = T) # sampling from nonsmokers
  s.smokers <- sample(bw.smokers, replace = T) # sampling from smokers

  m.nonsmokers <- median(s.nonsmokers) # calculate median of nonsmokers
  m.smokers <- median(s.smokers) # calculate median of nonsmokers

  v.mdiff[i] <- m.nonsmokers - m.smokers # difference in median
}

# use percentiles to calculate 95% CI, top and bottom 2.5%
CI.95 <- quantile(v.mdiff, probs = c(0.025, 0.975))
print(CI.95)
```

```
2.5% 97.5%
0.16  1.17
```

5 The Kruskal-Wallis test

two or more unrelated groups

Kruskal-Wallis test is an extension of the Wilcoxon rank sum test for unrelated k groups, where $k \geq 2$. Under the null hypothesis of no differences in the distribution between the groups, the sums of the ranks in each of the k groups should be comparable after allowing for any differences in sample size.

In R one can use `kruskal.test()` to compute the test. Otherwise, the procedure is outlined below.

5.1 Define the null and alternative hypothesis

H_0 : each group has the same distribution of values in the population

H_1 : at least one group does not have the same distribution of values in the population

5.2 Calculate the value of the test statistics

Rank all n values and calculate the sum of the ranks in each of the groups: these sums are $R_1, R_2, ..., R_k$. The test statistics is given by:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{n} \frac{R_i^2}{n_i} - 3(n+1)$$

which follows a χ^2 distribution with $(k-1)$ degrees of freedom.

6 Correlation

6.1 Pearson correlation coefficient

Pearson correlation coefficient, or rather more correctly Pearson product moment correlation coefficient, gives us an idea about the strength of association between two numerical variables. Its true value in the population, ρ, is estimated in the sample by r, where:

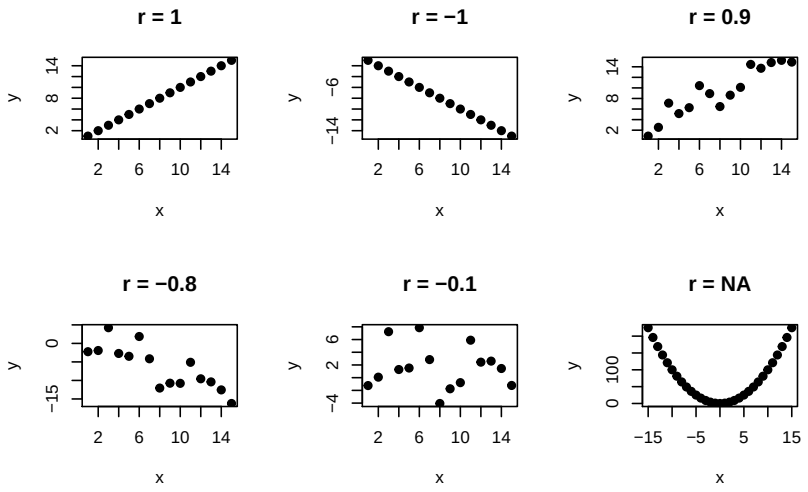
$$r = \frac{\sum(x - \bar{x})(x - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(x - \bar{y})^2}} \tag{6.1}$$

Properties:

- the value of r range between -1 to 1
- the sign indicates whether, in general, one variable increases as the other variable increases ($r > 0$) or whether one variable increases while the other variables decreases ($r < 0$)
- the magnitude indicates how close the points are to the straight line, in particular $r = 1$ for a perfect positive correlation, $r = -1$ for a perfect negative correlation and $r = 0$ for no correlation

```
# simulate data with different levels of correlation
# no. of observations to generate
n <- 15

# perfect positive correlation
x1 <- 1:n
y1 <- 1:n
cor1 <- cor(x1, y1) %>% round(2)
```

```r
# perfect negative correlation
x2 <- 1:n
y2 <- -1*(1:n)
cor2 <- cor(x2, y2) %>% round(2)

# positive correlation
set.seed(123)
x3 <- 1:n
y3 <- x3 + rnorm(n, mean = 1, sd = 2)
cor3 <- cor(x3, y3) %>% round(1)

# negative correlation
set.seed(123)
x4 <- 1:n
y4 <- x4*(-1) + rnorm(n, mean = 1, sd = 4)
cor4 <- cor(x4, y4) %>% round(1)

# no correlation
set.seed(123)
x5 <- 1:n
y5 <- rnorm(n, mean = 1, sd = 4)
cor5 <- cor(x5, y5) %>% round(1)

# quadratic relationship
x6 <- -n:n
y6 <- x6^2

par(mfrow = c(2,3))
plot(x1, y1, xlab="x", ylab="y", main = paste("r = ", cor1, sep=""), pch=19)
plot(x2, y2, xlab="x", ylab="y", main = paste("r = ", cor2, sep=""), pch=19)

plot(x3, y3, xlab="x", ylab="y", main = paste("r = ", cor3, sep=""), pch=19)
plot(x4, y4, xlab="x", ylab="y", main = paste("r = ", cor4, sep=""), pch=19)

