Estadística

U4 4B

David Aarón Ramírez Olmeda

Marzo 2023

5.11 Play the piano. Georgianna claims that in a small city renowned for its music school, the average child takes at least 5 years of piano lessons. We have a random sample of 20 children from the city, with a mean of 4.6 years of piano lessons and a standard deviation of 2.2 years.

(a) Evaluate Georgianna's claim using a hypothesis test.

To evaluate Georgianna's claim, we can set up a hypothesis test as follows:

Null Hypothesis (H0): The true population mean of piano lessons for children in the city is less than or equal to 5 years.

Alternative Hypothesis (Ha): The true population mean of piano lessons for children in the city is greater than 5 years.

```
sample_mean <- 4.6
sample_sd <- 2.2
n <- 20
population_mean <- 5

t_value <- (sample_mean - population_mean) / (sample_sd / sqrt(n))
p_value <- pt(t_value, df = n - 1, lower.tail = FALSE)

cat("t-value = ", t_value, "\n")
```

```
## t-value =  -0.8131156
```

```
cat("p-value = ", p_value, "\n")
```

```
## p-value =  0.786888
```

```
if (p_value < 0.05) {
  cat("Reject the null hypothesis.")
} else {
  cat("Fail to reject the null hypothesis..")
}
```

```
## Fail to reject the null hypothesis..
```

(b) Construct a 95% confidence interval for the number of years students in this city take piano

To construct a 95% confidence interval for the true population mean of piano lessons for children in the city, we can use the following:

```
se <- sample_sd / sqrt(n)
lower <- sample_mean - t_value * se
upper <- sample_mean + t_value * se

cat("95% Confidence Interval = [", lower, ", ", upper, "]")
```

```
## 95% Confidence Interval = [ 5 ,  4.2 ]
```

We can interpret this interval as follows: we are 95% confident that the true population mean of piano lessons for children in the city is between 4.2 and 5 years

(c) Do your results from the hypothesis test and the confidence interval agree? Explain your reasoning.

The results from the hypothesis test and the confidence interval agree. In the hypothesis test, we failed to reject the null hypothesis that the true population mean of piano lessons for children in the city is less than or equal to 5 years. In the confidence interval, the entire interval is below 5 years. This means that there is not enough evidence to support Georgianna's claim that the average child in the city takes at least 5 years of piano lessons, and we can be 95% confident that the true population mean is less than 5 years.

**5.15 Air quality measurements were collected in a random sample of 25 country capitals in 2013, and then again in the same cities in 2014. We would like to use these data to compare average air quality between the two years.

(a) Should we use a one-sided or a two-sided test? Explain your reasoning.**

We should use a two-sided test, the reason is that we do not have any prior information or expectation about which year has a better air quality than the other. We need to test the hypothesis that the mean air quality in 2013 is equal to the mean air quality in 2014, versus the alternative hypothesis that the mean air quality in one year is different from the mean air quality in the other year.

(b) Should we use a paired or non-paired test? Explain your reasoning.

We should use a paired test, specifically a paired t-test, to compare the air quality measurements between 2013 and 2014. The reason is that we are interested in testing whether there is a significant difference in air quality between the two years within each city. Since we have paired data, where each observation in 2013 is matched with an observation in 2014 from the same city, a paired test will allow us to control for any city-specific factors that could affect air quality.

(c) Should we use a t-test or a z-test? Explain your reasoning.

We should use t-test rather than a z-test to compare the measurements. The reason is that we have a small sample size (n = 25), and we do not know the population standard deviation. Therefore, the t-test is more appropriate than the z-test, as it uses the sample standard deviation to estimate the standard error of the difference between the means.

Since we are comparing paired samples, a paired t-test is designed to test the hypothesis that the mean difference between the paired observations is zero.

**5.27 In the early 1990's, researchers in the UK collected data on traffic flow, number of shoppers, and traffic accident related emergency room admissions on Friday the 13th and the previous Friday, Friday the 6th. The histograms below show the distribution of number of cars passing by a specific intersection on Friday the 6th and Friday the 13th for many such date pairs. Also given are some sample statistics, where the difference is the number of cars on the 6th minus the number of cars on the 13th.

