R Intermediate Short Course
Session 4 - Hypothesis Testing
(Two sample Proportions)

Facilitator: Rajesh Lakhan

The University of the West Indies, St Augustine

Thursday 21th July 2022
(5:00pm - 7:00pm)
(online)

For the two-sample hypothesis test we compare the difference between two proportions that are assumed to come from a normal population.

The types of hypotheses tested are:

1. **Two tailed test** $H_0 : p_1 = p_2$ vs $H_1 : p_1 \neq p_2$
   or written as $H_0 : p_1 - p_2 = 0$ vs $H_1 : p_1 - p_2 \neq 0$
2. **Upper tailed test** $H_0 : p_1 \leq p_2$ vs $H_1 : p_1 > p_2$
3. **Lower tailed test** $H_0 : p_1 \geq p_2$ vs $H_1 : p_1 < p_2$

where the null hypothesis can be written as H_0: $p_1 = p_2$ regardless of the type of test.

$\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$ are the sample estimates for the population proportions, p_1 and p_2.

We reject the null hypothesis H_0 if,

1. The confidence interval does not contain the value at H_0
2. P-value $< \alpha$ (level of significance)

Note: R modifies the p-value for two-tailed test in it's output, so no adjustment needed.

Performing two-sample proportion tests in R:

Two tailed test:

```
prop.test(x, n, p=NULL, alternative = "two.sided",
          conf.level = CL, correct = T/F)
```

Upper tailed test:

```
prop.test(x, n, p=NULL, alternative = "greater",
          conf.level = CL, correct = T/F)
```

Lower tailed test:

```
prop.test(x, n, p=NULL, alternative = "less",
          conf.level = CL, correct = T/F)
```

where,
x = a vector of the number or successes
n = a vector of the number of trials
p = NULL since the value at H_0 is 0
alternative = " " depends on the type of test
and correct = TRUE/FALSE represents Yates correction for continuity.

Example 1:

In a sample of 200 imported vehicles, 90 vehicles cost more than \$100,000. In another sample of 275 imported vehicles, 125 vehicles cost more than \$100,000. Test at $\alpha = 0.03$ whether there is a difference in the proportion of vehicles that cost more than \$100,000 from both samples.

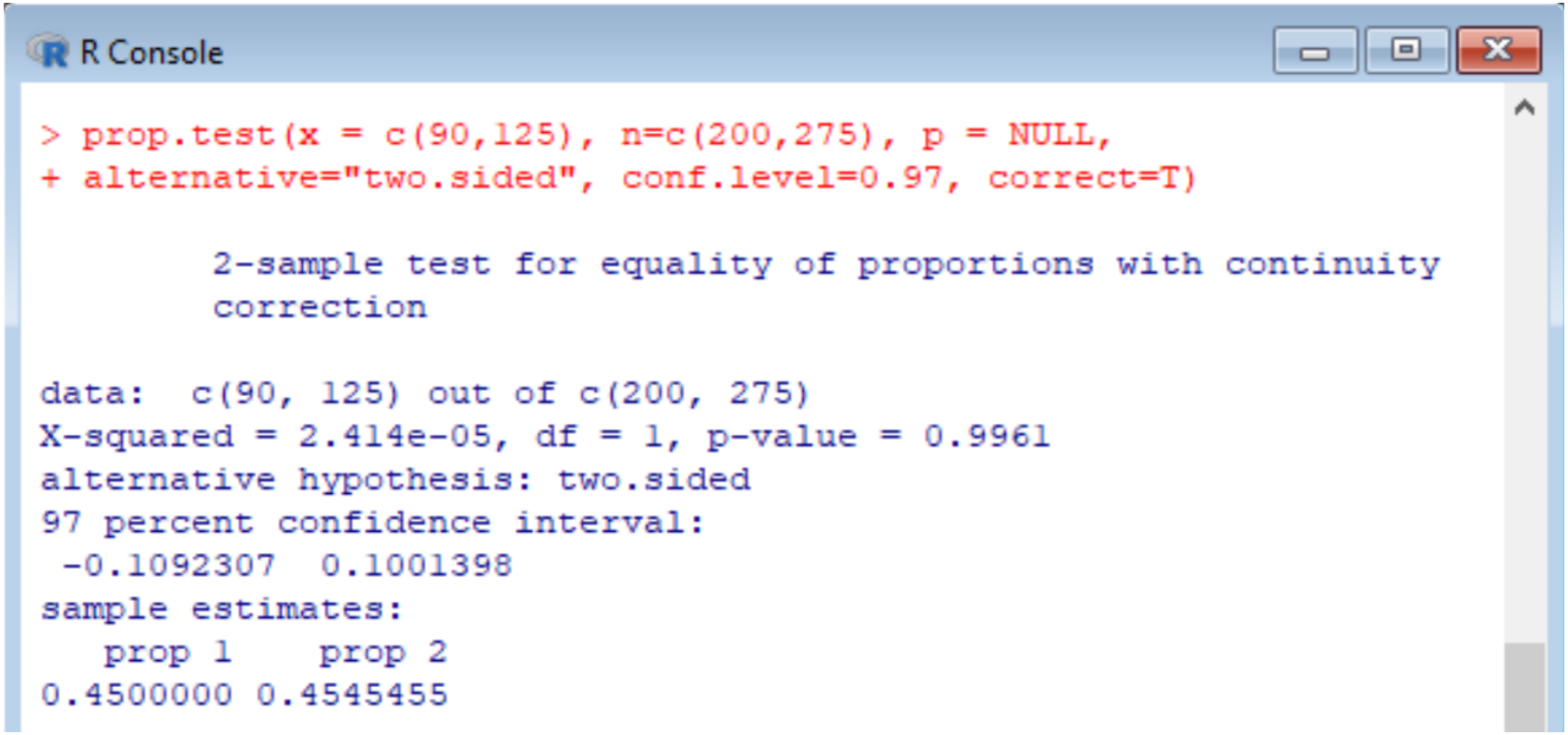
Required to test H_0: $p_1 = p_2$ vs H_1: $p_1 \neq p_2$