plot(x5, y5, xlab="x", ylab="y", main = paste("r = ", cor5, sep=""), pch=19)
plot(x6, y6, xlab="x", ylab="y", main = paste("r = NA", sep=""), pch=19)
```



Limitations

It may be misleading to calculate correlation coefficient, r, when:

- there is a non-linear relationship between the two variables, e.g. quadratic
- outliers are present
- the data include more than one observation on each individual (grouped data)

Spearman correlation and Kendall's tau use try to overcome some of the above limitations, by operating on ranks, to measure the strength of the association.

6.2 Spearman correlation

To calculate Spearman's rank correlation between two variables X and Y we:

- rank the values of X and Y independently
- follow the formula to calculate the Pearson correlation coefficient using ranks (Equation 6.1)

6.3 Kendall's tau

To calculate Kendall's tau, τ, we compare ranks of X and Y between every pair of observation. (There are n(n-1)/2 possible pairs). The pairs of ranks for observation i and j are said to be:

- concordant: if they differ in the same direction, i.e. if both the X and Y ranks of subject i are lower than the corresponding ranks of subject j, or both are higher
- discordant: otherwise

$$\tau = \frac{n_C - n_D}{n(n-1)/2}$$

where

n_C, number of concordant pairs n_D, number of discordant pairs

For instance, in the below data, the ranks of subjects #1 and #2 are concordant as subject #1 has a lower rank than subject #2 for both the variables. The ranks of subjects #3 and #6 are discordant as subject #3 has a more highly ranked X value than subject #6 but a lower ranked $Y value$.

```
# input data
x <- c(58, 70, 74, 63.5, 62.0, 70.5, 71, 66)
y <- c(2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12)
n <- length(x)
df <- data.frame(subject = 1:n, x = x, rank_x = rank(x), y = y, rank_y = rank(y))

df %>%
  kable() %>%
  kable_styling(full_width = TRUE)
```

6.4 In `R` we use `cor()` function

Apart from following the above equations by hand, in R we can use `cor()` function to calculate Pearson, Spearman and

Table 6.1: **?(caption)**

(a)

subject	x	rank_x	y	rank_y
1	58.0	1	2.75	2
2	70.0	5	2.86	4
3	74.0	8	3.37	7
4	63.5	3	2.76	3
5	62.0	2	2.62	1
6	70.5	6	3.49	8
7	71.0	7	3.05	5
8	66.0	4	3.12	6

Kendall's tau correlation coefficients.

```
# Pearson
cor(x,y, method = "pearson")
```

```
[1] 0.7591266
```

```
# Spearman
cor(x,y, method = "spearman")
```

```
[1] 0.8095238
```

```
# Kendall's tau
cor(x,y, method = "kendall")
```

```
[1] 0.6428571
```

Exercises

Exercise 6.1 (Wilcoxon signed rank test)**.** Try repeating the calculations for the Wilcoxon signed rank test. Could you check if the median sleeping time without taking the drug is significantly different than 6 h?

Below is the code to create the data.

```
# input sleep data
data.sleep <- data.frame(id = 1:10,
                         drug = c(6.1, 6.0, 8.2, 7.6, 6.5, 5.4, 6.9, 6.7, 7.4, 5.8),
                         placebo = c(5.2, 7.9, 3.9, 4.7, 5.3, 7.4, 4.2, 6.1, 3.8, 7.3))
```

Exercise 6.2 (Wilcoxon rank sum test)**.** You've collected more data on the newborn babies born to smokers and non-smokers. Is the enough evidence to reject the null hypothesis of the difference between the medians of the two groups equals to zero? Use the code below to input new data.

```
# input babies weights
bw.nonsmokers <- c(3.99, 3.89, 3.6, 3.73, 3.31, 3.7, 4.08, 3.61, 3.83, 3.41, 4.13, 3.36, 3.5
bw.smokers <- c(3.18, 2.74, 2.9, 3.27, 3.65, 3.42, 3.23, 2.86, 3.6, 3.65, 3.69, 3.53, 2.38,
```

Exercise 6.3 (Kruskall Wallis)**.** Can you double-check your calculations using Kruskall-Wallis test instead via `kruskal-test()` function? Would you expect to get different or similar results? And finally, imagine that you've repeated

the experiment again, this time collecting data for three groups, non-smokers, occasional smokers and regular smokers. Is there enough evidence to reject the null hypothesis of each group having the same distribution of values in the population?

Data data are below.

```
# input babies weights
bw.nonsmokers <- c(3.99, 3.89, 3.6, 3.73, 3.31, 3.7, 4.08, 3.61, 3.83, 3.41, 4.13, 3.36, 3.5
bw.smokers <- c(3.18, 2.74, 2.9, 3.27, 3.65, 3.42, 3.23, 2.86, 3.6, 3.65, 3.69, 3.53, 2.38,
bw.occsmokers <- c(3.65, 3.53, 2.34, 3.70, 3.42, 2.71, 3.83, 3.60, 3.18, 3.65)
```

References

Wilcoxon, Frank. 1945. "Individual Comparisions by Ranking Methods." *Biometrics Bulletin* 1 (6): 80–83.