Statistics and R short course

Session 4 - Practical (solutions)

Vester Gunsaru

28 January 2026

Session 4 - Practical

Exercise 1

Take the iris dataset that we worked with during Sessions 1 and 2.

Explain how you would, in a formal statistical way, compare the following:

1. `Petal.Width` between the flower species `virginica` and `setosa`.
2. `Sepal.Length` between all 3 flower species.

For each comparison, state which test you will use (there may be more than one valid option!), state the null and alternative hypotheses, do the test and interpret the results.

Exercise 1 (solution)

1. `Petal.Width` between the flower species `virginica` and `setosa`.

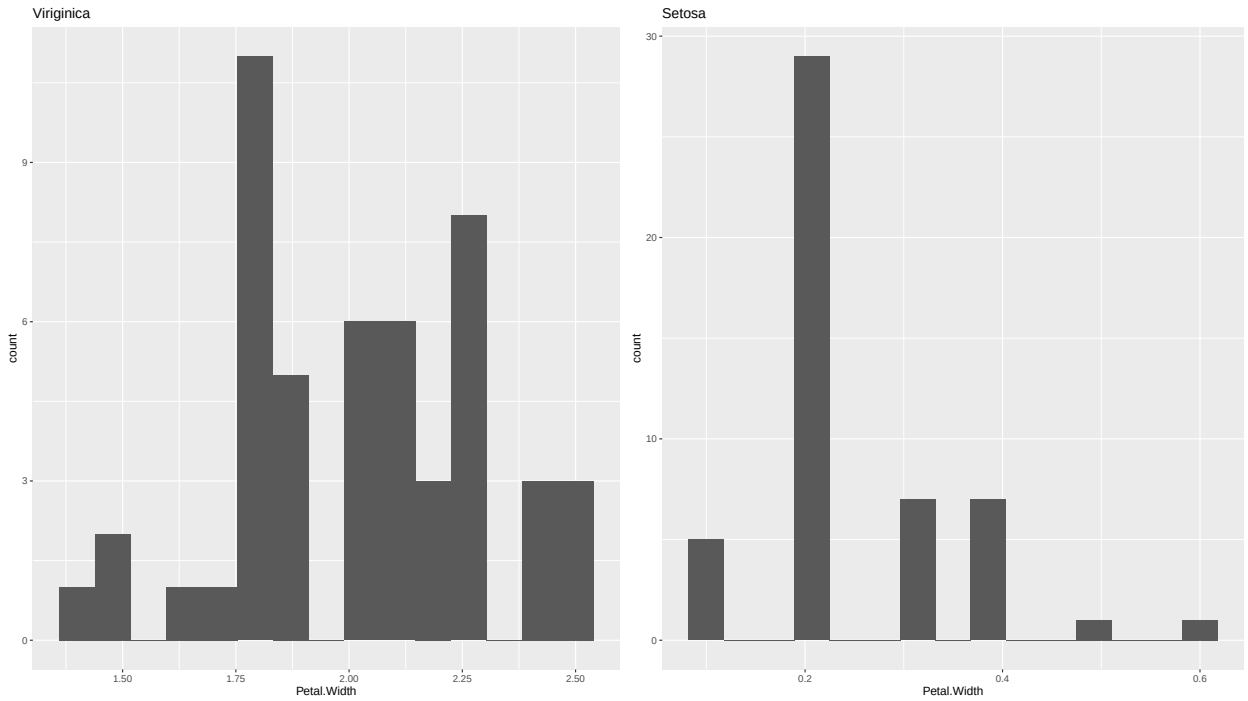
We have 50 observations for each flower type – large enough for the Central Limit Theorem (CLT) to guarantee that sample means are approximately normally distributed as long as the data are not too severely non-normal (outliers etc).

Let's quickly check the distribution of `Petal.Width` in the 2 flower species:

```
g1<-iris %>%
  filter(Species=="virginica") %>%
  ggplot(mapping=aes(x=Petal.Width)) +
  geom_histogram(bins=15) +
  ggtitle("Viriginica")

g2<-iris %>%
  filter(Species=="setosa") %>%
  ggplot(mapping=aes(x=Petal.Width)) +
  geom_histogram(bins=15) +
  ggtitle("Setosa")

grid.arrange(g1,g2,nrow=1)
```



The data do not look particularly normally distributed, but there is no instance of severe non-normality either. The t-test should be OK to use, given the CLT.

