HSCI 50 LAB 6: Bivariate Analysis for Continuous Variables ANSWER KEY

INSTRUCTIONS

The data set that we will be using for the entire course is resampled data from the MIT COVID-19 Beliefs, Behaviors & Norms Survey (https://covidsurvey.mit.edu/api.html). This is a multi-country, online survey that examined different COVID-19 perceptions across time, from July 6, 2020 to March 28, 2021. We will be using data from the Philippines aged 20 to 60. **Assume the data are randomly sampled.**

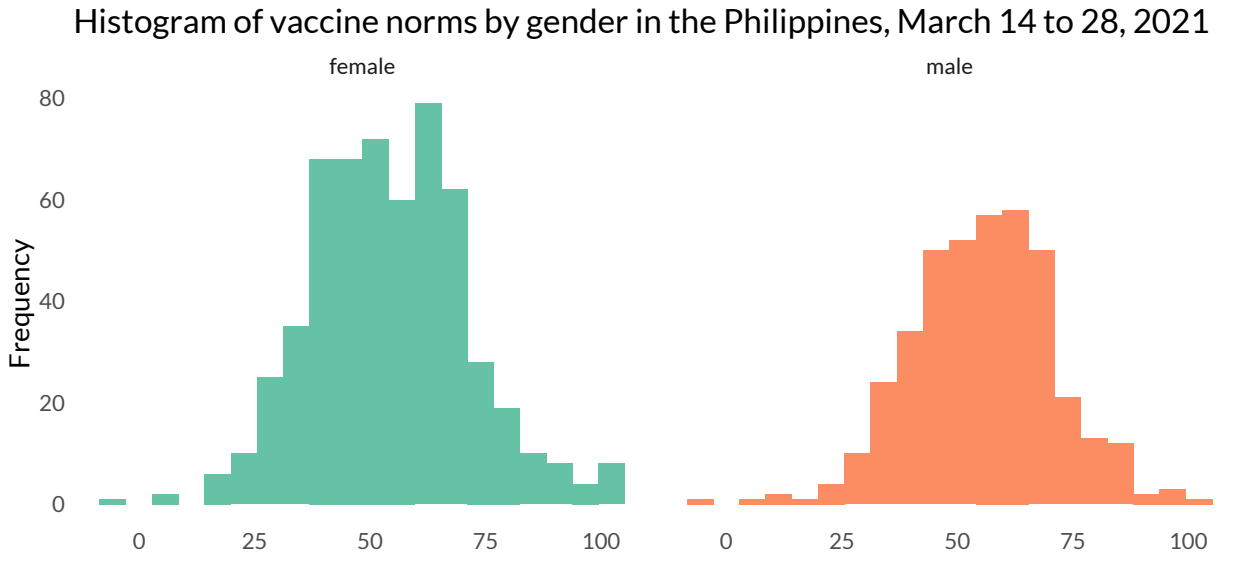
In this lab, you will learn how to conduct and interpret statistical outputs from an independent samples *t*-test, a paired samples *t*-test, and a one-way analysis of variance (ANOVA).

