Lab 07 – Non-parametric tests & Chi-squared tests

ENVX1002 Handbook

Semester 1, 2026

Learning Outcomes

- Learn to use R to perform a Wilcoxon signed-rank test and a Mann-Whitney U test.
- Learn to use R to calculate a chi-squared test for:
  - Test of proportions
  - Test of independence
- Learn how to interpret statistical output.

Before we begin

Create your Quarto document and save it as `Lab-07.qmd` or similar. There are no data files to download for this lab.

Quick introduction to non-parametric tests

Non-parametric tests are statistical tests that do not assume a specific distribution for the data. They are often used when the assumptions of parametric tests (such as normality) are not met. Non-parametric tests are also useful when dealing with ordinal data or when the sample size is small.

The two non-parametric tests we will cover in this lab are: 1. Wilcoxon signed-rank test - used to compare one sample to a known value or to compare two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (similar to a one sample t-test or paired t-test).

Exercise 1: dog pain (walk-through)

Background

A study measures dogs' pain using the Glasgow CMPS-SF scale before and after medication.

```
before <- c(7, 6, 8, 5, 9, 7, 6, 8, 7, 5)
after  <-  c(4, 5, 6, 3, 7, 5, 4, 6, 5, 3)

## create a data frame
```

```r
pain <- data.frame(
  dog_ID = 1:length(before),
  before = before,
  after = after,
  diff = after - before
)
```

EDA

```r
library(ggpubr)
# Create a comparative boxplot of the before and after pain scores
pain_long <- pivot_longer(pain, cols = c(before, after),
                          names_to = "time", values_to = "score")

ggboxplot(pain_long, x = "time", y = "score",
          color = "time", palette = c("#00AFBB", "#E7B800"),
          add = "jitter") +
 labs(title = "Pain Scores Before and After Medication",
       x = "Time", y = "Pain Score")
```

It looks like the pain scores are lower after medication. But we need to test this statistically.

Hypothesis (H)

The null hypothesis is that there is no difference in pain scores before and after treatment. The alternative hypothesis is that there is a difference in pain scores before and after treatment.

Assumptions (A)

```r
ggqqplot(pain$diff, main = "QQ-Plot of Differences"
         , ylab = "Differences")
```

We can see that the pain scores are not normally distributed, they are highly kurtotic (you can try `shapiro.test` on the difference).

The Wilcoxon signed-rank test is a non-parametric test that does not assume normality, so we can use it to compare the pain scores before and after treatment.

The assumptions of the Wilcoxon signed-rank test are: 1. In this case the data are paired (i.e., the same subjects are measured before and after treatment). 2. The differences between the paired observations are continuous/ordinal and symmetric about the median (looking at the boxplots, this is the case).

Test (T)

```r
# Perform the Wilcoxon signed-rank test
wilcox.test(pain$before, pain$after, paired = TRUE)
```

The output shows the test statistic (V) and the p-value. The test statistic is the sum of the ranks of the positive differences (after - before). The p-value is the probability of observing a test statistic as extreme as the one obtained, assuming that the null hypothesis is true.

We also see that the test is Wilcoxon signed rank test with continuity correction. This is because the test is based on ranks, and the continuity correction is applied to adjust for the fact that we are using a discrete distribution to approximate a continuous distribution.

P-value (P)

We can see that the p-value is less than 0.05, so we reject the null hypothesis and conclude that there is a significant difference in pain scores before and after treatment.

Conclusion (C)

Given that the p-value is significant (P<0.05) and looking at the output of the boxplot, we can conclude that the pain scores are significantly lower after treatment.

Comparison

Now analyse the data with a paired t-test. What do you find? Key is to compare the p-values. If you notice the p-value in the t-test is smaller, this indicates that the t-test is more powerful than the Wilcoxon signed-rank test. This is because the t-test assumes normality, and in this case, the data is not normally distributed. So you are more likely to make a type I error (reject the null hypothesis when it is true) with the t-test than with the Wilcoxon signed-rank test.

Exercise 2: Chi-squared tests

Quick introduction

The chi-square test is used to compare the *observed* distribution to an *expected* distribution, in a situation where we have **two or more categories** in discrete data. In other words, it compares multiple observed proportions to expected probabilities.

The formula is:

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

where O_i is the observed frequency, E_i is the expected frequency for each category, and k is the number of categories.

For more information about the technique, consult your lecture slides and tutorial 7.

Wild tulips (walk-through)

Background

Suppose we collected wild tulips and found that 81 were red, 50 were yellow and 27 were white. Are these colours equally common?

If these colours were equally distributed, the expected proportion would be 1/3 for each of the colour. Therefore, we want to test if the observed proportions are significantly different from the expected proportions.

The data is below.

```
tulip <- c(81, 50, 27)
```

Instructions

Utilise the **HATPC** process and test the hypothesis that the proportion of flower colours of tulips are equally common, assuming that the samples are independent. We can explore the data as we check the assumptions of the test.

HATPC:

- Hypothesis
- Assumptions
- Test (statistic)
- P-value
- Conclusion

> **i** Level of significance
>
> **The level of significance is usually set at 0.05.** This value is generally accepted in the scientific community and is also linked to Type 2 errors, where choosing a lower significance increases the likelihood of failing to reject the null hypothesis when it is false.