As we only want to assess whether `Petal.Width` is the same or not across the 2 flower species, we will do a two-sided test. We have no reason to believe one or the other flower species should have larger values.

For a 2 sample t-test, the null and alternative hypotheses are:

$$H_0 : \mu_v = \mu_s$$
$$H_1 : \mu_v \neq \mu_s$$

where μ_v, μ_s are the population means for the virginica and setosa flower species respectively.

We can now proceed to do the two-sided, two-sample t-test:

```
t.test(Petal.Width~Species,data=iris %>% filter(Species %in% c("virginica","setosa")))
```

```
##
##  Welch Two Sample t-test
##
## data:  Petal.Width by Species
## t = -42.786, df = 63.123, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group setosa and group virginica is not equa
## 95 percent confidence interval:
##  -1.863133 -1.696867
## sample estimates:
##    mean in group setosa mean in group virginica
##                   0.246                   2.026
```

The p-value is essentially 0, so we reject the null hypothesis that the mean `Petal.Width` is the same in both groups. Under the null hypothesis H_0 it is very unlikely that we would have observed the data we collected, hence we H_0.

Note: it would also be OK to do a Wilcoxon rank-sum test and this gives the same result (p-value essentially 0, reject H_0):

```
wilcox.test(Petal.Width~Species,data=iris %>% filter(Species %in% c("virginica","setosa")))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Petal.Width by Species
## W = 0, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

2. `Sepal.Length` between all 3 flower species.

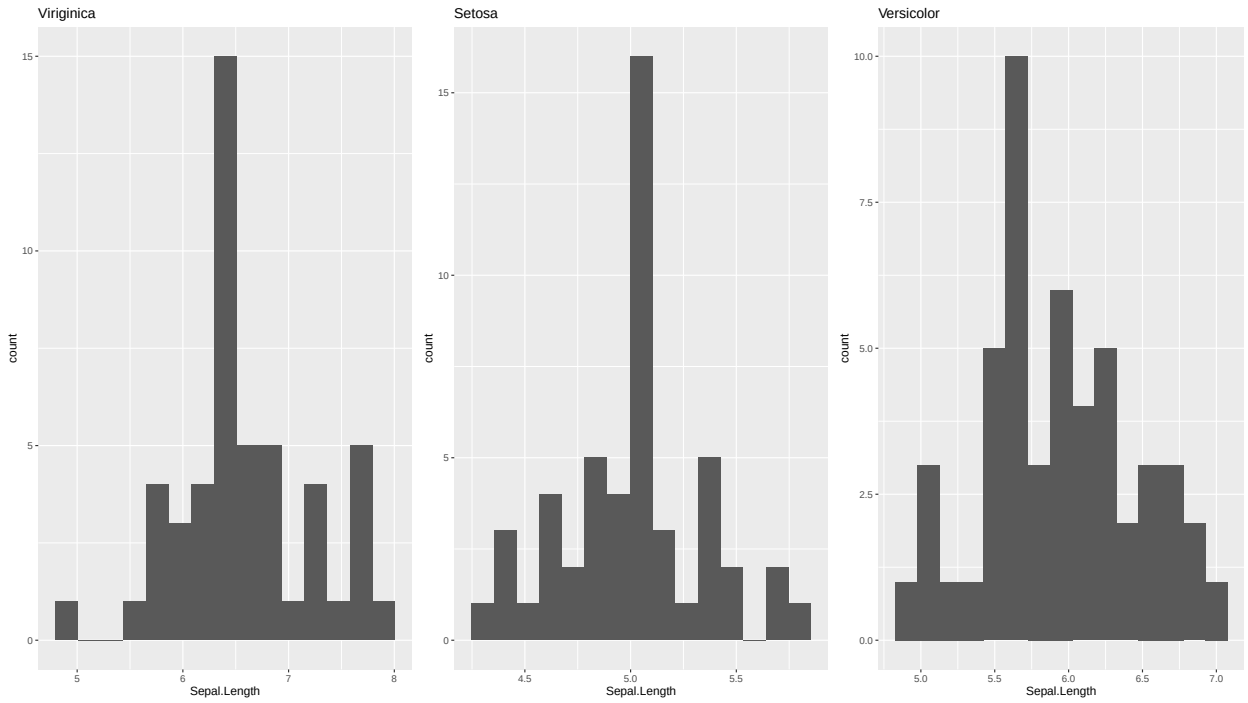
Now we are comparing 3 groups, not 2. To check whether we can use ANOVA or need to use the Kruskal-Wallis test, we need inspect that the data are not severely non-normal:

