STAT 614 - HW 3

Due: Thursday, October 1, 2020 in Blackboard (go to the Homework folder under the Homework/Classwork content area) by 11:59 pm.

Instructions: Please type your solutions to these FIVE problems and upload the document as a pdf file in Blackboard. There is only one file to submit for this assignment.

Notes: This homework continues our discussions of sampling distributions and statistical inferences, especially using the t-procedures. You will need some concepts that will be discussed in class next week (hence the two-week due date). This is also the last homework before our first exam!!!

1. Triceps skinfold thickness is an upper arm measurement that has been used as a proxy measure of body fat. The table below gives the mean and standard deviations of tricep skinfold thickness (in cm) for two populations of adult males, those with chronic airflow limitation (such as COPD, a type of obstructive lung disease) and those without any airflow limitation. A study comparing tricep skinfold thickness is being planned in these populations using the respective sample sizes (n), also given in the last column of the table.

Population	μ	σ	n
Chronic airflow limitation	0.92	0.4	32
No airflow limitation	1.35	0.5	40

a. Consider a random sample $y1$, $y2$, $y3$, ... , yn from the chronic airflow limitation population with mean μ and standard deviation σ as given in the table. What is the **standard deviation of the sample mean**, $\bar{y}$? (Note, this is often called the "standard error" of the mean, especially when an estimate for σ is used.)

SE = Standard error

$$SE(EST) = \frac{\sigma}{\sqrt{n}} = \frac{0.4}{\sqrt{32}} = \frac{0.4}{5.66} = 0.07$$

b. Assume the Central Limit Theorem is applicable. What does it suggest about potential values of the sample mean, $\bar{y}$, the researchers can expect in their study?

The Central Limit Theorem indicates that if we take samples of n from any populatuion, and we know the population has mean μ and a finite standard deviation σ, then, for large n, we get the Y-bar, which is sample mean. And the distribution of Y-bar is approximately normal, and Y-bar's mean is equal to μ, and the standard error of Y-bar is σ over the square root of n.

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ or } \bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

2. A human resources manager for a large company takes a random sample of 50 employees from the company database. She calculates the mean time that they have been employed. She records this value and then repeats the process: She takes another random sample of 50 names and calculates the mean employment time. After she has done this 1000 times, she makes a histogram of the mean employment times. Is this histogram a display of the population distribution, the distribution of a sample, or the sampling distribution of mean?

According to the Central Limit Theorem, the histogram is the display of the smapling distribution of mean.

3. In most software, the default p-values are computed based on a 2-sided alternative hypothesis (HA: μ ≠ H0: μ0). However, we may want to use a 1-sided alternative in some problems. Hence, we need to be able to compute the correct 1-sided p-values from the reported 2-sided version. Sketch a graph for each of the following to demonstrate that each is the correct procedure. (You can take a photo of your sketch to include in your HW.)

(a) For H_A: $\mu > \mu_0$:

i. if test statistic $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ is positive, (1-sided p-value) = 0.5*(2-sided p-value).

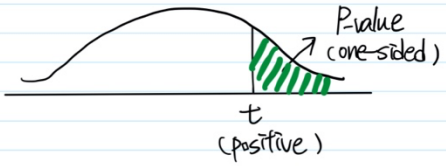
ii. if test statistic $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ is negative, (1-sided p-value) = 1 – 0.5*(2-sided p-value).

(a) For H_A $\mu > \mu_0$

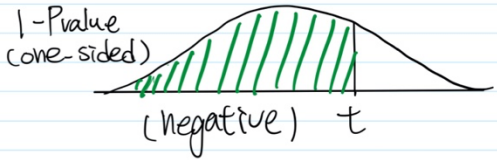
Given a normal distribution, $P_{two\text{-}sided} = P(>t) + P(<-t)$ (t is positive)

P-value is the probability of a true null. Therefore, it represents the area to the right of the normal distribution. Because of symmetry of a normal distribution, we have the following

(i) $P(>t) = 0.5 * P_{two\text{-}sided}$



(ii) $P(<-t) = 1 - 0.5 * P_{two\text{-}sided}$



(b) For H_A: $\mu < \mu_0$:

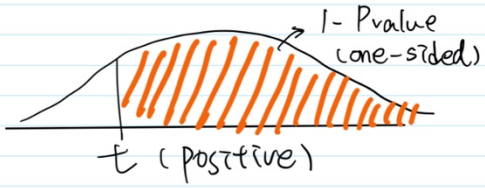
i. if test statistic $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ is positive, (1-sided p-value) = 1 – 0.5*(2-sided p-value).

ii. if test statistic $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ is negative, (1-sided p-value) = 0.5*(2-sided p-value).