Hypotheses

What are the null hypothesis and alternative hypotheses?

Click here to view answer

- H_0: There is no significant difference between the observed and the expected proportions of flower colours.
- H_1: There is a significant difference between the observed and the expected proportions of flower colours.

Assumptions

Recall that the assumptions of the χ^2 test are:

1. No cell has expected frequencies less than 1
2. No more than 20% of cells have expected frequencies less than 5

> **Note**
>
> In the case that the above assumptions are violated then the probability of a type 1 error occurring (rejecting the null hypothesis when it is true, i.e. false positive) increases.

To calculate expected frequencies, we first calculate the total number of tulips and then divide by the number of categories.

```
expected <- rep(sum(tulip) * 1 / 3, 3) #rep function replicates the value
we're calculating inside the brackets
expected
```

Does the data satisfy the assumptions of a χ^2 test?

Click here to view answer Yes, as expected frequencies > 5.

Test statistic

The `chisq.test()` function in R is used to calculate the chi-squared test.

```
res <- chisq.test(tulip, p = c(1 / 3, 1 / 3, 1 / 3))
res
```

Note that we could check our assumptions post-analysis by checking the expected frequencies stored in the `expected` object of the output:

```
res$expected
```

P-value

Write down how you should report the critical value, p-value and df in a scientific paper?

Click here to view answer $\chi^2 = 27.9, d.f. = 2, p < 0.01$

Conclusions

Based on the p-value, do we accept or reject the null hypothesis?

Click here to view answer We reject the null hypothesis as the p-value is less than 0.05.

Now write a scientific (biological) conclusion based on the outcome.

Click here to view answer There is a significant difference in the proportion of flower colours of tulips ($\chi^2 = 27.9, d.f. = 2, p < 0.01$).

Exercise 3: hermit crabs

Background

In a study of hermit crab behaviour at Point Lookout, North Stradbroke Island, a random sample of 3 types of gastropod shells was collected. Each shell was then scored as being either occupied by a hermit crab or empty. Do hermit crabs prefer a certain shell?

Shell species	Occupied	Empty
Austrochochlea	47	42
Bembicium	10	41
Cirithid	125	49

The data is stored in a `table` object in R below. Note that it is different from a `data.frame` object. You can verify this by using the `str()` or `class()` functions.

```
crabs <- as.table( #Make a table with values for each row
  rbind(
    Aus = c(47, 42),
    Bem = c(10, 41),
    Cir = c(125, 49)
  )
)

colnames(crabs) <- c("Occupied", "Empty") #Add column names to table
str(crabs)
crabs
```

Data exploration

Since we have a multi-dimensional dataset, we can try to plot the data to visualise it.

A mosaic plot is a graphical representation of the data in a two-way contingency table. It is a way of visualising the relationship between two categorical variables.

Try the code below. Can you interpret the plot?

```
# mosaic plot of crabs
mosaicplot(crabs, main = "Hermit crabs and shell species")
```

Click here to view interpretation The plot shows that the distribution of hermit crabs in the different shell species is not equal. The majority of hermit crabs are found in the *Cirithid* shell species, followed by *Austrochochlea* and *Bembicium*. This can be observed by the **width** of the boxes in the plot.

There are also differences in the number of empty shells in the different shell species. The *Bembicium* shell species has the highest number of empty shells, followed by *Austrochochlea* and *Cirithid*. This can be observed by the **height** of the boxes in the plot.

HATPC Analysis

Now it is your turn to test the hypothesis that the three shell species are equally preferred by hermit crabs. Follow the HATPC process with the following questions in mind (but you don't have to answer them individually):

1. What are the null hypothesis and alternative hypotheses?
2. Does the data satisfy the assumptions of a χ^2 test?
3. How should you report the critical value, p-value and df in a scientific paper?
4. Based on the p-value, do we accept or reject the null hypothesis?
5. Write a scientific (biological) conclusion based on the outcome.

Take your time, and when you are ready, check your answers with your demonstrators.

Exercise 4: pregnancies

Quick introduction

The Mann-Whitney U test (also called the Wilcoxon rank-sum test) is a non-parametric alternative to the independent samples t-test. It's used to compare two independent groups when the data doesn't meet the assumptions for a t-test, particularly when:

- The data is not normally distributed
- The sample sizes are small
- The data is ordinal rather than continuous

The test works by ranking all observations from both groups together, then comparing the sum of ranks between the two groups. If there's a significant difference in these rank sums, it suggests that the distributions of the two groups differ.

