Homework 1

Denis Ostroushko

2022-10-29

2.1

In this data we have:

- 223 Total observations for 42 unique participants in the study
- 72 observations for 13 unique E-cig smokers
- 151 observations for 29 unique non-smokers smokers

The goal of the study is to compare the level of two biomarkers between those who do and do not use E-cigarettes.

First biomarker is CEMA. CEMA is a highly reliable urinary biomarker to identify users of combustive tobacco products such as cigarettes as opposed to users of non-combustive products, medicinal nicotine, or nonusers of tobacco products.

Second, HOP1, referred to as 1-HOP in the literature, is another urinary biomarker.

We begin our comparison of groups with the two sample t-test for both measurements.

(1) Two-sample t-test

Test of average CEMA levels

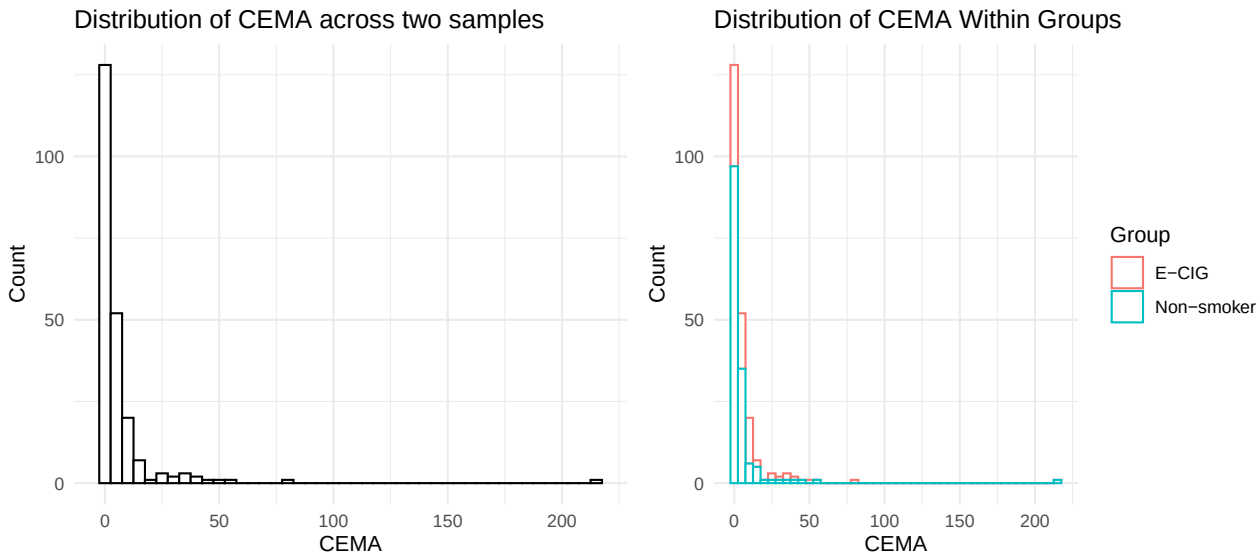
Before conducting a two-sample t-test, let's summarize the data. We need to know the average value of CEMA in two groups, standard deviation, sample size(number of measurements, not number of unique participants), and a standard error for the mean. We will calculate standard error using s.d. and sample size.

Table below presents these statistics.

Group	N	Mean CEMA	Median CEMA	Standard Deviation	Mean Standard Error
E-CIG	72	8.08	3.84	13.26	1.56
Non-smoker	151	5.39	1.56	19.03	1.55

We can see that the average CEMA measurement is higher for a group of Smokers. However, the difference in median values is even greater. Standard deviation is quite high in these samples, while sample size brings the standard error down. The two samples do not have equal variance, which violates one of t-test's assumptions.

The proportional difference in mean and median values implies that the distribution could be skewed. We can examine distribution shape before conducting the test.



While we can see that the distributions of two samples are heavily skewed, and has a lot of outliers. We would want to transform this distribution to the logarithmic scale to achieve 'normal shape', however, the t-test is robust, so we can conduct it anyway.

