A tutorial on comparing two independent samples in R

Version 1.0.0

Konstantinos I. Bougioukas

04/11/2021

Contents

Objectives

- Applying hypothesis testing
- Compare two independent samples
- Interpret the results

Packages for these notes:

We need to download the `rstatix`, `PupillometryR` and `EnvStats` packages for the notes.

Next, we need to load the following packages:

```
library(rstatix)      # Pipe-Friendly Framework for Basic Statistical Tests
library(PupillometryR)
library(here)         # A Simpler Way to Find Your Files
library(tidyverse)    # Easily Load the 'Tidyverse'
```

1 Introduction

1.1 The variables

In these notes, we discuss situations where we investigate possible associations between a binary random variable and a numerical random variable. In these situations, the binary variable typically represents two different experimental conditions (e.g., treatment vs. placebo) or two different groups (diabetes vs. non-diabetic patients) from the population. We will treat the binary variable as the independent variable in our analysis. The binary variable is also known as the **grouping** variable or **factor** variable. The numerical variable, on the other hand, is regarded as the dependent (response) variable (e.g., systolic blood pressure, urinary β thromboglobulin excretion, age)

1.2 The samples

Two samples are independent if the sample values selected from one population are not related to or somehow paired or matched with the sample values selected from the other population. For example, one group of subjects is treated with a drug, while a second and separate group of subjects is given a placebo. These two sample groups are independent because the individuals in the treatment group are in no way paired or matched with corresponding members in the placebo group.

Note Throughout these notes, we assume that the individuals from which we collect data are sampled randomly. Moreover, we hope that the individuals in the two samples are quite comparable except for the characteristic that defines the groups.

1.3 Parametric and non-parametric statistical tests

There are two types of significance test and they are described as parametric and non-parametric. For the parametric tests (e.g., t-test, ANOVA), certain conditions should exist for the population being tested. For non-parametric tests (e.g., Wilcoxon tests), no such conditions are laid down. For example, one condition which must hold before a (parametric) t-test can be used is that the population from which the sample under observation is drawn must be normally distributed. That is, the population distribution can be fitted by a normal curve.

Note Permutation tests can also be used to compare distributions. The permutation test can be used with any measure of location (e.g., mean, median), but regardless of which measure of location is used, in essence the goal is to test the hypothesis that the groups under study have identical distributions. There are many extensions and variations of the method including a range of techniques.

2 Two-sample t-test (Student's t-test)

Two sample t-test (Student's t-test) can be used if we have two independent (unrelated) groups (e.g., males-females, unmatched case-controls, treatment-non treatment) and one quantitative variable of interest.

2.1 Research question

In an experiment designed to test the effectiveness of paroxetine for treating bipolar depression, the participants were randomly assigned into two groups (intervention Vs placebo). The researchers used the Hamilton Depression Rating Scale (HDRS) to measure the dpression state of the participants and wanted to find out if the HDRS score is different in paroxetine group as compared to placebo group at the end of the experiment. The significance level α was set to 0.05.

Note A score of 0–7 in HDRS is generally accepted to be within the normal range, while a score of 20 or higher indicates at least moderate severity.

2.2 H0 and H1 Hypotheses

- H_0: the means of HDRS in the two groups are equal ($\mu_1 = \mu_2$)
- H_1: the means of HDRS in the two groups are different ($\mu_1 \neq \mu_2$)

2.3 Preraring the data

We import the data:

```
library(readxl)
depression <- read_excel(here("data", "depression.xlsx"), col_names=TRUE)
depression
```

Table 1: Depression Data (first and last 5 rows)

group	HDRS
placebo	19
placebo	21
placebo	28
placebo	22
placebo	22
NA	...
paroxetine	16
paroxetine	25
paroxetine	22
paroxetine	19
paroxetine	24

We inspect the data:

```
glimpse(depression)
```

```
## Rows: 76
## Columns: 2
## $ group <chr> "placebo", "placebo", "placebo", "placebo", "placebo", "pla~
## $ HDRS  <dbl> 19, 21, 28, 22, 22, 28, 23, 17, 19, 20, 26, 23, 23, 22, 19,~
```

