

Assignment #2**Due: Wednesday January 26, 2022 by 5 PM ET**

Please sign your name to the **appropriate** space below. Remember that you are permitted to receive (and provide) authorized assistance but must acknowledge it if you do.

I received assistance on this assignment and/or discussed it with fellow classmates or a tutor.

I received no assistance on this assignment and/or did not discuss it with anyone other than Professor Miller.

Brooke Wheeler

Brooke Wheeler
Homework_2

Part 1: Variance

Instructions: In these problems we are going to focus on variance. Provide complete justification for each problem. You may type your work or include work done by hand. Please make sure that all work is neat, organized, and shows your complete understanding of the problem.

1. Let's start by looking at the sample variance. Generally, we think of the sample variance as $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$. Show that $\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} [\sum_{i=1}^n X_i^2 - n\bar{X}^2]$.

$$\begin{aligned}
 & \textcircled{1} \quad \frac{1}{n-1} \sum (x_i - \bar{x})^2 \\
 & \quad \frac{1}{n-1} \left[\sum (x_i - \bar{x})(x_i - \bar{x}) \right] \\
 & \quad \frac{1}{n-1} \left[\sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \right] \\
 & \quad \frac{1}{n-1} \left[\sum x_i^2 - \sum 2x_i\bar{x} + \sum \bar{x}^2 \right] \quad \begin{array}{l} \rightarrow \sum x_i = \\ \sum x_1 + \sum x_2 + \dots \\ = n\bar{x} \end{array} \\
 & \quad \frac{1}{n-1} \sum x_i^2 - \frac{1}{n-1} \sum 2x_i\bar{x} + \frac{1}{n-1} \sum \bar{x}^2 \\
 & \quad \frac{1}{n-1} \sum x_i^2 - \frac{2\bar{x}}{n-1} \sum x_i + \frac{\bar{x}^2}{n-1} \\
 & \quad \downarrow \quad - \frac{2\bar{x} n\bar{x}}{n-1} + \frac{\bar{x}^2}{n-1} \\
 & \quad \downarrow \quad - \frac{n\bar{x}^2}{n-1} \\
 & \quad \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)
 \end{aligned}$$

2. In class we reviewed the sampling distribution of \bar{X} . Show that $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$. (Hint: Use the properties of variance we reviewed in class.)

$$\begin{aligned}
 & \textcircled{2} \quad \text{Var}(\bar{x}) = \frac{\sigma^2}{n} \quad \bar{x} = \frac{1}{n} \sum x_i \\
 & \quad \text{Var} \left[\frac{1}{n} \sum x_i \right] \quad \sum \text{var}(x_1) + \text{var}(x_2) + \text{var}(x_n) \\
 & \quad \frac{1}{n^2} \text{Var} \sum x_i \quad \begin{array}{l} \sigma^2 \quad \sigma^2 \quad \sigma^2 \\ n\sigma^2 \end{array} \\
 & \quad \frac{1}{n^2} n\sigma^2 \\
 & \quad \frac{\sigma^2}{n}
 \end{aligned}$$

3. What does it mean for s^2 to be an unbiased estimator of σ^2 ?

- a. For s^2 to be an unbiased estimator of σ^2 means that the sampling distribution of the sample variance has a mean that is equal to the population variance. This means the expected value of the sample variance is equal to the population variance.

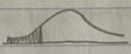
Part 2: Working with T Tests and F Tests

Instructions: For each question below, you must show all work. Use R to find p-values and critical values. If α is not given and it is difficult to make conclusions, use $\alpha = 0.05$. For two-sample problems, start with the F-test to decide which t-test to use. You may type your work or include work done by hand.