Parameters: $x_1 = 90$ and $x_2 = 125 \Rightarrow$ x = c(90, 125)

$n_1 = 200$ and $n_2 = 275 \Rightarrow$ n = c(200, 275)

CL = 1 - α = 1 - 0.03 = 0.97

R Output: (see next slide)



The 97% confidence interval (-0.109, 0.100) contains the value 0 and the p-value ($= 0.996$) is greater than the level of significance α ($= 0.03$).

Therefore we fail to reject H_0 concluding that there is no significant difference in the proportion of vehicles from both samples.

Example 2:

The *mtcars* built-in data set in R shows various features of 32 cars. Conduct a hypothesis test at $\alpha = 0.05$ to show that the proportion of vehicles having 4 gears is greater than the proportion of vehicles with 5 gears in the data.

Required to show that the test H_0: $p_1 \leq p_2$ vs H_1: $p_1 > p_2$ rejects H_0 at a 5% level of significance.

Parameters:

We know $n_1 = 32$ and $n_2 = 32$ since the sample size is 32.
We also know alternative = "greater" and CL = 0.95.

Now we must find x_1 and x_2 from the data set where x_1 is the number of vehicles that have 4 gears and x_2 is the number of vehicles with 5 gears.

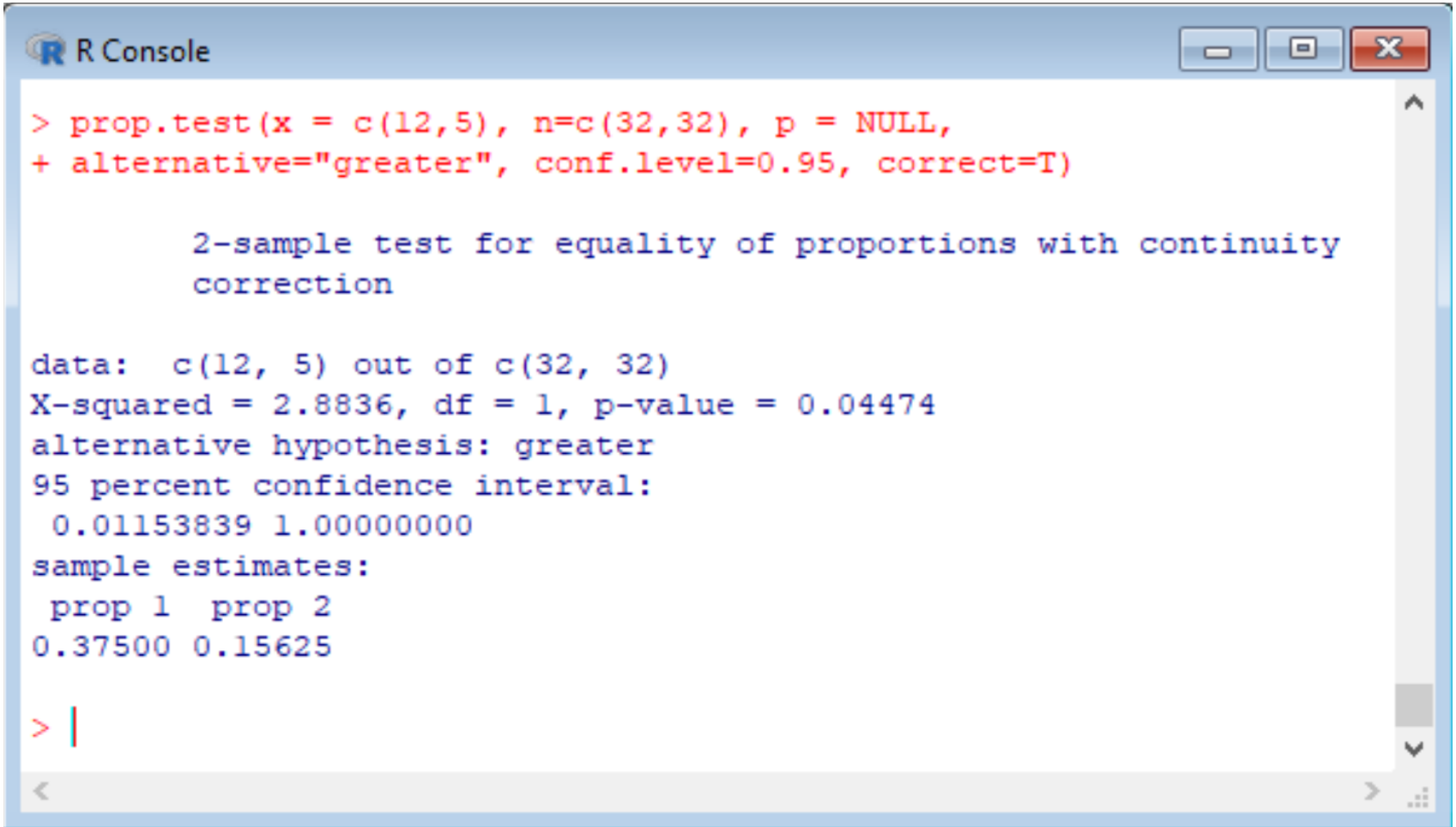
```
> attach(mtcars)
> library(data.table)
> x1 = which(gear == 4)
```

```
> length(x1)
[1] 12
> x2 = which(gear == 5)
> length(x2)
[1] 5
```

$x_1 = 12$ and $x_2 = 5 \Rightarrow$ x = c(12, 5)
Also n = c(n_1, n_2) = c(32, 32)

```
> prop.test(x=c(12,5), n=c(32,32), p=NULL,
alternative="greater", conf.level=0.95, correct=T)
```

R Output: (see next slide)



The 95% confidence interval (0.011, 1.000) does not include 0 and the p-value (= 0.045) is less than α. We can reject H_0 to show that the proportion of vehicles with 4 gears is greater than those with 5 gears.

In-class Exercises

Question 1

(15min + 7min)

The built-in data set *iris* gives the measurements of 4 features for different flowers each having 3 species of iris i.e. *setosa*, *versicolor* and *virginica*.