The dataset `depression` has 76 patients and includes two variables. The numeric `HDRS` variable and the `group` variable which should be converted from character to a factor variable using the `factor()`:

```
depression <- depression %>%
  mutate(group = factor(group))
glimpse(depression)
```

```
## Rows: 76
## Columns: 2
## $ group <fct> placebo, placebo, placebo, placebo, placebo, placebo, place~
## $ HDRS  <dbl> 19, 21, 28, 22, 22, 28, 23, 17, 19, 20, 26, 23, 23, 22, 19,~
```

2.4 Assumptions

1. The data are **normally** distributed in both groups
2. The data in both groups have similar **variance** (also named as homogeneity of variance or homoscedasticity)

2.5 Explore the characteristics of distributions

The distributions can be explored visually with appropriate plots. Additionally, summary statistics and significance tests to check for normality (e.g., Shapiro-Wilk test) and for equality of variances (e.g., Levene's test) can be used.

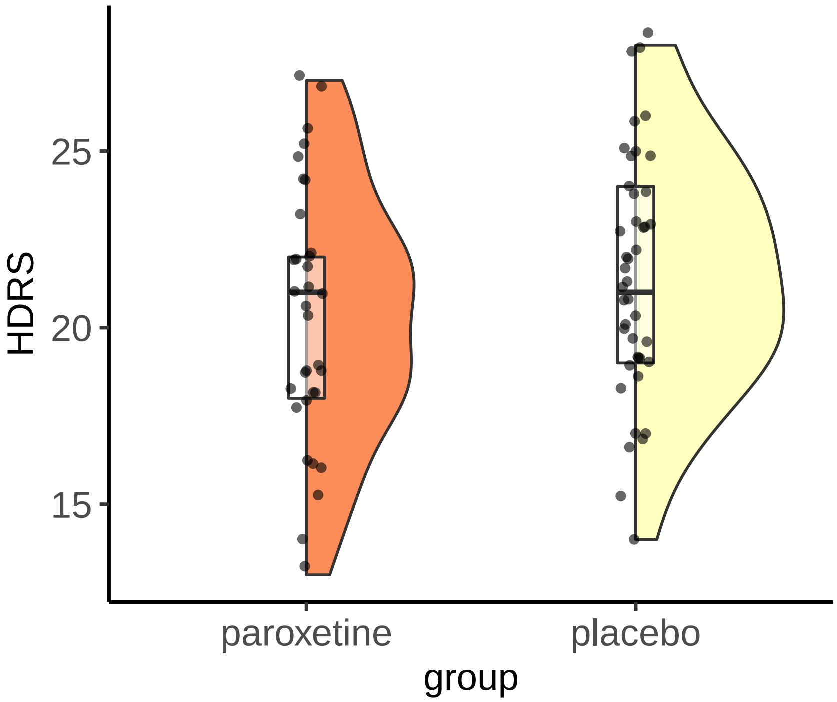
Note Normality tests are **not** very helpful guides as they are often non-significant for small samples even when the variables appear to be non-normal (this is because the test is underpowered in small samples), and they tend to be significant for large samples even for small deviation from perfect normality.

Note The Central Limit Theorem (CLT) is frequently used to justify the use of parametric statistical tests and confidence intervals when their assumptions seem to be violated based on the samples. For example, it is claimed that with a sample size of 30 or more, when testing Ho: $\mu_1 = \mu_2$, normality can be assumed for the sampling distributions. However, simulations have shown that larger sample sizes often are needed (>200 for light-tailed population distributions and >300 for skewed distributions having heavier tails).