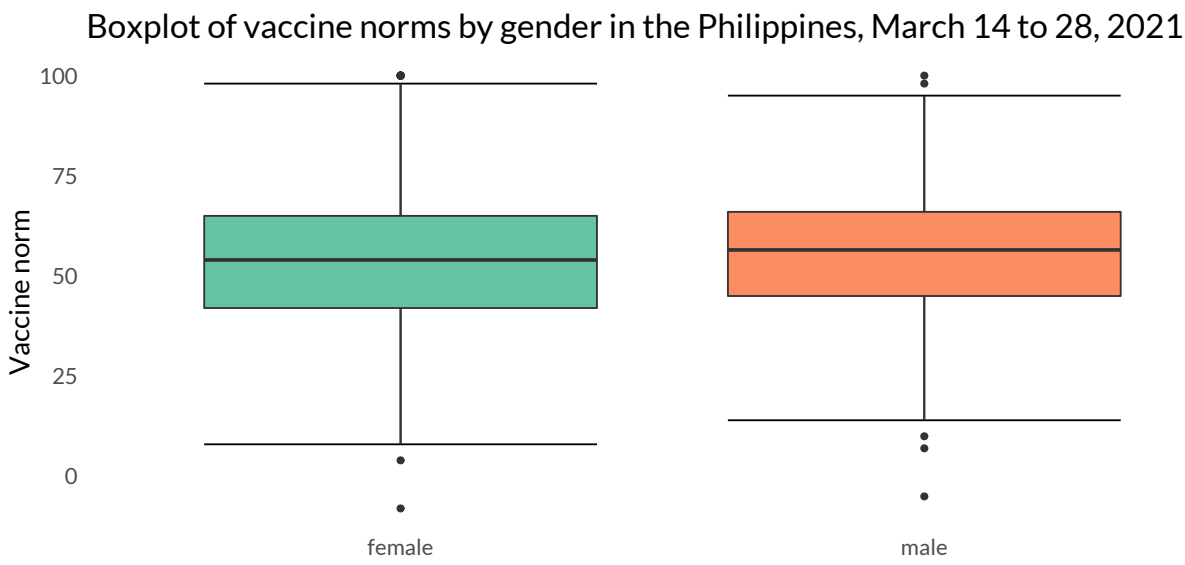
You have many options to submit this worksheet. Either you work on this by hand and scan/take a clear photo of your submission and save as PDF, or type your responses in a Word processor or PowerPoint presentation. You do not have to copy the questions again, but please number them accordingly.

Let's revisit the **vaccine norms** variable we had used in Labs 3 and 4. Individuals were asked about their perceptions of how other people will accept the COVID-19 vaccine, as a measurement of vaccine norms. They were asked, "Out of 100 people in your community, how many do you think would take a COVID-19 vaccine if it were made available?" Again, while this is technically a discrete variable, we will assume for the purposes of this class that this is a continuous variable.

PART A. Independent samples *t*-test

For this part, we want to check if there are any differences in vaccine norm by gender (male vs. female) during the last survey wave (March 14 to 28, 2021). The following are statistical outputs that will help you answer the following questions.





Note that there are a couple of outliers that are below zero. This is a data quality issue with the data in this study, and typically we would remove it before analysis

Summary statistics of vaccine norm by gender

Vaccine norm	Male	Female
Count	396	565
Mean	55.6	54.4
Standard deviation (SD)	15.5	17.1
Variance	239.3	292.2
Median	56.5	54
First Quartile (Q1)	45	42
Third Quartile (Q3)	66	65
Interquartile Range (IQR)	21	23
Mode	57	59

```
# male19 contains all the values of vaccine norms for males during the last survey wave (Wave 19)
# female19 contains all the values of vaccine norms for females during the last survey wave (Wave 19)

# Conduct independent samples t-test, two-sided, equal variances assumed
t.test(x = male19, y = female19, alternative = "two.sided", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  male19 and female19
## t = 1.1008, df = 959, p-value = 0.2713
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9286605  3.3013698
## sample estimates:
## mean of x mean of y
```

```
##  55.59343  54.40708
```

```
# Conduct independent samples t-test, two-sided, unequal variances assumed
t.test(x = male19, y = female19, alternative = "two.sided", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  male19 and female19
## t = 1.1202, df = 899.17, p-value = 0.2629
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8920958  3.2648052
## sample estimates:
## mean of x mean of y
##  55.59343  54.40708
```

1. What are the appropriate null and alternative hypothesis for this question?

The appropriate null hypothesis is that the mean vaccine norm among females is the same as males

$$H_0 : \mu_{\text{male}} = \mu_{\text{female}}; H_a : \mu_{\text{male}} \neq \mu_{\text{female}} \quad \text{or}$$

$$H_0 : \mu_{\text{male}} - \mu_{\text{female}} = 0; H_a : \mu_{\text{male}} - \mu_{\text{female}} \neq 0$$

2. What are the assumptions required to conduct an independent samples *t*-test, and how were they fulfilled?

The assumptions are:

- One variable is continuous (vaccine norm) while the other is binary (gender)
- The data are independently and identically distributed: we assume this to be the case since it was indicated in the worksheet.
- The distribution of vaccine norm is normally distributed for each category of gender: The histogram shows this to be true. There are a few outliers as indicated in the boxplot but the t-test is robust to these minor violations of the normal distribution.

