Enrico Dal Pos (2708533) & Osama Obaidat(2704856) & Justus Beck(2717047) - Group 64

Assignment 3

Exercise 3.1

a) The maximum margin of error E_{max} should be equal to 0.02 ($E_{max} = 0.2$), because the estimate has to be within 2% of the true population percentage.
Further, if we were to select 100 different samples of the same size and construct the corresponding confidence intervals, 95 of the would actually contain the true population proportion.
In order to be 95% confident, our Z-score has to be between -1.96 and 1.96(according to the table in the book):
$P(-1.96 \le Z \le 1.96) = 0.95$
Because we do not have any information regarding the sample population proportion we need to find a number of flights (n), so that we are still within a margin of error of 2%. For this, we used a formula from the slides:
$n \ge (\frac{1.96}{2 \cdot E_{max}})^2$
$n \ge (\frac{1.96}{2 \cdot 0.02})^2 = 2401$

b) From the task description we know that in the previous year the population proportion (p) was equal to 90% ($p = 0.9$).
We also know that E_{max} should be 0.2, hence 95% of our values shall be between the interval $I = [0.89; 0.91]$.
Therefore, the 95% confidence interval $[p - 1.96\sqrt{\frac{p*(1-p)}{n}}; p + 1.96\sqrt{\frac{p*(1-p)}{n}}]$ shall approximate I.

For n=3457:
$0.9 - 1.96\sqrt{\frac{0.9*(1-0.9)}{3457}} \approx 0.88\overline{9} < 0.89$
$0.9 + 1.96\sqrt{\frac{0.9*(1-0.9)}{3457}} \approx 0.91 > 0.90$
For n=3458:
$0.9 - 1.96\sqrt{\frac{0.9*(1-0.9)}{3458}} \approx 0.89$
$0.9 + 1.96\sqrt{\frac{0.9*(1-0.9)}{3458}} \approx 0.90$

Answer:
In order to be 95% confident that our estimate is within 2% of the true population percentage we should survey 3458 flights ($n = 3458$).

Exercise 3.2

Given statistics:

- Sample size n = 30
- Population mean for first born: $\overline{x_1} = 1124.3$
- Population mean for second born: $\overline{x_2} = 1118.1$
- Standard deviation first borns: $s_1 = 130.5$
- Standard deviation second borns: $s_2 = 124.7$
- $s_d = 57.8$
- Since we have a 90% confidence interval we know that $\alpha = 0.1$
- Degrees of freedom = n -1 = 29

The sample is dependent because two twins form a natural pair (matched pair).

For this task we used this formula for the $(1 - \alpha)$ confidence interval from the lecture slide 7 for dependent samples:

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

We also know that $\overline{d} = \frac{\overline{x_1} - \overline{x_2}}{2}$, so

$\overline{d} = \frac{1124.3 - 1118.1}{2} = 3.1$

With the information we have about α and the degrees of freedom, we can look up the t-score in the table provided by the book:

$t_{n-1,\alpha/2} = 1.699$

Now we can solve the formula for the $(1 - \alpha)$ confidence interval with the information we just gathered:

$[3.1 - 1.699\frac{57.8}{\sqrt{30}}; 3.1 + 1.699\frac{57.8}{\sqrt{30}}]$

$[- 14.83; 21.03]$

Exercise 3.3

Exercise 3.3 Consider a study about the brain volumes (in cm^3) of 25 first-borns and 20 (unrelated) second-borns. The aim of the study was to test whether there is a difference in the population mean volumes of first- and second-born babies. The average volume of the 25 first-borns was $\bar{x}_1 = 1131.3$, and that of the 20 second-borns was $\bar{x}_2 = 1123.8$. Some more statistics that you may or may not use are: $s_1 = 129.0, s_2 = 127.2, s_p = 128.2$. Conduct the statistical test of interest and use $\alpha = 5\%$. Do we need to make specific assumptions in order to be allowed to use the test?
(See the first page of the assignment for detailed instructions about testing)

We assume that the samples of the first-borns and second-borns are independent because their sample size differs and there is no relationship between them (unrelated second-borns).