```
g1<-iris %>%
  filter(Species=="virginica") %>%
  ggplot(mapping=aes(x=Sepal.Length)) +
  geom_histogram(bins=15) +
  ggtitle("Viriginica")

g2<-iris %>%
  filter(Species=="setosa") %>%
  ggplot(mapping=aes(x=Sepal.Length)) +
  geom_histogram(bins=15) +
  ggtitle("Setosa")

g3<-iris %>%
  filter(Species=="versicolor") %>%
  ggplot(mapping=aes(x=Sepal.Length)) +
  geom_histogram(bins=15) +
  ggtitle("Versicolor")

grid.arrange(g1,g2,g3,nrow=1)
```



This looks OK to use ANOVA.

The null and alternative hypotheses will be:

$$H_0 : \mu_s = \mu_{ve} = \mu_{vi}$$
$$H_1 : \mu_i \neq \mu_j \quad \text{for some i,j}$$

```
oneway.test(Sepal.Length~Species,data=iris)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  Sepal.Length and Species
## F = 138.91, num df = 2.000, denom df = 92.211, p-value < 2.2e-16
```

The p-value is again essentially 0 and so we reject the null hypothesis H_0. We conclude that there is enough evidence to suggest that the mean values for `Sepal.Length` are different across the 3 flower species.

Note: as above, we can always do the non-parametric test. Nonparametric tests have slightly less power than the parametric tests if the parametric assumptions are met, but that does not mean that it's not possible to use the non-parametric test when you can use an equivalent parametric test.

Here, if we used Kruskal-Wallis:

The null and alternative hypotheses are somehwat different:

$$H_0 : \text{Sepal.Length in all groups has the same distribution.}$$
$$H_1 : \text{The distribution of Sepal.Length is not the same across all groups.}$$

```
kruskal.test(Sepal.Length~Species,data=iris)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Sepal.Length by Species
## Kruskal-Wallis chi-squared = 96.937, df = 2, p-value < 2.2e-16
```

The p-value is essentially 0, so we reject the null hypothesis.

Exercise 2

In a drug trial, researchers are assessing overall in-hospital mortality as the primary outcome. The new drug is compared against the standard-of-care treatment (SOC). Patients are randomised 1:1 to the new drug and SOC. At trial conclusion, the researchers observe that out of 250 SOC patients, 61 have died and out of 250 patients on the new drug arm, 48 have died.

Perform a statistical test to conclude whether or not there is a difference between the new drug and the SOC. State the test you use, the null and alternative hypotheses, perform the test and interpret the results.

Exercise 2 (solution)

Here we need to compare the proportions of patients that die in hospital during the study period, so we will need to do a two-sample test for proportions.

Let p_{drug}, p_{SOC} be the proportion of patients dying on the new drug regime and on the SOC arm respectively.

The null and alternative hypotheses are:

$$H_0 : p_{drug} = p_{SOC}$$
$$H_1 : p_{drug} \neq p_{SOC}$$

Performing the test, we get:

```
res<-prop.test(x=c(48,61),n=c(250,250))
res
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(48, 61) out of c(250, 250)
## X-squared = 1.6894, df = 1, p-value = 0.1937
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.12823737  0.02423737
## sample estimates:
## prop 1 prop 2
##  0.192  0.244
```

The p-value is $0.1936817 > 0.05$, so we do not reject H_0, there is not enough evidence (at the 5% significance level) to suggest that the proportions in both groups are different,

Note that a Fisher's exact test could also be used here (note the different specification of the data for this test):