3. Which t-test did you use - equal or unequal variances assumed? Explain your response.

The summary statistics tell us that the variances are not the same: for females, the variance of vaccine norms is 292.2 while for males it is 239.3. Therefore we use the independent samples t-test with unequal variances assumed.

However, if we decide to use equal variances t-test, then there's no problem as well. The conclusions are going to end up the same. Unless the variances are egregiously far apart from one another, like say 4x apart, then we do not normally expect differences in the conclusion depending on the assumption regarding the variances. There is a formal statistical test where the null hypothesis is that the variances are equal, but we do not recommend this test because a visual inspection of the variances should suffice.

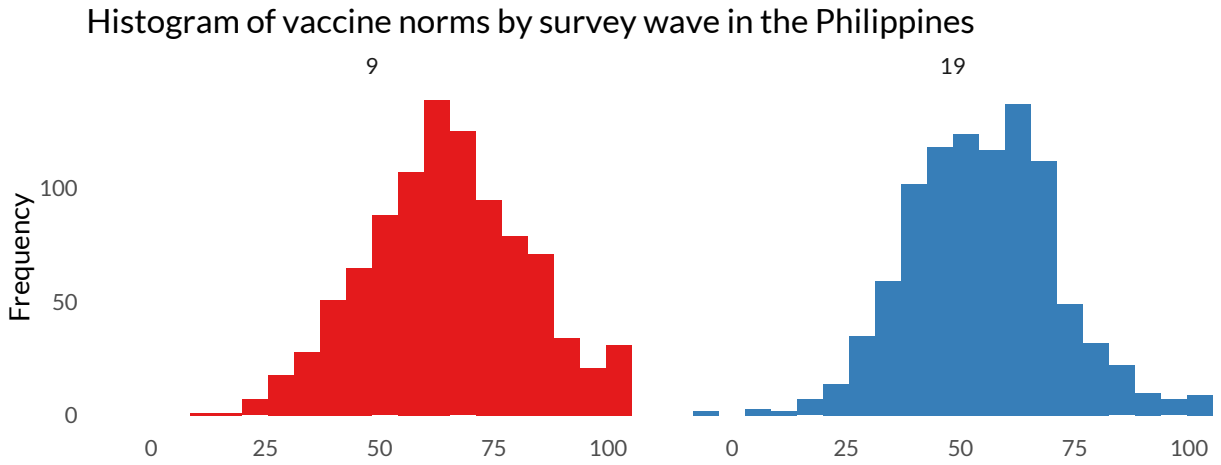
4. What can you conclude? Either use the appropriate test statistic or the 95% confidence interval (CI).

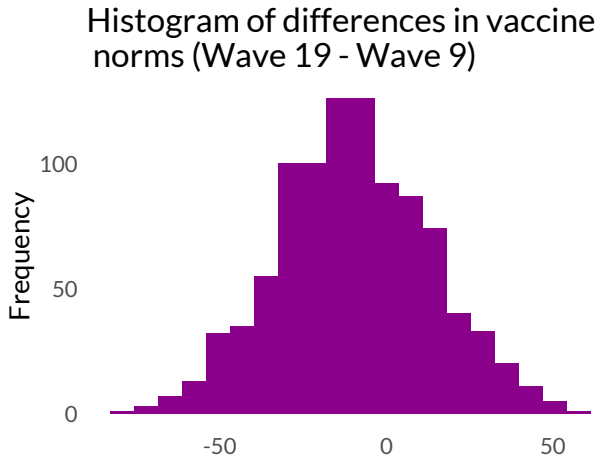
Using the test statistic, we use the *t*-statistic provided: 1.1202. The corresponding *p*-value is 0.2629. Because $p > 0.05$, we fail to reject the null hypothesis and conclude that the mean vaccine norm between males and females are the same. The mean score for males is 55.6, while for females 54.4.