2.5.1 Visualization of the distributions

We can visualize the distribution of HDRS for the two groups:

```
ggplot(depression, aes(x=group, y=HDRS)) +
  geom_flat_violin(aes(fill = group), scale = "count") +
  geom_boxplot(width = 0.11, outlier.shape = NA, alpha = 0.5) +
  geom_point(position = position_jitter(width = 0.05),
             size = 1.2, alpha = 0.6) +
  scale_fill_brewer(palette = "Spectral") +
  theme_classic(base_size = 14) +
  theme(legend.position="none",
        axis.text = element_text(size = 14))
```



The above figure shows that the data are close to symmetry and the assumption of a normal distribution is reasonable.

2.5.2 Summary statistics

The HDRS summary statistics for each group are:

```
HDRS_summary <- depression %>%
  group_by(group) %>%
  dplyr::summarise(
    n = n(),
    min = min(HDRS, na.rm = TRUE),
    q1 = quantile(HDRS, 0.25, na.rm = TRUE),
    median = quantile(HDRS, 0.5, na.rm = TRUE),
    q3 = quantile(HDRS, 0.75, na.rm = TRUE),
    max = max(HDRS, na.rm = TRUE),
    mean = mean(HDRS, na.rm = TRUE),
    sd = sd(HDRS, na.rm = TRUE),
    skewness = EnvStats::skewness(HDRS, na.rm = TRUE),
    kurtosis= EnvStats::kurtosis(HDRS, na.rm = TRUE)
  ) %>%
  ungroup()

HDRS_summary
```

group	n	min	q1	median	q3	max
paroxetine	33	13	18	21	22	27
placebo	43	14	19	21	24	28

group	mean	sd	skewness	kurtosis
paroxetine	20.33333	3.654335	0.0016663	-0.5738236
placebo	21.48837	3.411271	0.0276307	-0.4033681

The means are close to medians (20.3 vs 21 and 21.5 vs 21) and the two standard deviations (3.65 vs 3.41) are also similar. The skewness is approximately zero (symmetric distribution) and the (excess) kurtosis is close to zero (mesokurtic distribution) indicating normal distributions for both groups.

Alternatively, we can download the `{dlookr}` package and use the `describe()` function:

```
depression %>%
  group_by(group) %>%
  dlookr::describe(HDRS) %>%
  select(variable,  group, n, mean, sd, p25, p50, p75, skewness, kurtosis) %>%
  ungroup()
```

```
## # A tibble: 2 x 10
##   variable group          n  mean    sd   p25   p50   p75 skewness kurtosis
##   <chr>    <fct>      <int> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>    <dbl>
## 1 HDRS     paroxetine    33  20.3  3.65    18    21    22  0.00167   -0.574
## 2 HDRS     placebo       43  21.5  3.41    19    21    24  0.0276    -0.403
```

2.5.3 Shapiro-Wilk test for normality

The `Shapiro-Wilk` test for normality for each group is:

```
depression %>%
  group_by(group) %>%
  shapiro_test(HDRS) %>%
  ungroup()
```

group	variable	statistic	p
paroxetine	HDRS	0.976	0.670
placebo	HDRS	0.979	0.614

The tests of normality suggest that the data for the `HDRS` in both groups are normally distributed (p=0.67 >0.05 and p=0.61 >0.05, respectively).

2.5.4 Levene’s test for equality of variances

The `Levene's test` for equality of variances is:

```
depression %>%
  levene_test(HDRS ~ group)
```

df1	df2	statistic	p
1	74	0.176	0.676

Since the p-value = 0.676, which is higher than 0.05, the null hypothesis that the variances of HDRs in two groups are equal is not rejected.

2.6 Run the t-test

We will perform a pooled variance t-test (Student’s t-test) to test the null hypothesis that the mean HDRS score is the same for both groups (paroxetine and placebo).

```
depression %>%
    t_test(HDRS ~ group, var.equal = T, detailed = T)
```

estimate	estimate1	estimate2	.y.	group1	group2	n1	n2
-1.155039	20.33333	21.48837	HDRS	paroxetine	placebo	33	43

statistic	p	df	conf.low	conf.high	method
-1.418505	0.16	74	-2.777497	0.46742	T-test

The p-value = 0.16 is higher than 0.05. There is no evidence of a significant difference in mean HDRS between the two groups (failed to reject Ho). The difference between means (20.33 - 21.49) equals to -1.16 units of the HDRS and note that the 95% confidence interval of the difference in means (-2.78 to 0.47) includes the hypothesized null value of 0. Based on these results, there is not evidence that paroxetine is effective as a treatment for bipolar depression.

