Assignment 8 Submission deadline: None

1. **Vaccine:** It is well established that vaccines are critical in fighting and eradicating infectious diseases in the world. ***Vaccine efficacy*** is one commonly used metric to evaluate how good a vaccine is. It is defined as:

$$\text{vaccine efficacy} = \frac{\text{ARU} - \text{ARV}}{\text{ARU}} \times 100\%$$

where

- ARU = attack rate of unvaccinated people, which is basically the proportion of people with the disease among unvaccinated people.
- ARV = attack rate of vaccinated people, which is the proportion of people with the disease among vaccinated people.

We can easily see that:

$$\text{vaccine efficacy} = \left(1 - \frac{\text{ARV}}{\text{ARU}}\right) \times 100\%$$

1.1) **(5 points)** Read the following statements, and tick all the statements that are correct:

- √ **ARU is the risk of the disease in unvaccinated people**
- ◯ ARU is the odds of the disease in unvaccinated people
- √ **ARV is the risk of the disease in vaccinated people**
- ◯ ARV is the odds of the disease in vaccinated people
- √ **The ratio of ARV/ARU is the relative risk of the disease for the vaccinated people compared to the unvaccinated people**
- ◯ The ratio of ARV/ARU is the relative risk of the disease for the unvaccinated people compared to the vaccinated people
- ◯ The ratio of ARV/ARU is the odds ratio of the disease for the vaccinated people compared to the unvaccinated people
- ◯ The ratio of ARV/ARU is the odds ratio of the disease for the unvaccinated people compared to the vaccinated people
- √ **For a good vaccine, the ratio ARV/ARU should be small**

In late 2019 and early 2020, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection and the resulting coronavirus disease 2019 (Covid-19) started to spread the world and affect the lives of all people in a worldwide pandemic. In early 2020, the company Moderna announced the development of their vaccine (mRNA-1273) against SARS-CoV-2. Based on their publication in the journal *The New England Journal of Medicine* about mRNA-1273 (Baden *et al.* NEJM, 384;5), the results of 30,351 participants were recorded during the clinical trial. 15,170 of them received the placebo, and 15,181 received the mRNA-1273 vaccine. 185 participants showed symptomatic Covid-19 illness in the placebo group and 11 participants showed symptomatic Covid-19 illness in the mRNA-1273 group.

1.2) **(5 points)** Finish the following contingency table based on the information from the publication:

	mRNA-1273	**Placebo**	**Total**
COVID-19	**11**	**185**	**196**
No symptoms	**15,170**	**14,985**	**30,155**
Total	**15,181**	**15,170**	**30,351**

1.3) **(5 points)** Calculate the efficacy of the mRNA-1273 vaccine based on the data.

Solution:

$$\text{efficacy} = \left(1 - \frac{\text{ARV}}{\text{ARU}}\right) \times 100\% = \left(1 - \frac{\frac{11}{15181}}{\frac{185}{15170}}\right) \times 100\% = 94\%$$

1.4) **(15 points)** There are multiple ways of assessing whether the protection offered by the vaccine is statistically significant or not. Here, let's use the **Chi-squared test**. Write the null and alternative hypotheses, compute the expected table, the test statistic and the $p-$value.

Solution: H_0: there is no association/relationship between the disease status and the vaccine status
H_1: there is a association/relationship between the disease status and the vaccine status
Now, in order to assess the significance, we need to compute the expected

values for each cell. Remember, under the H_0 hypothesis, there is no association/relationship, meaning they are independent. Check the lecture slides if you are not sure. You should be able to calculate the expected values for each cell as follows:

	mRNA-1273	**Placebo**	**Total**
COVID-19	98.04	97.96	196
Healthy	15082.96	15072.04	30155
Total	15181	15170	30351

Once you get the expected table, it should be straightforward to get the χ^2 score and the p-value. If you decide to use Yates' correction, then they are $\chi^2 = 153.8177$ and $p = 2.54 \times 10^{-35}$. If you decide not to use the continuity correction, then they are $\chi^2 = 155.6004$ and $p = 1.04 \times 10^{-35}$. Both are correct. In addition, if you use excel to calculate the p-value, you may not have enough accuracy, in this case, getting $p = 0$ is also okay.

1.5) **(2½ points)** What assumptions do you need to check before you actually perform the **Chi-squared test**.

Solution: The sample is randomly and independently drawn, and all values in the expected table are more than 10.

2. **Effect size:** Read the paper **"Using Effect Size - or Why the P Value Is Not Enough"** by Sullivan and Feinn. Answering the following questions:

2.1) **(5 points)** During the lectures, we repeatedly used the example from the our medical school, where the authors used the **Chi-squared** test to figure out whether there is any relationship between ABO blood groups and COVID-19 susceptibility. What is the effect size if we look at the blood type A group in their test?

Solution: In the context of a Chi-squared test, the odds ratio is a measurement of the effect size. When looking at the type A blood, you should be able to get the effect size from the lecture slide, which is 1.279.