We define $\overline{X}_1$ = mean CEMA for smokers and $\overline{X}_2$ = mean CEMA for non-smokers.

Then, the null hypothesis is: $H_0 : \overline{X}_1 = \overline{X}_2$

And the alternative hypothesis is: $H_a : \overline{X}_1 \neq \overline{X}_2$

Results of T-test are given below:

Test results summary and interpretation:

- The average CEMA levels for smokers were 8.08 and 5.39 for non-smokers.
- Test statistic: 1.22 with 191.5496251 degrees of freedom
- Estimated difference between sample averages is 2.68. 95% Confidence interval: (-1.65 , 7.02).
- P-value: 0.2238. P-value was greater than 0.05 and confidence interval included 0.
- Concision: we do not have enough evidence to reject the null hypothesis. There is not enough evidence to suggest that the average levels of CEMA are difference between smokers and non-smokers.

Test of average HOP1 levels

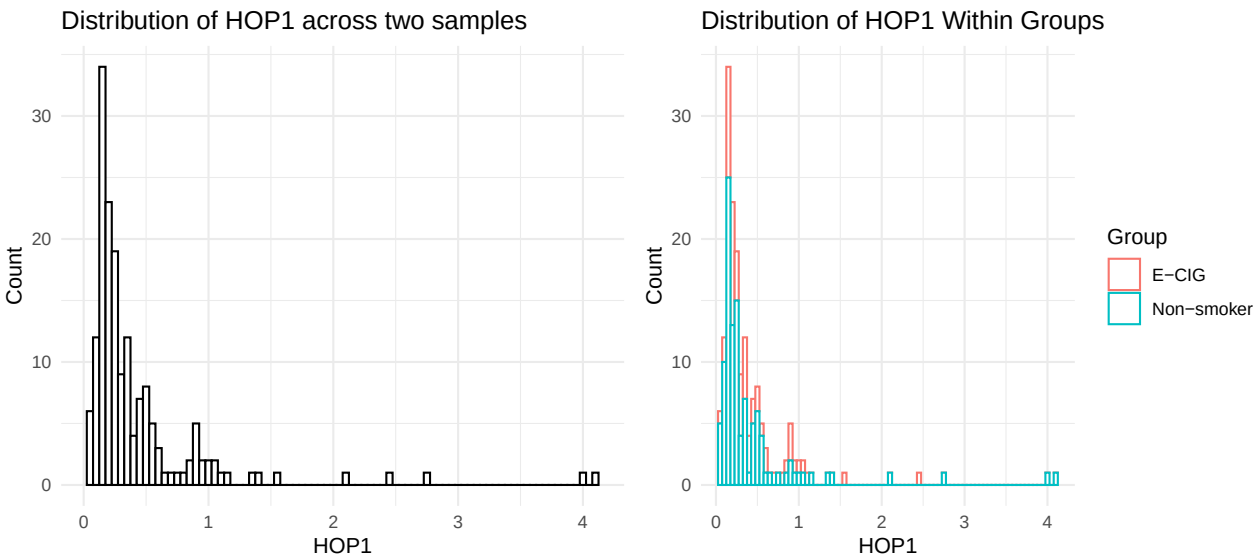
Again, we begin this section with the examination of data. Table below shows all statistics of interest.

Group	N for Analysis	N Missing	Mean HOP1	Median HOP1	Standard Deviation	Mean Standard Error
E-CIG	58	14	0.44	0.30	0.42	0.01
Non-smoker	111	40	0.43	0.24	0.63	0.01

It appears that we have some missing values of HOP1 in the data, so we present the number of data points available for analysis. Standard error was estimated using the number of data point available for analysis.

We should not expect any statistically significant results here, the average values are quite similar, while the median values are slightly further apart. The shape of the distribution may be skewed in this case too.

Once more, let's examine the shape and visual properties of distribution before the test.