4. A researcher wants to examine if women lie about how much they weigh. She suspects that women will tend to give a weight less than their true weight. She looks at a large population of women whose average weight is 145 pounds. She then takes a random sample of 56 women from this population and asks them their weight (after giving them the opportunity to weigh themselves beforehand). She finds that the 56 women report an average weight of 142.2 pounds with standard deviation 9.4 pounds.
- a. Does the researcher's data provide sufficient evidence to support her claim? Be specific.
- b. Construct a 97% confidence interval for the true mean reported weight of the women.

```
13 qt(.05,55, lower.tail = TRUE)
14 pt(-1.673, 9.4)
15
16 ^ ` ` `
    [1] -1.673034
    [1] 0.06360903
```

④ $\mu = 145$ $\bar{x} = 142.2$ $\alpha = .05$
 $n = 56$
 $s = 9.4$

a) $H_0: \mu = 145$ 
 $H_a: \mu < 145$
 $df = 55$

critical $t = q(.05, 55) = -1.673$
 $p\text{-value} = p(-1.673, 9.4) = .064$

b) 97% Confidence Interval
 $\bar{x} \pm (t_{crit}) \frac{s}{\sqrt{n}}$
 $142.2 \pm (-1.673) \frac{9.4}{\sqrt{56}}$
 $142.2 \pm (-2.102)$
 $[144.302, 140.098]$

Since the p-value is .064 we fail to reject the null hypothesis. Meaning we do not have statistically sufficient evidence to say that women will give a weight less than their true weight.

5. There are two manufacturing processes, old and new, that produce the same product. The defect rate has been measured for 20 days for the old process, and for 14 days for the new process, resulting in the following sample summaries.

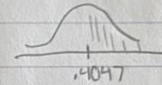
	OLD	NEW
Average defect rate	4.7	2.3
Standard deviation	6.8	5.0

The firm is interested in switching to the new process only if it can be demonstrated convincingly that the new process reduces the defect rate. Is there significant evidence of that? Use $\alpha = 5\%$ and assume that the collected data represent two random samples from Normal distributions. Use the method of testing that is appropriate for this situation.

115) $H_0: \mu_1 = \mu_2$	new 1	old 2	$\alpha = .05$
$H_a: \mu_1 < \mu_2$	$\bar{x} = 2.3$	$\bar{x} = 4.7$	
	$s = 5$	$s = 6.8$	
	$n = 14$	$n = 20$	

F test $H_0: \sigma_1^2 = \sigma_2^2$ $df_1 = 14 - 1 = 13$
 $H_a: \sigma_1^2 \neq \sigma_2^2$ $df_2 = 20 - 1 = 19$

critical F = $qf(.05, 13, 19, \text{lower.tail} = \text{TRUE})$
 $= .4047$



p-value = $pf(.4047, 13, 19, \text{lower.tail} = \text{FALSE})$
 $= .95$

we fail to reject H_0 , assume the variances are equal

cont.

$$\begin{aligned} \text{pooled variance} &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} \\ &= \frac{(13)(25) + (19)(46.24)}{32} \\ &= 37.611 \end{aligned}$$

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{(2.3 - 4.7) - 0}{\sqrt{37.611 \left(\frac{1}{14} + \frac{1}{20} \right)}} = \frac{-2.4}{2.137} = -1.123$$

p-value = $pt(-1.123, 32, \text{lower.tail} = \text{TRUE})$
 $= .135$

fail to reject the null hypothesis

```

17 # 5
18 ```{r}
19 qf(.05, 13,19, lower.tail=TRUE)
20 pf(.4047, 13,19, lower.tail = FALSE)
21
22 ((13*25) + (19*46.24))/32
23
24 ((1/14) + (1/20)) * 37.611
25 sqrt(4.567)
26 2.4 / 2.137054
27
28 pt(-1.123, 32, lower.tail = TRUE)
29 ```

```

```

[1] 0.4047157
[1] 0.950007
[1] 37.61125
[1] 4.56705
[1] 2.137054
[1] 1.123041
[1] 0.1348963

```

- a. At a p-value of .135 we fail to reject the null hypothesis that the two means are equal. Meaning that we do not have statistically significant evidence to prove that there is a difference between the old and new manufacturing processes.

Part 3: Mini-Project

6. Data on 522 recent home sales is posted on Canvas with this assignment (*"HOME_SALES.csv"*). The following variables are included.

