Name:____ **Date:**____________________

PSTAT 5A: Discussion Worksheet 05

Summer Session A 2023, with *Ethan P. Marzban*

1. Ecologists near the remote village of *Statsville* in the nation of *Gauchonia* would like to determine whether the village's drinking water has the same amount of bacteria as drinking water found near the capital city. To that end, they take 15 samples of water near *Statsville* and 20 samples of water near the capital, and record the bacteria levels (in cells per liter) in each sample. The ecologists' results are summarized below:

	Sample Mean	**Sample Std. Dev.**
Statsville	75	12.3
Capital City	63	18.7

Let "Population 1" refer to the drinking water in *Statsville* and "Population 2" refer to the drinking water near the capital city. Additionally, assume all independence and normality conditions are satisfied; also use a 5% level of significance and a two-sided alternative wherever necessary.

(a) Define the parameters of interest, μ_1 and μ_2.

Solution: Let μ_1 denote the average bacteria level in water near *Statsville* and let μ_2 denote the average bacteria level in water near the capital city.

(b) State the null and alternative hypotheses.

Solution:

$$\left[\begin{array}{ll} H_0: & \mu_1 = \mu_2 \\ H_A: & \mu_1 \neq \mu_2 \end{array}\right. = \left[\begin{array}{ll} H_0: & \mu_2 - \mu_1 = 0 \\ H_A: & \mu_2 - \mu_1 \neq 0 \end{array}\right.$$

(c) Compute the observed value of the test statistic.

Solution:

$$\mathrm{ts} = \frac{\overline{y} - \overline{x}}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}} = \frac{63 - 75}{\sqrt{\frac{12.3^2}{15} + \frac{18.7^2}{20}}} \approx -2.29$$

(d) Assuming the null is correct, what is the approximate distribution of the sampling distribution? Be sure to include any/all relevant parameters.

Solution: Since we are told to assume all relevant independence and normality conditions hold, we know that, under the null, the test statistic will approximately be distributed according to a t−distribution with degrees of freedom given by the Sattherthwaite equation:

$$\begin{aligned} \mathrm{df} &= \mathrm{round}\left\{ \frac{\left[\left(\frac{s_X^2}{n_1} \right) + \left(\frac{s_Y^2}{n_2} \right) \right]^2}{\frac{\left(\frac{s_X^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left(\frac{s_Y^2}{n_2} \right)^2}{n_2 - 1}} \right\} \\ &= \mathrm{round}\left\{ \frac{\left[\left(\frac{12.3^2}{15} \right) + \left(\frac{18.7^2}{20} \right) \right]^2}{\frac{\left(\frac{12.3^2}{15} \right)^2}{15 - 1} + \frac{\left(\frac{18.7^2}{20} \right)^2}{20 - 1}} \right\} = \mathrm{round}\left\{ 32.54532 \right\} = 33 \end{aligned}$$

Hence, we have

$$\mathrm{TS} \overset{H_0}{\sim} t_{33}$$

(e) What is the p−value of the observed test statistic? (As a reminder, you may need to use Python for this part.)

Solution: Recall that (as was mentioned in Lecture) we cannot usually use our t−table to find p−values, as our t−table gives us *percentiles* (whereas p−values are *areas*). Hence, we must use Python: after importing `scipy.stats` as sps we run

```
2 * sps.t.cdf(-2.29, 33)
```

which gives us a p−value of around 0.03.

(f) What is the critical value of the test?

Solution: Here we could use either our table, or Python. If we use our table, we find the critical value to be 2.03. If we use Python, we would run

```
-sps.t.ppf(0.025, 33)
```

which gives us a critical value of around 2.0345. Either way, our critical value is approximately 2.03.

(g) Now, carry out the test. Be sure to phrase your conclusions in terms of the context of the problem.

Solution: We reject the null when the absolute value of the test statistic exceeds the critical value, or equivalently when the p−value is smaller than the level of significance. Hence, since $|-2.29| = 2.29 > 2.03$ and $p = 0.03 < 0.05$, we reject the null:

> At a 5% level of significance, there was sufficient evidence to reject the null that the drinking water in *Statsville* has the same average amount of bacteria as the water found near the capital city, in favor of the alternative that the two locations have different average bacterial concentrations.