While we can see that the distributions of two samples are heavily skewed, and has a lot of outliers. We would want to transform this distribution to the logarithmic scale to achieve 'normal shape', however, the t-test is robust, so we can conduct it anyway.

We define $\overline{X}_1$ = mean HOP1 for smokers and $\overline{X}_2$ = mean HOP1 for non-smokers.

Then, the null hypothesis is: $H_0 : \overline{X}_1 = \overline{X}_2$

And the alternative hypothesis is: $H_a : \overline{X}_1 \neq \overline{X}_2$

Results of T-test are given below, due to the scale of these measurements we round estimates to 5 decimal points:

Test results summary and interpretation:

- The average HOP1 levels for smokers were 0.44336 and 0.42639 for non-smokers.
- Test statistic: 0.21 with 157.9459302 degrees of freedom
- Estimated difference between sample averages is 0.01697. 95% Confidence interval: (-0.14267 , 0.17661).
- P-value: 0.833987. P-value was greater than 0.05 and confidence interval included 0.
- Concision: we do not have enough evidence to reject the null hypothesis. There is not enough evidence to suggest that the average levels of HOP1 are difference between smokers and non-smokers.

(2) Wilcoxon Test.

Test of median CEMA levels

As we saw in the previous section, the distribution of measurement is skewed and has a lot of extreme values. Perhaps, it had an impact on the t-test. I do not expect that we will see a meaningful change in result for HOP1 comparison. Summary statistics suggest that the biomarker is distributed almost identically for smokers and non-smokers. However, we saw that the mean CEMA levels were actually quite different, so applying a non-parametric test to these data can have an effect on our results.

Wilcoxon test allows us to test for difference in median values. We begin our tests by comparing median values of CEMA levels.

We define $M(X_1)$ = median CEMA levels for smokers and $M(X_2)$ = median CEMA levels for non-smokers.

Therefore, the null hypothesis is $H_0 : M(X_1) = M(X_2)$

And the alternative hypothesis is $H_a : M(X_1) \neq M(X_2)$

Test results are given below:

Test results summary and interpretation:

- Test statistic: 7110
- Estimated median CEMA levels for non-smokers was 1.56 and 3.84 for smokers.
- Estimated difference was 1.35 with a (0.55 , 2.65) 95% confidence interval
- P-value was 0.000201
- Conclusion: the p-value was well below 0.05, and the confidence interval did not include 0. Therefore, we have enough statistical evidence to reject the null hypothesis and conclude that the median levels of CEMA for smokers are higher than median levels of non-smokers, by a magnitude of two.

Wilcoxon test allowed us to see that for these skewed distributions the center of the distribution, and common values, were at a much higher levels for smokers than non-smokers. Therefore, CEMA can be used as a potentially useful predictor of smoking status, however, due to natural variance of the data and presence of extreme values and a long tail, it needs to be taken in context with other predictors and biomarkers.

Test of median HOP1 levels

We define $M(X_1)$ = median HOP1 levels for smokers and $M(X_2)$ = median HOP1 levels for non-smokers.

Therefore, the null hypothesis is $H_0 : M(X_1) = M(X_2)$

And the alternative hypothesis is $H_a : M(X_1) \neq M(X_2)$

Test results are given below:

Test results summary and interpretation:

- Test statistic: 3696
- Estimated median CEMA levels for non-smokers was NA and NA for smokers.
- Estimated difference was 0.05 with a (-0.01 , 0.11) 95% confidence interval
- P-value was 0.114609
- Conclusion: P-value was close to 0.05. The estimate was positive, and the confidence interval contained mostly positive values. However, the confidence interval did include 0. These results suggest that the median levels of HOP1 could be different between smokers and non-smokers. However, more data is needed to gather more evidence in order to reject the null hypothesis.

Overall, applying Wilcoxon tests to these two samples shows one case where non-parametric tests are more applicable than tests. He observe that we can extract more useful information using non-parametric tests when working with observational data. Such data are usually prone to variance that is outside of our control, and is different in groups that we wish to compare. Moreover, extreme values and outliers also make application of t-test difficult.