Note that the paroxetine sample (n= 33) has 32 (33-1) degrees of freedom and the placebo sample (n= 43) has 42 (43-1), so we have 74 (32 + 42) d.f. in total. Another way of thinking of this is to reason that the complete sample size as 76, and we have estimated two parameters from the data (the two means) so we have 76-2 = 74 d.f.

Note The Student t-test for two independent samples does not have any restrictions on n1 and n2 —they can be equal or unequal. However, equal samples are preferred because when a total of 2n subjects are available, their equal division among the groups maximizes the power to detect a specified difference. Although equal samples are desirable, this may not be a prudent allocation in many medical situations. In clinical trials, many times controls are easy to investigate, and more than one control per case could be a good strategy.

Note If the variance is different between the two groups then the degrees of freedom and the t-value associated with a two-sample t-test are calculated differently. In this case, we have to write **var.equal = F** (or write nothing because this is the default) in the function so the Welch-Satterthwaite approximation is applied to the degrees of freedom.

2.7 Present the results in a summary table

Characteristic	**Overall**, N = 76[1]	**paroxetine**, N = 33[1]	**placebo**, N = 43[1]	**p-value**[2]
HDRS score	21.0 (3.5)	20.3 (3.7)	21.5 (3.4)	0.16

[1] Mean (SD)

[2] Two Sample t-test

Hence, there is not evidence that HDRS score is significantly different in paroxetine group, mean (sd) 20.3 (3.7), as compared to placebo group, 21.5 (3.4), (mean difference= -1.16 units, 95% CI = -2.78 to 0.47, p = 0.16 >0.05).

3 Wilcoxon-Mann-Whitney (Mann-Whitney U) test

The Wilcoxon-Mann-Whitney (WMW) test (sometimes called Mann-Whitney U test or Wilcoxon Rank Sum test) is used to compare two independent samples and is often considered the non-parametric alternative to the two-sample t-test when there is violation of normality or for small sample sizes. However, if the samples are very small (both smaller than four observations) then statistical significance is impossible.

Note The non-parametric tests are based on the **ranks** of the data rather than on the actual data values so, in general, they are about 5% less powerful than parametric tests when normality assumption is upheld. However, the non-parametric tests can be more powerful than parametric tests if the distributions are strongly skewed by the presence of outliers.

3.1 Research question

We consider the data in `thromboglobulin` dataset that contains the urinary β thromboglobulin excretion (pg/ml) measured in 12 non-diabetic patients and 12 diabetic patients. The researchers used α = 0.05 significance level to test the null hypothesis that the distribution of urinary β thromboglobulin (b_tg) is the same in the two groups.

3.2 H0 and H1 Hypotheses

- H_0: the distribution of urinary β thromboglobulin is the same in the two groups
- H_1: the distribution of urinary β thromboglobulin is different in the two groups

Note The null hypothesis is that the observations from one group do not tend to have a higher or lower ranking than observations from the other group. This test **does not** test the medians of the data as is commonly thought, it tests the **whole distribution**. In practice, however, we use the medians to present the results. Statistical speaking, if the distributions of the two groups have **similar shapes**, the Wilcoxon-Mann-Whitney test can be used to determine whether there are differences in the medians in the two groups.