2.2) **(5 points)** Which of the following is true when we look at the blood type A group?

A. The p-value is large, and the effect size is large
B. The p-value is large, and the effect size is small
C. The p-value is small, and the effect size is large
D. The p-value is small, and the effect size is small

2.3) **(5 points)** Now, look at the first question in this assignment regarding Moderna's mRNA-1273 vaccine. What is the effect size of mRNA-1273 and is it large, medium or small?

Solution: Similarly, the effect size in this context is the odds ratio of mRNA-1273 over the placebo. It can be calculated based on the definition in the lecture slides:

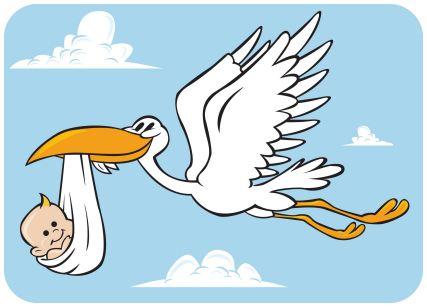
$$\frac{8 \times 21566}{162 \times 21712} = 0.04905064730660407$$

The effect size is really huge here, meaning there are much lower risk of getting COVID-19 in the mRNA-1273 group compared to the placebo group.

2.4) **(5 points)** Finally, in your own words, explain: what does effect size tell us, and why should we care about effect size?

Solution: This is an open question. As long as the answer covers that "the effect size is a measurement of the magnitude of the difference", full score can be given.

3. **Storks Deliver Babies:** The white stork (*Ciconia ciconia*) is a large bird in the stork family, *Ciconiidae*. Its plumage is manly white, with black on the bird's wings. According to European folklore, the stork is responsible for bringing babies to new parents. The legend is very ancient, but was popularised by a 19th-century Hans Christian Andersen story called "**The Storks**".



The above text is from the "White Stork" entry of Wikipedia. The lore appears in many countries in Europe, North Africa and the Middle East. Read the paper **"Storks Deliver Babies (p=0.008)"** by Robert Matthews and complete the following questions using **Table 1** from the paper:

3.1) **Reproduce the results from the paper:**

Solution: You should get the same results as the paper for (i), (ii), (iii), (iv) and (v). You might get a different linear equation if you swap x and y. You will get full points in this case.

i. **(5 points)** Write down the equation of Pearson's Correlation Coefficient r (just choose one), and calculate r between "**Storks (pairs)**" and "**Birth rate ($10^3/yr$)**" using the equation you choose.

ii. **(2½ points)** Is there a positive or negative linear relationship between "**Storks (pairs)**" and "**Birth rate ($10^3/yr$)**"?

iii. **(5 points)** To check if there is a linear relationship, you can perform a statistical test. Write down the null and alternative hypotheses, calculate the test statistic and the $p-$value. Based on the $p-$value, do you reject the null hypothesis?

iv. **(7½ points)** Perform a simple linear regression between "**Storks (pairs)**" and "**Birth rate ($10^3/yr$)**" using OLS. Write out the equations (choose one for each) for calculating the slope, the intercept and the r^2. Then calculate them based on the equations you choose.

v. **(2½ points)** Compare your reults to those in the paper. Are they the same? If not, explain why.

3.2) **(10 points)** Repeat the analyses of (i), (ii), (iii) and (iv) from **3.1** using "**Area (km^2)**" and "**Birth rate ($10^3/yr$)**" as variables.

Solution:

(i) Any equation for r is okay, and $r = 0.92$.

(ii) Positive linear relationship.

(iii) $H_0 : r = 0$ and $H_1 : r \neq 0$, the test statistic is

$$\boldsymbol{t} = \frac{r}{\sqrt{(1-r^2)/(n-2)}} = \frac{0.92}{\sqrt{(1-0.92^2)/(17-2)}} = 9.09$$

$p = 2 \times \mathbb{P}\left(\boldsymbol{t}_{15} \geqslant 9.09\right) = 1.73 \times 10^{-7}$, if you do not round the floating points and use a more precise calculation, you can also get $p = 1.36 \times 10^{-7}$. It is also correct. We reject H_0 at a significance level of $\alpha = 0.05$.

(iv) Choose the equation you like. $y = 0.0017x - 7.7755$ and $r^2 = 85.11\%$ or $y = 493.98x + 36553$ and $r^2 = 85.11\%$. Both are correct, it depends on which

one you put as x.

3.3) **(10 points)** Explore the data on your own. Then in your own words, try to explain why there seems to be a positive relationship between "**Storks (pairs)**" and "**Birth rate ($10^3/yr$)**"

Solution: Open question, you get full points as long as it is reasonable. For example, I see that there are some outliers in the data, which can cause the positive relationship.

3.4) **(5 points)** Hopefully, through the above practice, we can agree that correlation does not necessarily indicate causation. Using your own words, explain what a confounding variable (or sometimes simply called a confounder) is.

Solution: Open question, you get full points as long as it is reasonable.