```
res<-fisher.test(x=matrix(c(48,61,250-48,250-61),byrow=T,nrow=2))
res
```

```
##
##  Fisher's Exact Test for Count Data
##
```

```
## data:  matrix(c(48, 61, 250 - 48, 250 - 61), byrow = T, nrow = 2)
## p-value = 0.1935
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.4686843 1.1531506
## sample estimates:
## odds ratio
##  0.7367018
```

The p-value is 0.1935381, almost the same as for the two-proportion test (which uses a normal distribution approximation) and since this is > 0.05, we do not reject H_0. As for the two-proportion test, we conclude that there is not enough evidence (at the 5% significance level) to suggest that the proportions in both groups are different.

Exercise 3

Test whether the 2 variables from Table 1 below are independent or not. State the test you use, the null and alternative hypotheses, do the test and interpret the results.

Table 1: Summary of patient outcomes for different health centers.

	alive	dead
Hospital1	92	29
Hospital2	54	15
Hospital3	31	3

What about when you repeat your analysis for Table 2 below?

Table 2: Summary of patient outcomes for different health centers.

	alive	dead
Hospital1	920	290
Hospital2	540	150
Hospital3	310	30

Comment on the results from your analyses for both tables.

Exercise 3 (solution)

We are assessing whether 2 categorical variables are independent or not. If possible, we would use the exact Fisher test for this.

The null and alternative hypotheses are:

$$H_0: \text{ Health centre and outcome are independent.}$$

$$H_0: \text{ Health centre and outcome are not independent.}$$

Note that we could also express this as $H_0 : p_1 = p_2 = p_3$ and $H_1 : p_i \neq p_j$ for some i, j.

We can proceed to do the test:

```
res<-fisher.test(matrix(c(92,29,54,15,31,3),byrow=T,ncol=2))
res
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  matrix(c(92, 29, 54, 15, 31, 3), byrow = T, ncol = 2)
## p-value = 0.1483
## alternative hypothesis: two.sided
```

The p-value is $0.1483358 > 0.05$, so at the 5% significance level we do not reject H_0, there is not enough evidence to suggest that outcome depends on the health centre.

Everything (test to use, null and alternative hypotheses) remains to same for the table with larger counts (note that Table 2 has the same cell counts as Table 1, just multiplied by 10 – the sample size is 10 times larger), but we get a different result.

```
res<-fisher.test(matrix(c(920,290,540,150,310,30),byrow=T,ncol=2))
res
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  matrix(c(920, 290, 540, 150, 310, 30), byrow = T, ncol = 2)
## p-value = 5.149e-10
## alternative hypothesis: two.sided
```

Now the p-value is $5.1493727 \times 10^{-10} < 0.05$, so we reject H_0 at the 5% significance level. There is considerable evidence that the outcome depends on the health centre.

What has changed is simply the sample size: with the larger sample size we can conclude that the same differences in outcome proportions between health centres are not due to random chance, but are likely a real feature. For the table with less counts, we did not have enough evidence to conclude this – it was reasonably possible to observe the data even under the null hypothesis H_0.

Exercise 4 - sample size calculation

Researchers want to estimate the effect of malnutrition in early childhood on body height at adult age. For this, the researchers recruit former participants of a childhood malnutrition cohort study. Recruited participants are known to have been malnourished or not at any point in the previous cohort study.

Assume a standard deviation of 6cm for height, power of 90%, significance level of 5% and equal group sizes (malnourished and non-malnourished groups). How many former cohort study participants will need to be recruited, if the researchers want to be powered to detect an average difference in height of 2cm or more?

Exercise 4 (Solution)

```
library(pwr)

sampsize<-pwr.t.test(d=2/6,sig.level=0.05,power=0.9)
print(sampsize)
```

```
##
##      Two-sample t test power calculation
##
##               n = 190.0991
##               d = 0.3333333
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
```

```
##
## NOTE: n is number in *each* group
```

```
n<-ceiling(sampsize$n)
print(n)
```

```
## [1] 191
```

Per group 191 participants will be needed, i.e. 382 participants in total.