3.3 Preraring the data

We import the data:

```r
library(readxl)
thromboglobulin <- read_excel(here("data", "thromboglobulin.xlsx"))
thromboglobulin
```

We inspect the data:

Thromboglobin Data (first and last 5 rows)

status	b_tg
non-diabetic	4.1
non-diabetic	6.3
non-diabetic	7.8
non-diabetic	8.5
non-diabetic	8.9
NA	...
diabetic	40.7
diabetic	51.3
diabetic	56.2
diabetic	61.7
diabetic	69.2

```
glimpse(thromboglobulin)
```

```
## Rows: 24
## Columns: 2
## $ status <chr> "non-diabetic", "non-diabetic", "non-diabetic", "non-diabe~
## $ b_tg   <dbl> 4.1, 6.3, 7.8, 8.5, 8.9, 10.4, 11.5, 12.0, 13.8, 17.6, 24.~
```

The variable `status` can be converted to a factor variable:

```
thromboglobulin <- thromboglobulin %>%
  mutate(status = factor(status))

glimpse(thromboglobulin)
```

```
## Rows: 24
## Columns: 2
## $ status <fct> non-diabetic, non-diabetic, non-diabetic, non-diabetic, no~
## $ b_tg   <dbl> 4.1, 6.3, 7.8, 8.5, 8.9, 10.4, 11.5, 12.0, 13.8, 17.6, 24.~
```

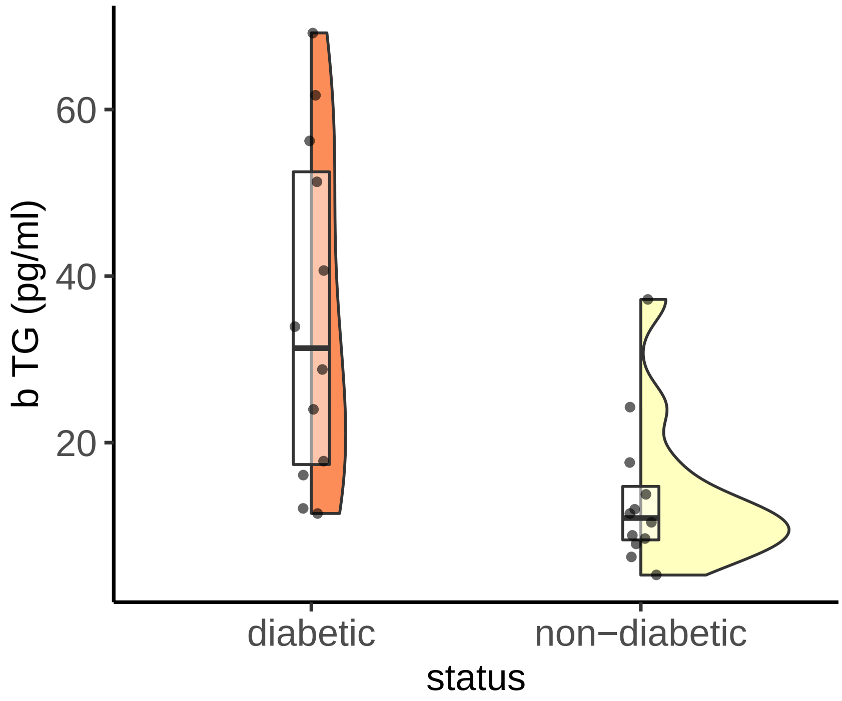
3.4 Explore the characteristics of distributions

The distributions of the two groups can be explored with appropriate plots and summary statistics.

3.4.1 Visualization of the distributions

We can visualize the distribution of beta-thromboglobulin for the two groups:

```
ggplot(thromboglobulin, aes(x = status, y = b_tg)) +
  geom_flat_violin(aes(fill = status), scale = "count") +
  geom_boxplot(width = 0.11, outlier.shape = NA, alpha = 0.5) +
  geom_point(position = position_jitter(width = 0.05),
             size = 1.2, alpha = 0.6) +
  scale_fill_brewer(palette = "Spectral") +
  labs(y = "b TG (pg/ml)") +
  theme_classic(base_size = 14) +
  theme(legend.position="none",
        axis.text = element_text(size = 14))
```



The above figure shows that the data in non-diabetic group are not symmetrical and the two groups have different shaped distributions.