Column	Variable
1	Identification number 1–522
2	Sales price of residence (×\$1000 dollars)
3	Finished area of residence (square feet)
4	Total number of bedrooms in residence
5	Total number of bathrooms in residence
6	Air conditioning: present or absent
7	Number of cars that garage will hold
8	Swimming pool: present or absent
9	Year property was originally constructed
10	Quality of construction: high, medium, or low
11	Architectural style. Three styles are coded as 1, 2, and 3
12	Lot size (square feet)
13	Location near a highway: yes or no

Note: In order to group a quantitative variable (such as Sales price) based on a categorical variable (such as Air conditioning) we can enter something like the following into R:

```
t.test(SALES_PRICE[AIR_CONDITIONER=="YES"], SALES_PRICE[AIR_CONDITIONER=="NO"])
```


- a. The sales price depends on the air conditioner in the house.

```

38 # null: mean sale price for AC = mean sale price NO AC
39 # Alt.: mean sale prices are not equal
40
41 #test for equal variances
42 var.test(SALES_PRICE[AIR_CONDITIONER=="YES"],SALES_PRICE[AIR_CONDITIONER=="NO"])
43 # since p-value is less than .05 we reject the null hypothesis that the variances
  are equal. This means we will use a t.test assuming the variances are not equal.
44
45 t.test(x= SALES_PRICE[AIR_CONDITIONER=="YES"], y=
  SALES_PRICE[AIR_CONDITIONER=="NO"], alternative = "two.sided", mu=0, var.equal =
  FALSE, conf.level = .95)
46 # since p-value is less than .05 we reject the null hypothesis that the mean sale
  prices for houses with AC and without AC are equal.
47 ```

```

F test to compare two variances

data: SALES_PRICE[AIR_CONDITIONER == "YES"] and SALES_PRICE[AIR_CONDITIONER == "NO"]

F = 3.749, num df = 433, denom df = 87, p-value = 1.017e-11

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

2.654162 5.108169

sample estimates:

ratio of variances

3.748998

Welch Two Sample t-test

data: SALES_PRICE[AIR_CONDITIONER == "YES"] and SALES_PRICE[AIR_CONDITIONER == "NO"]

t = 10.304, df = 241.5, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

85.91273 126.52249

sample estimates:

mean of x mean of y

295.8006 189.5830

b. On the average, homes with an air conditioner are more expensive.

```
50 # r}
51 # null: mean sale price with AC = mean sale price without AC
52 # alt.: mean sale price with AC > mean sale price without AC
53
54 t.test(x= SALES_PRICE[AIR_CONDITIONER=="YES"], y=
SALES_PRICE[AIR_CONDITIONER=="NO"], alternative = "greater", mu=0, var.equal = TRUE,
conf.level = .95)
55 # since p-value is less than .05 we reject the null hypothesis that the mean sale
prices of housing with AC are the same as those without AC. This means we have
statistically significant evidence to say homes that have an AC are more expensive
than those without.
56 #
```

Two Sample t-test

```
data: SALES_PRICE[AIR_CONDITIONER == "YES"] and SALES_PRICE[AIR_CONDITIONER ==
"NO"]
t = 6.8735, df = 520, p-value = 9.001e-12
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 80.75386      Inf
sample estimates:
mean of x mean of y
295.8006 189.5830
```

c. On the average, homes with an air conditioner are larger.

```

60 # null: mean square feet home with AC = mean square feet with homes without AC
61 # alt.: mean square feet home with AC > mean square feet with homes without AC
62
63 #test for equal variances
64 var.test(FINISHED_AREA[AIR_CONDITIONER=="YES"],FINISHED_AREA[AIR_CONDITIONER=="NO"])
65 # since p-value is less than .05 we reject the null hypothesis that the variances
   are equal. This means we will use a t.test assuming the variances are not equal.
66
67 t.test(x= FINISHED_AREA[AIR_CONDITIONER=="YES"], y=
   FINISHED_AREA[AIR_CONDITIONER=="NO"], alternative = "greater", mu=0, var.equal =
   FALSE, conf.level = .95)
68 # since p-value is less than .05 we reject the null hypothesis homes with an AC are
   larger than homes without an AC.
69 ```