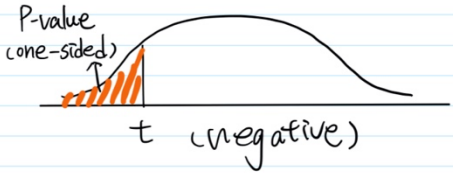
(b) For H_A $\mu < \mu_0$

P-value is still defined as the probability of a true null but it represents the area to the left of the t statistic. Similarly, we have the following

(i) $P(>t) = 1 - 0.5 * P_{two\text{-}sided}$



(ii) $P(<t) = 0.5 * P_{two\text{-}sided}$



4. Suppose the following statement is made in the conclusions section of a paper: "A comparison of breathing capacities of individuals in households with low nitrogen dioxide levels and individuals in households with high nitrogen dioxide levels indicated that there is no difference in means (two- sided p-value = 0.24)."

a) Give a reasonable null and alternative hypothesis that was being tested in this scenario. Carefully define the population parameters of interest being tested.

- μ = population mean
- H0 (Null hypothesis): $\mu 1 = \mu 2$
- HA (Alternative hypothesis): $\mu 1 \neq \mu 2$
- $\mu 1$: the population mean of breathing capacities of individuals in households with **low** nitrogen dioxide levels
- $\mu 2$: the population mean of breathing capacities of individuals in households with **high** nitrogen dioxide levels

b) Why is this statement an inaccurate summary of the hypothesis test?
With a p-value of 0.24, we fail to reject the null hypothesis because alpha = 0.05, which is smaller than 0.24. However, it does not mean the null hypothesis is true, it simply means the data (the sample used in this example) are consistent with that null hypothesis.

c) Re-write the statement so that it is properly summarizing the results of the hypothesis test.

In this case, p-value = 0.24, which means that given the data, there is 24% chance that there is not difference of means in breathing capacities of individuals in households with low nitrogen dioxide levels and individuals in households with high nitrogen dioxide levels.

5. **Use the lead.csv data set from HW 1.** Revisit HW 1 for a description of this data set and study. For this problem you will compare the Wechsler full-scale IQ scores (the variable IQF) between the different lead exposure groups, denoted by the GROUP variable.

Noted: GROUPS 1 and 2, those children with elevated blood-lead levels > 40 mug/ml and those with lower levels, < 40 mug/ml, respectively.

GROUPS	1	2
blood-lead levels	> 40 mug/ml	< 40 mug/ml

a. Compute the mean, standard deviation, standard error, and 95% confidence interval for the population mean IQ score for **each** lead exposure group, separately. Summarize each confidence interval.

There are mean, standard deviation, and standard error with each group below.

```r
# mean and standard deviation of group 1
favstats(~iqf, data = leadGroup1)
```

data.frame
1 x 9

data.frame
1 x 9

min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>	missing <int>
46	80	88	93.75	114	88.02174	12.20654	46	0

1 row

```r
# mean and standard deviation of group 2
favstats(~iqf, data = leadGroup2)
```

min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>	missing <int>
50	85	94	101	141	92.88462	15.34451	78	0

1 row

```r
# standard error of group 1, 2
library(plotrix)
std.error(leadGroup1$iqf, na.rm = TRUE)
std.error(leadGroup2$iqf, na.rm = TRUE)
```

```
[1] 1.799756
[1] 1.737424
```