Construct a table showing the simulated means of the 4 features for each of the species in the data set. Your table should be labelled like this:

```
> table
           Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
setosa
versicolor
virginica
```

Question 2

(13min + 5min)

Let X be a random variable that follows a binomial distribution with parameters n = 2000 and p = 0.001. Also, let Y be a random variable that follows a Poisson distribution with mean 2.

(i) Compute the probabilities: $P(X \leq 3)$ and $P(Y = 4)$

(ii) By generating a sample of 50000 values of X and 50000 values of Y, construct histograms for X and Y. Use set.seed(8).

(iii) Compute the covariance and correlation between the samples of X and Y generated. What conclusion can you draw from these values?

Question 3

(10min + 5min)

The Orange built-in data set in R gives the growth by circumference of 5 orange trees as they age. Test at a 5% level of significance, whether the mean circumference of trees is significantly different from 110 mm.

Question 4

(13min + 7min)

A sample of 12 employees were evaluated on a monthly basis, using an employee performance metric. The higher the value of the performance metric the better the employee performs. Their performance after the 1st and 2nd month of evaluation were recorded as follows:

1^{st} month	0.70	0.65	0.96	0.23	0.51	0.63
2^{nd} month	0.70	0.68	0.84	0.44	0.49	0.70

1^{st} month	0.88	0.12	0.36	0.92	0.47	0.09
2^{nd} month	0.89	0.18	0.56	0.92	0.72	0.12

(i) Determine the mean difference in employee performance (manually enter the formula in R).

(ii) Conduct a hypothesis test at $\alpha = 0.10$ to determine if there is a decrease in the performance of employees.

(iii) Given that the test statistic is:

$$\frac{(\bar{d} - \mu_0)}{s_d/\sqrt{n}}$$

Calculate the test statistic, using R to "manually" coding it. (-1.922285)