(h) Re-do the test, now using the alternative that the capital city has higher bacterial levels in its water than *Statsville*. Again use a 5% level of significance.

Solution: Our hypotheses are now

$$\left[\begin{array}{ll} H_0: & \mu_1 = \mu_2 \\ H_A: & \mu_1 < \mu_2 \end{array}\right. = \left[\begin{array}{ll} H_0: & \mu_2 - \mu_1 = 0 \\ H_A: & \mu_2 - \mu_1 > 0 \end{array}\right.$$

If we use the same test statistic

$$\mathrm{TS} = \frac{\overline{Y} - \overline{X}}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}}$$

then we would reject the null in favor of the alternative for large positive values of TS. That is, we reject when $\mathrm{TS} > c$ meaning our critical value is now computed as

```
sps.t.ppf(1 - 0.05, 33)
```

which is around 1.69 (which is also in agreement with the value obtained from the table). Now, $\mathrm{ts} = -2.29 \not> 1.69$, meaning we fail to reject:

> At a 5% level of significance, there was insufficient evidence to reject the null that the drinking water in *Statsville* has the same average amount of bacteria as the water found near the capital city, in favor of the alternative that the capital city has a higher bacterial concentration than *Statsville*.

Intuitively, this makes sense- the data does not support the alternative that the capital city has a higher concentration of bacteria than *Statsville*.

(i) Re-do the test, now using the alternative that *Statsville* has higher bacterial levels in its water than the capital city. Again use a 5% level of significance.

Solution: Our hypotheses are now

$$\left[\begin{array}{ll} H_0: & \mu_1 = \mu_2 \\ H_A: & \mu_1 > \mu_2 \end{array}\right. = \left[\begin{array}{ll} H_0: & \mu_2 - \mu_1 = 0 \\ H_A: & \mu_2 - \mu_1 < 0 \end{array}\right.$$

If we use the same test statistic

$$\mathrm{TS} = \frac{\overline{Y} - \overline{X}}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}}$$

then we would reject the null in favor of the alternative for large negative values of TS. That is, we reject when TS $< c$ meaning our critical value is now computed as

```
sps.t.ppf(0.05, 33)
```

which is around -1.69 (which is also in agreement with the value obtained from the table; **note the sign flip if we use the table!**). Now, ts $= -2.29 < -1.69$, meaning we once again reject:

> At a 5% level of significance, there was sufficient evidence to reject the null that the drinking water in *Statsville* has the same average amount of bacteria as the water found near the capital city, in favor of the alternative that *Statsville* has a higher bacterial concentration than the capital city.

2. The following ANOVA table has some entries missing. Fill in the missing entries, and provide justification as to how you found those missing values. You may need Python for certain entries.

	DF	**Sum Sq.**	**Mean Sq.**	**F−value**	$\mathbb{P}(> F)$
Btw. Groups	12	18	1.5	<???>	0.02976
Residuals	120	<???>	<???>		

Solution: To fill in this table, we must work from right to left. That is, let's first find the value of the F−statistic. We know that, under the null, the F−statistic follows an $F_{k-1,\ n-k-1}$ distribution where n is the total number of observations and k is the number of groups. We also know that these are the two entries in the "DF" column, meaning we know

$$F \overset{H_0}{\sim} F_{12,\ 120}$$

Name:____________________________________ **Date:**__________________

This allows us to find the $F-$statistic using Python: after importing `scipy.stats` as `sps` we run

```
sps.f.ppf(1 - 0.02976, 12, 120)
```

which gives us a value for the $F-$statistic of around 2.

Now, we know that the $F-$statistic is found as the ratio of mean-squared quantities. Thus, we have

$$2 = \frac{1.5}{\mathrm{MS_E}} \implies \mathrm{MS_E} = \frac{1.5}{2} = 0.75$$

Additionally, we have

$$\mathrm{MS_E} = \frac{\mathrm{SS_E}}{n-k-1}$$

which, plugging in values we know, yields

$$0.75 = \frac{\mathrm{SS_E}}{120} \implies \mathrm{SS_E} = 0.75 \cdot 120 = 90$$

Hence, our final filled-out table is

	DF	**Sum Sq.**	**Mean Sq.**	**$F-$value**	**$\mathbb{P}(> F)$**
Btw. Groups	12	18	1.5	2	0.02976
Residuals	120	90	0.75		