3.4.2 Summary statistics

Summary statistics can also be inspected in each group:

```
thromboglobulin_summary <- thromboglobulin %>%
  group_by(status) %>%
  dplyr::summarise(
    n = n(),
    min = min(b_tg, na.rm = TRUE),
    q1 = quantile(b_tg, 0.25, na.rm = TRUE),
    median = quantile(b_tg, 0.5, na.rm = TRUE),
    q3 = quantile(b_tg, 0.75, na.rm = TRUE),
    max = max(b_tg, na.rm = TRUE),
    mean = mean(b_tg, na.rm = TRUE),
    sd = sd(b_tg, na.rm = TRUE),
    skewness = EnvStats::skewness(b_tg, na.rm = TRUE),
    kurtosis= EnvStats::kurtosis(b_tg, na.rm = TRUE)
  ) %>%
  ungroup()

thromboglobulin_summary
```

status	n	min	q1	median	q3	max
diabetic	12	11.5	17.375	31.35	52.525	69.2
non-diabetic	12	4.1	8.325	10.95	14.750	37.2

status	mean	sd	skewness	kurtosis
diabetic	35.27500	20.270001	0.4054531	-1.303061
non-diabetic	13.53333	9.194498	1.8098075	3.471819

The means are not very close to the medians (35.3 vs 31.3 and 13.5 vs 11.0) and the two standard deviations (20.3 vs 9.2) are different. The skewness is close to zero (symmetric distribution) for the diabetic group but the kurtosis is lower than zero (platykurtic). Moreover, both the skewness and the kurtosis for the non-diabetic group are higher than zero indicating non-normal leptokurtic distribution.

Alternatively, we can download the `{dlookr}` package and use the `describe()` function:

```
thromboglobulin %>%
  group_by(status) %>%
  dlookr::describe(b_tg) %>%
  select(variable,  status, n, mean, sd, p25, p50, p75, skewness, kurtosis) %>%
  ungroup()
```

```
## # A tibble: 2 x 10
##   variable status           n  mean    sd   p25   p50   p75 skewness
##   <chr>    <fct>        <int> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>
## 1 b_tg     diabetic        12  35.3 20.3  17.4   31.4  52.5    0.405
## 2 b_tg     non-diabetic    12  13.5  9.19  8.32  11.0  14.8    1.81
##   kurtosis
##      <dbl>
## 1    -1.30
## 2     3.47
```

3.4.3 Shapiro-Wilk test for normality

Additionally, we can check the normality applying the `Shapiro-Wilk` test:

```
thromboglobulin %>%
  group_by(status) %>%
  shapiro_test(b_tg) %>%
  ungroup()
```

status	variable	statistic	p
diabetic	b_tg	0.921	0.292
non-diabetic	b_tg	0.817	0.015

We can see that the data for the non-diabetic group is not normally distributed (p=0.015 <0.05) according to the Shapiro-Wilk test.

3.5 Run the Wilcoxon-Mann-Whitney (WMW) test

The differences between the two groups can be tested using a rank test such as Wilcoxon-Mann-Whitney (WMW):

```
thromboglobulin %>%
     wilcox_test(b_tg ~ status)
```

.y.	group1	group2	n1	n2	statistic	p
b_tg	diabetic	non-diabetic	12	12	125.5	0.002

Based on the p-value (p <0.05) the result is significant.

3.6 Present the results in a summary table

The median (or another percentile) can be used as a summary measure. The choice is guided by the shape of the distributions.

Characteristic	**Overall**, N = 24[1]	**diabetic**, N = 12[1]	**non-diabetic**, N = 12[1]	**p-value**[2]
b_TG (pg/ml)	16.9 (11.2, 34.7)	31.4 (17.4, 52.5)	10.9 (8.3, 14.8)	0.002

[1] Median (IQR)

[2] Wilcoxon rank sum test

Hence, the urinary β thromboglobulin excretion is significantly higher in diabetic group, median (IQR) 31.4 (17.4, 52.5) pg/ml, as compared to non-diabetic group, 10.9 (8.3, 14.8) pg/ml, ($p = 0.002 <0.05$).