```

F test to compare two variances

```

data: FINISHED_AREA[AIR_CONDITIONER == "YES"] and FINISHED_AREA[AIR_CONDITIONER
== "NO"]
F = 1.8536, num df = 433, denom df = 87, p-value = 0.00062
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.312303 2.525642
sample estimates:
ratio of variances
      1.853624

```

Welch Two Sample t-test

```

data: FINISHED_AREA[AIR_CONDITIONER == "YES"] and FINISHED_AREA[AIR_CONDITIONER
== "NO"]
t = 7.756, df = 160.14, p-value = 4.817e-13
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 399.9772      Inf
sample estimates:
mean of x mean of y
2346.339 1837.909

```

d. The sales price depends on the proximity to a highway.

```

72 {r}
73 # null: mean sale price close to hw = mean sale price not close to hw
74 # alt.: mean sale price close to hw is not equal mean sale price not close to hw
75
76 #test for equal variances
77 var.test(SALES_PRICE[HIGHWAY=="YES"],SALES_PRICE[HIGHWAY=="NO"])
78 # since p-value is greater than .05 we fail to reject the null hypothesis that the
  variances are equal. This means we will use a t.test assuming the variances are
  equal.
79
80 t.test(x= SALES_PRICE[HIGHWAY=="YES"], y= SALES_PRICE[HIGHWAY=="NO"], alternative =
  "two.sided", mu=0, var.equal = TRUE, conf.level = .95)
81 # since p-value is greater than .05 we fail to reject the null hypothesis that the
  mean sale prices for houses close to the highway are the same to homes not close to
  the highway.
82 ```

```

F test to compare two variances

```

data: SALES_PRICE[HIGHWAY == "YES"] and SALES_PRICE[HIGHWAY == "NO"]
F = 0.37562, num df = 10, denom df = 510, p-value = 0.08588
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1811525 1.1622045
sample estimates:
ratio of variances
 0.3756203

```

Two Sample t-test

```

data: SALES_PRICE[HIGHWAY == "YES"] and SALES_PRICE[HIGHWAY == "NO"]
t = -1.1638, df = 520, p-value = 0.2451
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -131.44001  33.64546
sample estimates:
mean of x mean of y
 230.0273  278.9245

```

e. On the average, homes are cheaper when they are close to a highway.

```

86 # null: mean sale price close to hw = mean sale price not close to hw
87 # alt.: mean sale price close to hw < mean sale price not close to hw
88
89
90 t.test(x= SALES_PRICE[HIGHWAY=="YES"], y= SALES_PRICE[HIGHWAY=="NO"], alternative =
"less", mu=0, var.equal = TRUE, conf.level = .95)
91 # since p-value is greater than .05 we fail to reject the null hypothesis that the
mean sale prices for houses close to the highway are the same to homes not close to
the highway.
92 ^ ```

```

Two Sample t-test

```

data: SALES_PRICE[HIGHWAY == "YES"] and SALES_PRICE[HIGHWAY == "NO"]
t = -1.1638, df = 520, p-value = 0.1225
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
 -Inf 20.33688
sample estimates:
mean of x mean of y
230.0273 278.9245

```

f. On the average, homes are cheaper when they are far from a highway.

```

94 f. On the average, homes are cheaper when they are far from a highway.
95 ^ ```{r}
96 # null: mean sale price close to hw = mean sale price not close to hw
97 # alt.: mean sale price close to hw > mean sale price not close to hw
98
99
100 t.test(x= SALES_PRICE[HIGHWAY=="YES"], y= SALES_PRICE[HIGHWAY=="NO"], alternative =
"greater", mu=0, var.equal = TRUE, conf.level = .95)
101 # since p-value is greater than .05 we fail to reject the null hypothesis that the
mean sale prices for houses close to the highway are the same to homes not close to
the highway.
102 ^ ```

```

Two Sample t-test

```

data: SALES_PRICE[HIGHWAY == "YES"] and SALES_PRICE[HIGHWAY == "NO"]
t = -1.1638, df = 520, p-value = 0.8775
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -118.1314      Inf
sample estimates:
mean of x mean of y
230.0273 278.9245

```