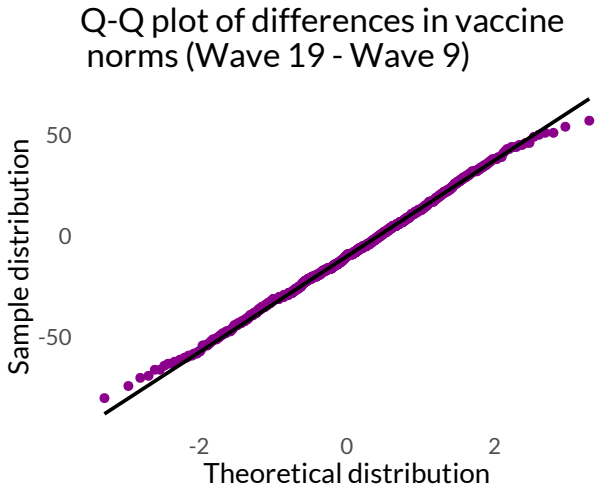
Using the confidence interval (CI), the 95% confidence interval of the difference in means is provided as well: (-0.89, 3.26). We see that the null value 0 crosses the 95% CI, therefore we also conclude that the mean vaccine norm between males and females are the same.

PART B. Paired samples *t*-test

In this part, we will compare the vaccine norms between Wave 9 of the survey (Oct 26 to Nov 10, 2020) to the last wave of the survey (Wave 19, Mar 14 to 28, 2021) and see if there was a significant change in vaccine norms over time. The following are statistical outputs that will help you answer the following questions:







Summary statistics of vaccine norm by survey wave

Vaccine norm	Wave 9	Wave 19	Difference (Wave 19 - Wave 9)
Count	961	961	961
Mean	64.7	54.9	-9.8
Standard deviation (SD)	17.3	16.4	23.1
Variance	298.0	270.5	534.9
Median	64	55	-10
First Quartile (Q1)	54	43	-26
Third Quartile (Q3)	76	66	6
Interquartile Range (IQR)	22	23	32
Mode	100	59	-9

```r
# wave9 contains all the values of vaccine norms for wave 9
# wave19 contains all the values of vaccine norms for wave 19
# data already sorted to align on the same row for the same ID number of survey participant

# Conduct paired samples t-test, two-sided
t.test(x = wave9, y = wave19, alternative = "two.sided", paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  wave9 and wave19
## t = 13.111, df = 960, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   8.317424 11.245532
## sample estimates:
## mean of the differences
##                9.781478
```

A paired samples t-test is effectively a one-sample means hypothesis test. This is because by subtracting the scores between the two categories of our binary variables, we effectively end up with one value, and we compare the mean of that value to zero. Clinical outcomes often look at pre- and post-event differences, and when the main independent variable is the time period (pre vs post), then we do a paired samples t-test. But we can also use the differences in an independent samples t-test when the main independent variable is something completely different, such as comparing a new drug vs placebo.

1. What are the appropriate null and alternative hypothesis for this question?

$$H_0 : \mu_{\text{wave 19 - wave 9}} = 0; H_a : \mu_{\text{wave 19 - wave 9}} \neq 0$$

2. What are the assumptions required to conduct a paired samples *t*-test, and how were they fulfilled?

- The samples are the same individuals measured twice
- The differences are normally distributed, as seen from the histogram and Q-Q plot