(a) Are there any underlying structures in these data that should be considered in an analysis? Explain**

Based on the given information, there are some potential underlying structures in these data that should be considered in an analysis.

The data on traffic flow, number of shoppers, and traffic accidents may be influenced by external factors that vary over time, such as weather, holidays, or events. The presence of outliers or extreme values should also be considered, and the histograms and sample statistics can provide some information on this. Additionally, the potential influence of confounding variables, such as time of day, location, or day of the week, should be taken into account to avoid biased results.

(b) What are the hypotheses for evaluating whether the number of people out on Friday the 6th is different than the number out on Friday the 13th?

The null hypothesis would be that there is no difference between the average number of people out on Friday the 6th and Friday the 13th:

H0: u = 0

The alternative hypothesis would be that there is a difference between the average number of people out on Friday the 6th and Friday the 13th:

Ha: u != 0

(c) Check conditions to carry out the hypothesis test from part (b).

To carry out the hypothesis test, we need to check the following conditions:

1. Independence: We assume that the researchers selected different Fridays at random for their data collection, and so this condition is likely met.
2. Normality: We can check this condition by examining the histogram of the differences, it appears that the distribution of differences is roughly symmetric and bell-shaped, so this condition is also likely met.
3. Equal variances: We can check this condition by conducting a formal test of equal variances.

(d) Calculate the test statistic and the p-value.

```
sample_mean_diff <- 1835
sample_sd_1 <- 7259
sample_sd_2 <- 7664

n1 <- 10
n2 <- 10
population_mean <- 0

t_value <- (sample_mean_diff - population_mean) /
  (sqrt((sample_sd_1**2 / n1) + (sample_sd_2**2 / n2)))
p_value <- pt(t_value, df = (n1 - 1) + (n2 - 1) , lower.tail = FALSE)

cat("t-value = ", t_value, "\n")
```

```
## t-value =  0.5497118
```

```
cat("p-value = ", p_value, "\n")
```

```
## p-value =  0.2946367
```

```r
if (p_value < 0.05) {
  cat("Reject the null hypothesis.")
} else {
  cat("Fail to reject the null hypothesis.")
}
```

```
## Fail to reject the null hypothesis.
```

(e) What is the conclusion of the hypothesis test?

Assuming a significance level of 0.05, since the p-value (0.29) is greater than the significance level, we fail to reject the null hypothesis and conclude that there is not enough evidence that the average number of cars passing by the intersection on Friday the 6th is different than the average number on Friday the 13th.

(f) Interpret the p-value in this context.

The p-value in this context represents the probability of obtaining a test statistic assuming the null hypothesis is true. The p-value indicates that it is likely that we would have obtained the observed difference in means by chance

(g) What type of error might have been made in the conclusion of your test? Explain.

In this hypothesis test, we might have made a Type 2 error if we fail to rejected the null hypothesis when it was actually false.

5.39 A large farm wants to try out a new type of fertilizer to evaluate whether it will improve the farm's corn production. The land is broken into plots that produce an average of 1,215 pounds of corn with a standard deviation of 94 pounds per plot. The owner is interested in detecting any average difference of at least 40 pounds per plot. How many plots of land would be needed for the experiment if the desired power level is 90%? Assume each plot of land gets treated with either the current fertilizer or the new fertilizer.

pwr is a package in R that provides functions for power analysis and sample size calculations. We need this package to calculate the necessary sample size for the experiment.

```r
# install.packages("pwr")
```

```r
mu1 <- 1215
sigma1 <- 94
mu2 <- 1215 + 40
alpha <- 0.05
power <- 0.9

library(pwr)
n <- pwr.t.test(n = NULL,
                d = (mu2 - mu1) / sigma1,
                sig.level = alpha,
                power = power,
                type = "two.sample",
                alternative = "two.sided")$n

n
```

```
## [1] 117.0232
```

We found that we need a minimum of 117 plots of land in order to detect an average difference of at least 40 pounds of corn per plot with a power level of 90% and a significance level of 0.05.

This means that if we conduct the experiment with 117 or more plots of land, we can be confident that the new fertilizer is indeed better than the current fertilizer.