<u>Population parameters:</u>

- Sample sizes: $n_1 = 25$ and $n_2 = 20$
- Sample mean for first- and second-borns: $\overline{x_1} = 1131.3$ and $\overline{x_2} = 1123.8$
- Sample standard deviation: $s_1 = 129.0$ and $s_2 = 127.2$
- Significance level $\alpha = 5\%$
- $s_p = 128.2$

<u>Null hypothesis:</u>
H_0: $\mu_1 = \mu_2$
<u>Alternative hypothesis:</u>
H_a: $\mu_1 \neq \mu_2$

<u>Test statistic:</u>
We already established that both samples are independent. Further, both samples come from a normal population and approximate a normal distribution.
Since we established those requirements, we can use the test statistic from lecture 8:
$T_2^{eq} = \frac{\overline{X_1} - \overline{X_2} - d_0}{\sqrt{S_p^2/n_1 + S_p^2/n_2}}$, and the pooled sample variance:
$S_p^2 = \frac{(n_1 - 1)S_2^1 + (n_2 - 1)S_2^2}{n_1 + n_1 - 2}$

Because we already specified in the null hypothesis that $\mu_1 = \mu_2$ we know that the difference between the population mean is 0 ($d_0 = 0$). Now we calculate the pooled sample variance with the statistics we already obtained:

$S_p^2 = \frac{(25-1)129 + (20-1)127.2}{25+20-2} = 128.2$
Lastly, we combine everything in the formula for the test statistic:

$T_2^{eq} = \frac{1131.3-1123.8-0}{\sqrt{S_p^2/25+S_p^2/20}} = 2.21$

<u>Critical region:</u>

Since we know the significance level and the degrees of freedom $(n_1 + n_2 - 2)$ we can look up the t-score in the table of the book, which corresponds roughly to $t = 2.014$.

<u>Conclusion:</u>

Because $t < T_2^{eq}$ we fail to reject H_0. There is not sufficient evidence to warrant rejection of the claim that first- and second-borns have different brain volumes.

Exercise 3.4

a) Mean number of hours spent by Alice: 3.957444
   Mean number of hours spent by Bob: 3.802945

   The Point estimate of the difference of mean working time per evening of Alice and Bob is 0.1544989

   Using the *t.test* function a 90% confidence interval is: [0.01568705, 0.29331081]

b) Manager Claim: on average they both work the same amount of time

   $H_0 : \mu_1 - \mu_2 = 0$

   $H_1 : \mu_1 - \mu_2 \neq 0$

   Significance level is $\alpha = 0.1$

   Since Alice sd is 0.314236 and Bob sd is 0.4994043 we use a test for two independent standard deviations and we assume that their work shifts are independent.
   The p-value derived from the t.test = 0.06767

   Since $\alpha > 0.06767$
   H_0 is true so the manager's claim is correct

c) Alice Claim: I work, on average, much more than Bob.
   H_0: $\mu_1 - \mu_2 = 0$
   H_0: $\mu_1 - \mu_2 > 0$
   Significance level is $\alpha = 0.01$

Since Alice sd is 0.314236 and Bob sd is 0.4994043 we use a test for two independent standard deviations and we assume that their work shifts are independent.
The p-value derived from the t.test = 0.03384
Since α > 0.03384

H_0 cannot be rejected, this means that using this test it cannot be derived if Alice Claim is true or not.

Exercise 3.5

Exercise 3.5 Alice from the previous exercise has another concern. By contract the employees are supposed to work 3.8 hours per evening. 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.
In contrast to Exercise 3.4, we now suppose that the data given in ***Assign3.RData*** *were all collected on different evenings for Suppose again that the data were collected on 100 different evenings. We will reuse that dataset in this exercise.*

a) Based on the data find an estimate for the difference in proportion of evenings that Alice and Bob have worked more than 3.8 hours.

b) Investigate Alice's claim with a suitable test. Take significance level 1%. Motivate your choice.
(See the first page of the assignment for detailed instructions about testing)

a) The proportion of theevenings where Alice worked longer than 3.8 hours is 0.7, and for Bob is 0.46.
So given this results the estimate for the difference in proportion of evenings that Alice and Bob worked more than 3.8 hours is 0.2

Appendix

3.4

A) Script:

```
mAlice = mean(Alice)
mBob = mean(Bob)

pointEstimate = mAlice - mBob
```

```
t.test(Alice, Bob, conf.level = 0.9)

t.test(Alice, Bob, alternative = "greater")
```

Output tests:

```
data:  Alice and Bob
t = 1.8515, df = 82.542, p-value = 0.06767
alternative hypothesis: true difference in means is not equal to
0
90 percent confidence interval:
 0.01568705 0.29331081
sample estimates:
mean of x mean of y
 3.957444  3.802945

Welch Two Sample t-test

data:  Alice and Bob
t = 1.8515, df = 82.542, p-value = 0.06767
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -0.01148235  0.32048021
sample estimates:
mean of x mean of y
 3.957444  3.802945
```

3.5
a)

proportionAlice = mean(Alice > 3.8)
proportionBob = mean(Bob > 3.8)
proportionDifference = abs(PAlice - PBob)