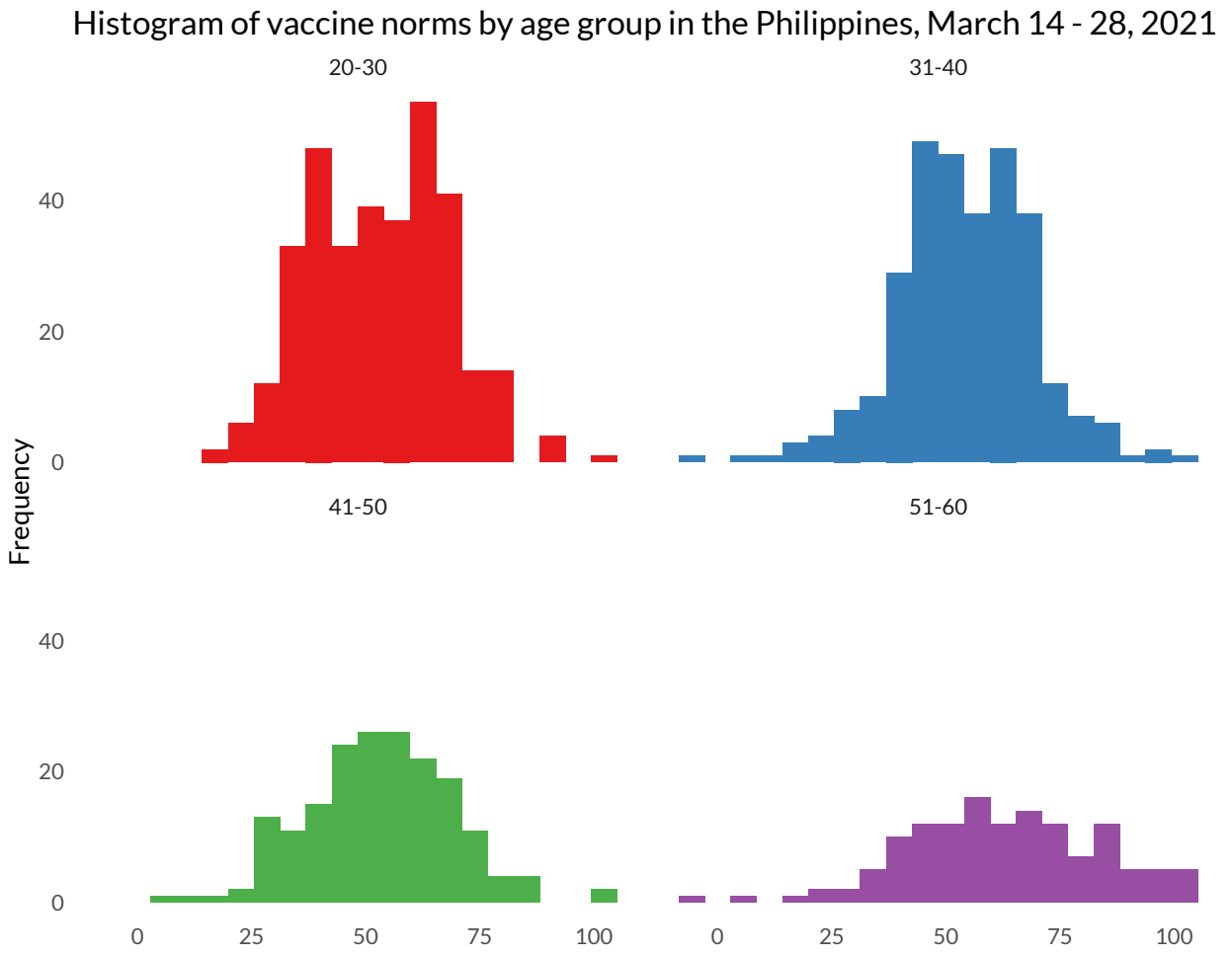
3. What can you conclude? Either use the appropriate test statistic or the 95% confidence interval (CI).

Using the test statistic, we use the t-statistic provided: 13.111. The corresponding *p*-value < 0.0005. Because $p < 0.05$, we reject the null hypothesis and conclude that the mean of the differences across survey wave is not equal to zero. The mean of the difference is 9.8, meaning the vaccine norm decreased on average by 9.8 points between Wave 9 and 19.