The 95% confidence interval defines a range of values that we can be 95% certain contains the **population mean**. Firstly, the result shows that IQ score is 84.4 ~91.6, which has 95% of the intervals would contain the population mean in group 1.

```r
# 95% confidence interval of group 1
t.test(~iqf,data = leadGroup1, conf.level = 0.95)
```

```
        One Sample t-test

data:  iqf
t = 48.908, df = 45, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 84.39685 91.64663
sample estimates:
mean of x
 88.02174
```

Secondly, the result shows that IQ score is 89.4 ~ 96.3, which has 95% of the intervals would contain the population mean in group 2.

```r
# 95% confidence interval of group 2
t.test(~iqf,data = leadGroup2, conf.level = 0.95)
```

```
        One Sample t-test

data:  iqf
t = 53.461, df = 77, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 89.42496 96.34427
sample estimates:
mean of x
 92.88462
```

b. Researchers were interested in assessing the difference in the mean IQ score between the two exposure group populations. Give the estimate mean difference, the standard error, and the 95% confidence interval for the **difference** in population mean IQ scores. Summarize the confidence interval.

```{r}
# group1 and group2 two sample t-test
tout1and2_01 <- t.test(iqf~GROUP, mu = 0, data = bloodlead, conf.level = 0.95) # iqf by each Group
tout1and2_01
```

```
        Welch Two Sample t-test

data:  iqf by GROUP
t = -1.9439, df = 111.41, p-value = 0.05442
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9.81966762  0.09391511
sample estimates:
mean in group 1 mean in group 2
       88.02174        92.88462
```

```{r}
tout1and2_01$stderr # standard error
```

```
[1] 2.501552
```

- H0: mean IQ score GROUP1 = GROUP2
- Ha: mean IQ score GROUP1 $\neq$ GROUP2

According to Welch's two sample t-test, the code shows that 95 % confidence interval for the difference in mean running from negative 9.82 to 0.09, we also see the sample group means of 88.02 and 92.88, as well we find the sample mean difference (Est) is 92.88 - 88.02 = 4.86.

p-value = 0.05442, which is geather than 0.05, so we fail to reject the null hypothesis. Thus, blood-lead levels are not affect their IQ.

c. Researchers hypothesized that the exposed group (GROUP = 1) would **have a lower population mean IQ score** than the control group (GROUP = 2). Set up and conduct a statistical hypothesis test to address the research hypothesis. Carefully state the null and alternative hypotheses to be tested. Give the parameter of interest, the estimate of this parameter, the standard error of the estimate, the test statistic, and the p-value. Summarize the results of the test.

Parameter of interest: difference in population mean
GROUP = 1: population mean in the exposed group
GROUP = 2: population mean in the control group.
H0: mean IQ score GROUP1 = GROUP2 (mu2 – mu1 = 0)
Ha: mean IQ score GROUP2 > GROUP1 (mu2 – mu1 > 0)

- T-procedure below:

```r
# group1 and group2 two sample t-test
tout1and2_02 <- t.test(iqf~GROUP, mu = 0, data = bloodlead,
                       conf.level = 0.95, alternative = "greater") # iqf by each Group
tout1and2_02
```

```
        Welch Two Sample t-test

data:  iqf by GROUP
t = -1.9439, df = 111.41, p-value = 0.9728
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -9.012065       Inf
sample estimates:
mean in group 1 mean in group 2
       88.02174        92.88462
```

```r
tout1and2_02$stderr
```

```
[1] 2.501552
```

In summary, we fail to reject the null hypothesis for the following three reasons. 1) the p-value is greater than 0.05, which is the signufiance level under 95% C.I.; 2) the t-stats (-1.9439) is greater than -1.96, which is the critical value. 3) the calculated difference in mean is 4.86, which falls into the confidence interval (-9.0121 to positive infinity). Therefore, we don't have evidence that the exposed group (GROUP = 1) would have a lower population mean IQ score than the control group (GROUP = 2).

References

Confidence Intervals & Hypothesis Testing: STAT 200. (n.d.). Retrieved September 29, 2020, from https://online.stat.psu.edu/stat200/lesson/6/6.6

The Role of Probability. (n.d.). Retrieved September 20, 2020, from https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/BS704_Probability12.html

2 Sample t-test Calculator. https://www.usablestats.com/calcs/2samplet