Background

In a medical study, researchers measured the permeability constants of the human chorioamnion (a placental membrane) between different gestational ages. They collected samples from:

- Term pregnancies (x): pregnancies at full term (around 40 weeks)
- Early pregnancies (y): pregnancies between 12 to 26 weeks gestational age

The researchers wanted to test if the permeability of the membrane is greater at term compared to earlier in pregnancy. Higher permeability might indicate changes in the membrane's function as pregnancy progresses.

```
# Permeability constants at term (x)
term <- c(0.80, 0.83, 1.89, 1.04, 1.45, 1.38, 1.91, 1.64, 0.73, 1.46)

# Permeability constants at 12-26 weeks (y)
```

```r
early <- c(1.15, 0.88, 0.90, 0.74, 1.21)

# Create a data frame for visualisation
placenta <- data.frame(
  permeability = c(term, early),
  group = factor(c(rep("Term", length(term)), rep("Early", length(early))))
)
```

EDA

Let's visualise the data using boxplots to see if there are visible differences between the two groups:

```r
# Create a comparative boxplot
ggplot(placenta, aes(x = group, y = permeability, fill = group)) +
  geom_boxplot() +
  geom_jitter(width = 0.2, alpha = 0.6) +
  labs(title = "Placental Membrane Permeability by Gestational Age",
       x = "Pregnancy Stage",
       y = "Permeability Constant") +
  theme_minimal()
```

Also, let's check if the data is normally distributed using QQ plots:

```r
# QQ plots for both groups
par(mfrow = c(1, 2))
qqnorm(term, main = "QQ Plot for Term Pregnancies")
qqline(term)
qqnorm(early, main = "QQ Plot for Early Pregnancies")
qqline(early)
```

Hypothesis (H)

Based on the research question, we can formulate the following hypotheses:

- $\mathbf{H_0}$: The permeability of the membrane at term is not greater than the permeability at earlier stages (12-26 weeks).
- $\mathbf{H_1}$: The permeability of the membrane at term is greater than the permeability at earlier stages (12-26 weeks).

Note that this is a one-sided test, as we're specifically interested in whether the permeability is *greater* at term, not just whether it differs.

Assumptions (A)

The Mann-Whitney U test has the following assumptions:

1. The observations from both groups are independent of each other.
2. The observations are ordinal (i.e., can be ranked).

3. The distributions of both populations have the same shape (though this is relaxed for large samples).

From our EDA, we can see that:

- The sample sizes are small (10 for term and 5 for early pregnancies).
- The QQ plots suggest that normality might be questionable, especially with such small sample sizes.
- The observations are certainly independent as they come from different individuals.

Given these observations, the Mann-Whitney U test is an appropriate choice over a t-test. Note that we may run a formal test for normality (e.g., Shapiro-Wilk test) if we wanted to be thorough, but the QQ plots already suggest that normality is *not* a strong assumption here and the Mann-Whitney U test is robust to deviations from normality. Even if the data were normally distributed, the Mann-Whitney U test would still be valid – just *less* powerful than a t-test (i.e., it would have a higher chance of making a type II error, or an error of failing to reject the null hypothesis when it is false).

Test (T)

We'll use the `wilcox.test()` function to perform the Mann-Whitney U test. Since we're interested in whether the permeability is greater at term, we'll use a one-sided alternative hypothesis:

```
# Perform the Mann-Whitney U test
wilcox.test(term, early, alternative = "greater")
```

We can also use the large-sample approximation method (without continuity correction) as described by Hollander & Wolfe:

```
# Mann-Whitney U test with large sample approximation
wilcox.test(term, early, alternative = "greater",
            exact = FALSE, correct = FALSE)
```

P-value (P)

The p-value from the test is 0.1914, which is greater than our significance level of 0.05.

Conclusion (C)

Based on the results of the Mann-Whitney U test, we fail to reject the null hypothesis. There is insufficient evidence to conclude that the permeability of the human chorioamnion is greater at term compared to earlier in pregnancy (12-26 weeks).

Despite what we might observe visually in the boxplots, the statistical test doesn't support the claim of higher permeability at term. This could be due to several factors:

1. Small sample sizes (only 10 term and 5 early samples)
2. High variability within the term group
3. The actual biological relationship might be more complex than our simple hypothesis

Question 1

Why might researchers choose a Mann-Whitney U test over a t-test in this situation? Consider both the sample size and what you observed in the QQ plots.

Question 2

What does the value "W = 31" in the test output represent? What information does it give us about the comparison between the two groups?

Bonus take home exercises

Exercise 1: National RSPCA statistics

The RSPCA releases statistics on the number of animals they receive, reclaim and rehome every year. In the 2023-24 financial year, the RSPCA received 17468 dogs, 26704 cats, and 37497 other animals. The "other" category includes horses, small animals, livestock and wildlife.

Using the HATPC framework, test whether these animals were received in equal proportions.

```
received <- c(17468, 26704, 37497)
```

Exercise 2: UC Berkeley Admissions

For the two exercises we're going to use simplified versions of the inbuilt dataset 'UCBAdmissions', which has data on student admissions for Berkeley. The dataset shows how many students were rejected and admitted to the university by both department and gender.

2.1 Admissions by department

Did every department at UC Berkeley admit students in equal proportions?

```
dept_admissions <- c(601, 370, 322, 269, 147, 46)
```

2.2 Gender differences in admissions

Are male and female students admitted and rejected in the same proportion?

```
gender <- as.table( #Make a table with values for each row
  rbind(
    admitted = c(1198, 557),
    rejected = c(1493, 1278)
  )
)

colnames(gender) <- c("Male", "Female") #Add column names to table
```