(3) Confidence intervals

In the previous section I have provided confidence intervals in the results interpretation section. In this section we validate those results by plugging estimates from the data into the formula.

First, we will have confidence intervals for comparison of average levels of CEMA

We need samples sizes, standard deviations of the two samples, and a coefficient for our desired level of confidence, which is 1.96.

The formula is given below:

Standard Error of difference = $$\sqrt{\frac{SD^2_{smokers}}{N_{smokers}} + \frac{SD^2_{non-smokers}}{N_{non-smokers}}}$$

Resulting standard error for the difference estimate is 2.2000759

The upper bound is 2.68 + 1.96 * 2.2000759 = 6.9921488

The lower bound is 2.68 - 1.96 * 2.2000759 = -1.6321488

Then we need to weight this result by

- Estimated difference between sample averages is 2.68. 95% Confidence interval: (-1.65 , 7.02).

This supports our results obtained in the first section. We saw the estimate for difference in means was not statistically significant, because the p-value was above 0.05. Confidence intervals includes 0, which is another piece of evidence in favor of null hypothesis.

Now we examine confidence interval for HOP1 difference

Similarly, we calculate confidence interval using values from the data

Resulting standard error for the difference estimate is 0.0813453

The upper bound is 0.02 + 1.96 * 0.0813453 = 0.1794368

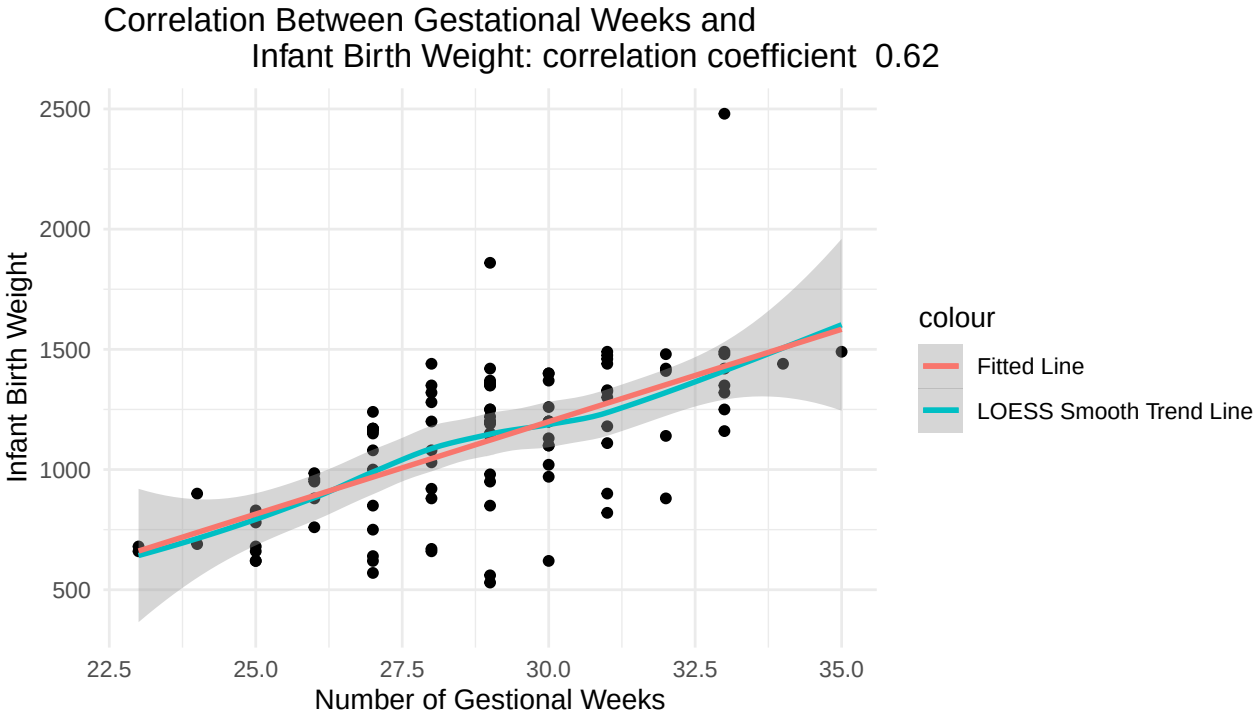
The lower bound is 0.02 - 1.96 * 0.0813453 = -0.1394368

