Teo Rebull - 2721042
Jacquiline Rose Roney - 2761539
Selma Kocabiyik - 2759524

Assignment 2 - Group 84

Exercise 1

a)
Since nothing is known about the percentage of on-time flights of the airline, we can use the following formula to estimate the required sample sampled size for a proportion with a specific margin of error and confidence level (in this case 95%).

The formula:
$n = z^2 \cdot p(1 - p)/E^2$
n = $(1.96)^2 \cdot 0.5(1 - 0.5)/0.02^2$
n ≈ 2401

z = 1.96 as the confidence interval is 95% thus this is equivalent to a 1.96 z-value.
p = 0.5 as it yields the maximum product thus the maximum sample size.
E=0.02 as the estimate is within two percentage points of the true population percentage.

We get that there needs to be at least 2401 flights in order to be 95% confident that the estimate E=0.02(2%).

b)
Here, using the same formula but knowing that p = 0.90;
$n = z^2 \cdot p(1 - p)/E^2$
n = $(1.96)^2 \cdot 0.9(1 - 0.9)/0.02^2$
n ≈ 865

We get that there needs to be at least 865 flights in order to be 95% confident that the estimate E=0.02(2%).

Exercise 2

Calculating a 90% confidence interval for the difference of means is as follows:
The following information is provided:

- The study is conducted on 30 pairs of twins
- Average volume of first born twins: $\bar{x}_1 = 1124.3$
- Average volume of second born twins: $\bar{x}_2 = 1118.1$
- $s_1 = 130.5$
- $s_2 = 124.7$
- $s_d = 57.8$

Based on the information provided it can be derived that the two samples are dependent. Based on this and the information provided the confidence interval formula:

$$\left[\overline{d} - t_{n-1,\alpha/2}\frac{s_d}{\sqrt{n}}, \overline{d} + t_{n-1,\alpha/2}\frac{s_d}{\sqrt{n}}\right]$$

The components in the formula can be calculated as such

- $\bar{d} = \bar{x}_1 - \bar{x}_2$: difference of mean

  $\bar{d} = 1124.3 - 1118.1$

  $\bar{d} = 6.2$

- $t_{29,\frac{0.1}{2}} = 1.699$: calculated at 0.01 significance level and 29 degrees of freedom

These values can be substituted into the formula above to achieve the confidence interval:

$$\left[6.2 - 1.699 \times \frac{57.8}{\sqrt{30}}, 6.2 + 1.699 \times \frac{57.8}{\sqrt{30}}\right]$$

$$\left[-11.7, 24.1\right]$$

So as shown above calculating based on the formula yields the value [-11.7,24.1] for a 90% confidence interval for the difference of means of the data.

Exercise 3

We are comparing the mean brain volumes of two populations: first-borns and second-borns.

- **Null Hypothesis (H_0):** The mean brain volume of first-borns (μ_1) is equal to second-borns(μ_2). Which means there is no difference between the mean brain volumes of first and second borns.
  $H_0 : \mu_1 = \mu_2$
- **Alternative Hypothesis (H_0):** The mean brain volume of first-borns (μ_1) is not equal to second-borns(μ_2). Which means there is a difference between the mean brain volumes of first and second borns
  $H_a : \mu_1 \neq \mu_2$
- **Significance Level (α):** 0.05.
- **Sample means:** $\bar{x}_1$ = 1131.3 (first-borns), $\bar{x}_2$ = 1123.8 (second-borns).
- **Sample standard deviations:** s_1 = 129.0 (first-borns), s_2 = 127.2 (second-borns).
- **Pooled sample variance:** s_p = 128.2
- **Sample sizes:** n_1 = 25 (first-borns), n_2 = 20(second-borns)

In the exercise it is clearly indicated that first and second borns are independent samples. There are no specific assumptions in the exercise stated, whether the volumes are less than or greater than each other. So we choose a two-sample t-test statistic for independent samples with equal variances, since because our sample standard deviations (s_1 = 129.0 and s_2 = 127.2) are relatively close we assumed that σ1 = σ2.
Since our null hypothesis states that there is no difference between the means of the samples:

$$\mu_1 - \mu_2 = d_0 \,(\textit{where } d_0 \textit{ is a specified difference, often zero})$$
$$\mu_1 - \mu_2 = 0$$

Thus this is implicitly assumed to be zero under the null hypothesis. So:

$$T_{eq} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}}$$

Two-sample t-test statistic for independent samples follows a t-distribution with n1 + n2 -2 degrees of freedom under the null hypothesis.
Calculation of test statistic is:

$$degrees\ of\ freedom = 25 + 20 - 2 = 43$$
$$S_p^2 = 128.2^2 = 16435.24$$
$$t_{43}^{eq} = \frac{113.3 - 1123.8}{\sqrt{\frac{16435.24}{25} + \frac{16435.24}{20}}} = 0.1950078$$