Using the confidence interval (CI), the 95% confidence interval of the difference in means is provided as well: (8.32, 11.25). We see that the null value 0 does not cross the 95% CI, therefore we also conclude that the mean of the differences across survey wave is not equal to than zero.

PART C. One-way analysis of variance

For this part, we want to check if there are any differences in vaccine norm by age group during the last survey wave (March 14 to 28, 2021). The following are statistical outputs that will help you answer the following questions.



Summary statistics of vaccine norm by age group

Vaccine norm	20-30 y/o	31-40 y/o	41-50 y/o	51-60 y/o
Count	339	306	182	134
Mean	53.3	54.3	53.2	62.5
SD	15.0	15.1	16.3	20.7
Variance	225.0	227.5	265.5	428.7
Median	54	54.5	54	61.5
Q1	40.5	45.0	43.0	48.2
Q3	64.5	64.0	63.5	77.2
IQR	24.0	19.0	20.5	29.0

```r
# The "vaccine_norm" variable contains the vaccine norm scores
# The "age_grp" variable contains the age group categories

# Conduct one-way ANOVA
summary(aov(vaccine_norm ~ age_grp, data = data19))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## age_grp       3   9167  3055.7   11.67 1.62e-07 ***
## Residuals   957 250503   261.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Conduct post hoc pairwise t-test with Bonferroni correction
# The results shown are the Bonferroni-inflated p-values, meaning the p-values have been inflated
# with the Bonferroni correction factor already. Compare with the original alpha.
pairwise.t.test(data19$vaccine_norm, data19$age_grp, p.adj = "bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  data19$vaccine_norm and data19$age_grp
##
##       20-30   31-40   41-50
## 31-40 1       -       -
## 41-50 1       1       -
## 51-60 2.2e-07 8.2e-06 3.5e-06
##
## P value adjustment method: bonferroni
```

1. What are the appropriate null and alternative hypothesis for this question?

$H_0 : \mu_{20-30} = \mu_{31-40} = \mu_{41-50} = \mu_{51-60}; H_a$: at least one mean is different (or age group is associated with vaccine norm)

2. What are the assumptions required to conduct a one-way analysis of variance (ANOVA), and how were they fulfilled?

- One variable is continuous (vaccine norm), while the other is categorical (age group)
- The data are independently and identically distributed: we assume this to be the case since it was indicated in the worksheet
- The variances are fairly homogenous as shown in the summary statistics table (Recall the rule of thumb that as long as the highest standard deviation (19.3) is not twice as large or larger than the smallest stnadard deviation (16.6))
- The vaccine norm values are normally distributed within each age category as shown in the histograms

3. What can you conclude from the one-way ANOVA AND the post-hoc pairwise *t*-test with Bonferroni correction (if applicable)? Mention the relevant test statistic and *p*-values, as well as the mean values of vaccine norm in each group

We use the *F*-statistic provided: 11.67. The corresponding *p*-value < 0.0005. Because $p < 0.05$, we reject the null hypothesis and conclude that at least one of the age groups had a different mean vaccine norm than the other age groups. We can also say that age group is associated with vaccine norm.

A one-way ANOVA tells us that the variables are associated, but it does not tell us which pairwise means are different. This is the purpose of the pairwise independent samples t-test. When we want to check which mean is higher or lower, we check the point estimates of the mean.

The summary statistics table shows that the mean values of vaccine norm per age group are as follows:

- 21-30 years old: 53.3
- 31-40 years old: 54.3
- 41-50 years old: 53.2
- 51-60 years old: 62.5

Looking at the post-hoc pairwise t-test with Bonferroni correction, the Bonferroni-adjusted *p*-values that were significant were 51 vs. 60 vs. each of the other age groups.

Synthesizing all the results, vaccine norm is associated with age group, and the mean vaccine norm was similar across ages 21 to 50, while significantly higher among 51-60 years old.

END OF LAB