- Estimated difference between sample averages is 0.02. 95% Confidence interval: (-0.14 , 0.18). Note that there are minor discrepancies due to rounding.

This supports our results obtained in the first section. We saw the estimate for difference in means was not statistically significant, because the p-value was above 0.05. Confidence intervals includes 0, which is another piece of evidence in favor of null hypothesis.

3.1

(1) Scatterplot



The relationship between gestational weeks and birth weight of infant have a linear relationship. Using LOESS smooth averaged trend line (in teal) we look for any non-linear curvature in the data. However, it is clear that LOESS smooth line is very close to a fitted trend line (in red). Therefore, we have visual conformation that the relationship is indeed linear.

We can see that the data is distributed in a fairly narrow cluster of point around the ends of the fitted line, and with more variance around the middle of the cluster. This is a very typical distribution of two continuous variables.

(2) Test of independence

We need to define hypothesis for any test that we perform. So,

the null hypothesis assumes no correlation between variables and is given by $H_0 : r = 0$

the alternative hypothesis is given by $H_a : r \neq 0$

The results of test for independence are given below:

Test results:

- Test statistic is 7.8132
- Estimated Pearson's correlation coefficient is 0.6195 with a (0.4817 , 0.7274) 95% confidence interval
- P-value is 0 on 98 degrees of freedom
- Conclusion: p-value is essentially 0 and the confidence interval does not include 0. Therefore, we have enough statistical evidence to reject null hypothesis and conclude that the number of gestational weeks and infant's birth weight are statistically positively correlated.

(3) Confidence Interval for Test of Independence

In this section we calculate a confidence interval for Pearson's correlation coefficient using data and formulas.

Step one is to obtain a coefficient from Fischer Transformation

z transformation = $\frac{1}{2} \times ln(\frac{1+correlationestimate}{1-correlationestimate}) = 0.7241933$

Step two is to obtain standard error

standard error = $\sqrt{\frac{1}{N-3}} = 0.1015346$

So, the raw upper bound is 0.7241933 + 1.96 * 0.1015346 = 0.9232011

So, the raw lower bound is 0.7241933 - 1.96 * 0.1015346 = 0.5251854

The last step is to transform back from Fischer Z back to the original scale

original scale lower or upper bound = $\frac{e^{2*bound}-1}{e^{2*bound}+1}$, which results in (0.4816922, 0.7274083)

Confidence interval is (0.4817 , 0.7274), it does not include 0 and gives more evidence in favor of rejection of null hypothesis

(4) Spearman's rho correlation coefficient

We had to use two separate functions to both do the Spearman's ρ test and obtain a confidence interval for it.

We keep the same null and alternative hypotheses, just this time we use a different method to calculate a correlation coefficient and perform test of independence.

Test results are below:

- Test statistic is 61987.1901
- Estimated Pearson's correlation coefficient is 0.628 with a (0.4924 , 0.7339) %95 confidence interval
- P-value is 0
- Conclusion: p-value is essentially 0 and the confidence interval does not include 0. Therefore, we have enough statistical evidence to reject null hypothesis and conclude that the number of gestational weeks and infant's birth weight are statistically positively correlated.

The difference between Pearson's coefficient and Spearman's coefficient is -0.0085, rounded to 4 decimal points.

The data did not have any notable problems with outliers or non-linear patterns, and I presume that these are the reasons why parametric and non-parametric methods produce the same results. We can also see that the lower and upper bounds of confidence intervals are similar for both methods, which gives us more confidence in obtained results.