We have a two-tailed test at a 0.05 significance level. To find the critical value, we search through table 3 from the book. In table 3, there is no degree of freedom as 43. So we take 2.021 as our critical value from 40 degrees of freedom.
Critical region:

$$t_{43}^{eq} \leq -2.021 \textit{ or } t_{43}^{eq} \geq 2.021 \,(\textit{Table } 3, \textit{df} = 40)$$
To be sure and exact we also used R to compute our critical vlaue.
$(qt(p = .05 / 2, df = 43, lower.tail = FALSE))$ which give us, critical t − value: 2.016692
$t_{43}^{eq} \leq -2.016692$ or $t_{43}^{eq} \geq 2.016692$

With these findings we cannot reject the null hypothesis, since our t is not greater or less than the critical value. So there is no statistically significant difference in the mean brain volumes of first and second borns. This also means that the data we have does not provide sufficient evidence to suggest that the mean brain volumes between these two groups are different.

Exercise 4

a)

```
shapiro.test(data$birthweight)
```

We check normality with the shapiro test, which in this case tells us the data is not normally distributed as the p-value is greater than the commonly chosen significance level of 0.05.

```
result <- t.test(data$birthweight, conf.level = 0.96)
```

Used to construct the confidence interval for the mean.

```
margin_of_error <- 100 / 2
```

We rearranged the formula to get the margin of error.

```
z_score <- qnorm(0.98)
```
We set the z-score for 98% confidence interval.

```
sample_size <- ceiling((z_score * sd(data$birthweight) / margin_of_error)^2)
```
Here, we are calculating the required sample size for estimating the mean of a population with normality assumption and with a specified margin of error in a confidence interval.
We get 821 as the sample size.
We can conclude that based on these assumptions, there needs to be a sample size of 821 to achieve the desired precision in estimating the population mean.

b)
```
t_test_result <- t.test(data$birthweight, mu = 2800, alternative = "greater")
```
With the data of the file, we get that

```
One Sample t-test

data:  data$birthweight
t = 2.2271, df = 187, p-value = 0.01357
alternative hypothesis: true mean is greater than 2800
95 percent confidence interval:
 2829.202      Inf
sample estimates:
mean of x
 2913.293
```
From this data, we can conclude that:
-Since the p-value < 0.05, the null hypothesis would be rejected.
-The true mean birthweight is greater than 2800 grams.
-The 95% confidence interval supports this by indicating a range of values above 2800.

c)
```
num_less_than_2600 <- sum(data$birthweight < 2600)
> n <- nrow(data)
> sample_proportion <- num_less_than_2600 / n
> standard_error <- sqrt(sample_proportion * (1 - sample_proportion) / n)
> p_left <- 0.25
> z <- (p_left - sample_proportion) / standard_error
> z
[1] -2.326961
> confidence_level <- 1 - pnorm(z)
> a <- pnorm(abs(z))
> a
[1] 0.9900163
> confidence_level
[1] 0.9900163
> margin_of_error <- abs(z) * standard_error
> p_right <- sample_proportion + margin_of_error
```

> p_right
[1] 0.4095745

From the calculations performed, we get that our recovered interval for p would be approximately [0.25, 0.41] at a confidence level of 99%. This means that we are 99% confident that the true proportion of newborns with a birthweight lower than 2600 grams falls within this interval.

Exercise 5

a)An estimate for the difference of mean can be calculated by calculating the means and then subtracting the means from each other. Which when executed in R is as follows:

Estimate:

```
> path <- "C:/Users/jacqu/Downloads/Alice.txt"
> data <- read.table(path, header=TRUE)
> alice_mean = mean(data$Alice)
> bob_mean = mean(data$Bob)
> D = alice_mean - bob_mean
> D
```

0.1544989

The second part of the question asks to calculate a confidence interval. A T test can be used to find the confidence interval in R at confidence interval 0.90. Which is as follows:

Confidence interval:

```
> t_test <- t.test(data$Alice-data$Bob, mu =0, conf.level = 0.90)
> t_test$conf.int
```

0.0194331 0.2895647

b) The manager claims that on average both Alice and bob both work the same amount. This claim can be checked through a t_test. The following is a systematic approach to check whether the claim is true or not.

Hypothesis:

- **Null hypothesis(H0) =** The average working hours of Alice and bob are the same $H_0 : Alice_\mu = Bob\mu$.
- **Alternative hypothesis(H1) =** The average working hours of Alice and bob are NOT the same. $H_1 : Alice_\mu \neq Bob_\mu$

It is also stated in the question that α = 0.1. Due to which the confidence interval is 90%. So a two sample t-test can be use to compare the means with confidence. To calculate the p value. The t.test function can be used with data$Alice and data$Bob as arguments along with specifying that the test is two-sided and paired is true.

```
> t_test <- t.test(data$Alice, data$Bob, alternative = "two.sided", paired = TRUE)
```

> t_test$p.value
0.06097823

As the p value is 0.06 it is less than the significance level 0.1. Due to which the null hypothesis must be rejected. So, the hypothesis that the average working hours of Alice and Bob are the same is rejected. Perhaps, there is a lack of statistical evidence to conclude that the manager's claim that the average working hours of Alice and Bob are the same or in other word H0 is rejected.

c)
To investigate Alice's claim with a suitable test we have to observe the means of Alice's and Bob's working hours.
Null Hypothesis (H0): There is no difference in the average working hours of Alice and Bob or Alice does not work more then Bob on their working hours average.
H0: μAlice - μBob ≤ 0.
Alternative Hypothesis (H1): Alice works more hours than Bob on average of their working hours.
H1: μAlice - μBob > 0.
Significance Level (α): 0.01

```
data <- read.table(file="Alice.txt", header=TRUE, sep=" ")
alice <- data$Alice
bob <- data$Bob
```

The sigma of the given two samples are unknown so we use t-distribution here. They are dependent samples so we use t-statistic for paired samples to identify the distribution under the null hypothesis. Also, since Alice claims that she works more than Bob, we use the right-tailed test.

```
result_c <- t.test(alice, bob, mu=0, alternative="greater", paired=TRUE)
```

result_c$statistic gives us 1.917771.
To make our conclusion based on the critical value:

```
critical_value <- qt(p=0.01, df=length(alice)-1, lower.tail=FALSE)
```

lower.tail=FALSE is FALSE since it is a right-tailed test. *critical_value* gives us 2.404892 value for a right tailed test at 0.01 significance level.

To make our conclusion based on the p-value:

```
p_value_c <- result_c$p.value
```

Since with the t.test we got p-value and t together, *p_value_c* gives us 0.03048911.

Conclusion:

For critical value, if the test statistic is greater than the critical value, we can reject the null hypothesis. Our t value, 1.917771, is not greater than the critical value, 2.404892.
For p-value, if p-value is less or equal to alpha(in this case it's 0.01), we can reject the null hypothesis. Our p-value from the data is approximately 0.03, which is not less than or equal to 0.01. So both cases demonstrate that we cannot reject the null hypothesis based on the test statistic, critical value and p-value that we got. Thus there is not enough statistical evidence to suggest Alice works more hours on average than Bob.

d)

Alice claims that the proportion of evenings on which she worked more than 3.8 hours is larger than the proportion of evenings during which Bob worked more than 3.8 hours.

Null Hypothesis (H0): The proportion of evenings Alice worked more than 3.8 hours is equal to or less than the proportion of evenings Bob worked more than 3.8 hours.
H0: pAlice ≤ pBob.
Alternative Hypothesis (H1): The proportion of evenings Alice worked more than 3.8 hours is greater than the proportion of evenings Bob worked more than 3.8 hours.
H1: pAlice > pBob.
Significance Level (α): 0.01

```
data <- read.table(file="Alice.txt", header=TRUE, sep=" ")
alice_work <- data$Alice > 3.8
bob_work <- data$Bob > 3.8
```

For the test statistic we choose z-statistic for two-proportion test. Since the question asked for different workings of different evenings, the samples are independent.
It is normally distributed because the sample size(n1 = n2 = 50 > 30) is large enough for Central Limit Theorem and for each sample there are greater or equal to 5 successes and failures.

```
> # Perform two-proportion z-test
> test_result <- prop.test(x = c(sum(alice_work), sum(bob_work)), n =
c(length(alice_work), length(bob_work)), alternative = "greater")
```

```
X-squared = 4.9672, df = 1, p-value = 0.01292
alternative hypothesis: greater
95 percent confidence interval:
 0.0625058 1.0000000
sample estimates:
prop 1 prop 2
  0.70   0.46
```

test_result$p.value gives us 0.01291652, which is greater than the significance level of 0.01.
For p-value, if p-value is less or equal to alpha(in this case it's 0.01), we can reject the null hypothesis. Thus we cannot reject the null hypothesis. Therefore, there is not enough statistical evidence at 0.01 significance level to support Alice's claim that she works more than 3.8 hours on a greater proportion of evenings than Bob.