Name:____________________________ Date:______________________

**HW #1 Ch13: Test of Significance: Comparing Two Population Means and Proportions**

1. What is a test statistics? Explain it in context to this chapter.

2. Does watching Youtube cooking shows help you with your culinary skills. A random sample of 15 amateur chefs were chosen and their culinary skills are measured on a scale of 1 to 100 before and after having them watch a series of Youtube cooking shows. The individual results are as follows:

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Before* | 32 | 27 | 24 | 36 | 41 | 54 | 40 | 18 | 29 | 50 | 43 | 36 | 24 | 35 | 44 |
| *After* | 38 | 31 | 27 | 35 | 39 | 60 | 39 | 16 | 26 | 49 | 48 | 34 | 29 | 39 | 50 |

a) Is this a One Sample or Two Sample Z or T-test? (paired?) Explain

b) Write the Hypothesis in context

c) Perform a test on the 5% significance level on whether if watching Youtube cooking shows help improve your culinary skills. Remember to state the conditions required. Are both populations approximately normal? Explain

d) Interpret and State your conclusion

e) Is there a correlation between culinary skills before and after watching Youtube cooking shows. Ie: Can culinary skills after watching Youtube shows be predicted from culinary skills before watching Youtube shows? Explain:

3. Pat wants to compare the cost of one- and two-bedroom apartments in the area of her college campus. She collects data for a random sample of 10 advertisements of each type. Here are the rents for the two-bedroom apartments (in dollars per month):

   595, 500, 580, 650, 675, 675, 750, 500, 495, 670

   Here are the rents for the one-bedroom apartments:

   500, 650, 600, 505, 450, 550, 515, 495, 650, 395

a) Is this a One Sample or Two sample T-test? Explain

b) Is there a significant difference between the rent for a one bedroom vs two bedroom apartment? Perform a significance test to answer this question.

c) Find a 95% confidence interval for the difference in rental cost. Does your C.I. provide evidence “for” or “against” the Null Hypothesis? Explain:

4. A medical research was conducted to investigate whether taking a daily low dose of aspirin reduces the chance of developing colon cancer. In a study of 1000 volunteers, subjects were randomly assigned to one of two groups: Half were assigned to an experimental group that took a small daily dose of aspirin and the other half were given a daily placebo. At the end of 10 years, 15 people in the experimental group developed colon cancer and 28 from the placebo group develop colon cancer. Does this data provide convincing statistical evidence that a small daily dose of aspirin help reduce the chance of developing colon cancer?
   a) Is this an One sample, Two sample, T-test, or Proportion test? Explain

   b) Write the Null and Alternative Hypothesis in context.

   c) Perform a 5% level significance test. State your conditions.

   d) Calculate and Interpret P-value in the context of this study. State your results.

   e) What is a Type I and Type II error. Which is more serious?

   f) Can the results from this test be generalized to all populations? Explain:

`

5. Patients with heart-attack symptoms arrive at an emergency room either by ambulance or self-transportation provided by themselves, family or friends. When a patient arrives at the emergency room, the time or arrival is recorded. The time when the patient's diagnostic treatment begins is also recorded. An administrator of a large hospital wanted to determine whether the mean wait time (time between arrival and diagnostic treatment) for patients with heart attack symptoms differ according to the mode of transportation. A random sample of 150 patients with heart attack symptoms who had reported to the emergency room was selected. For each patient, the mode of transportation and wait time were recorded.

| T*ransportation* | *n* | *Mean Wait time* | *Sample std dev* |
|---|---|---|---|
| *Ambulance* | 80 | 35.22 min | 15.23 min |
| *Self* | 65 | 52.55 min | 15.83 min |

i) Is this a One tail or two tail test? State your Null and Alternative Hypothesis in context

ii) Perform an appropriate significance test to see if there is a difference in mean wait time between patient arriving by ambulance vs self transportation. Follow the Inference Toolbox

iii) Calculate and Interpret P-value in the context of this study. Are your results statistically significant? State your results.

**Q#6: AP 2015**

A researcher conducted a medical study to investigate whether taking a low-dose aspirin reduces the chance of developing colon cancer. As part of the study, 1,000 adult volunteers were randomly assigned to one of two groups. Half of the volunteers were assigned to the experimental group that took a low-dose aspirin each day, and the other half were assigned to the control group that took a placebo each day. At the end of six years, 15 of the people who took the low-dose aspirin had developed colon cancer and 26 of the people who took the placebo had developed colon cancer. At the significance level $\alpha = 0.05$, do the data provide convincing statistical evidence that taking a low-dose aspirin each day would reduce the chance of developing colon cancer among all people similar to the volunteers?

**Q#7) AP 2014**

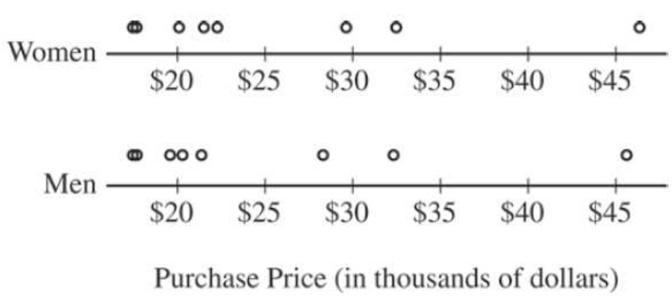
A researcher conducted a study to investigate whether local car dealers tend to charge women more than men for the same car model. Using information from the county tax collector's records, the researcher randomly selected one man and one woman from among everyone who had purchased the same model of an identically equipped car from the same dealer. The process was repeated for a total of 8 randomly selected car models.

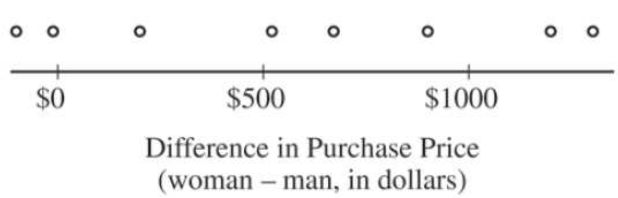
The purchase prices and the differences (woman – man) are shown in the table below. Summary statistics are also shown.

| Car model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Women | $20,100 | $17,400 | $22,300 | $32,500 | $17,710 | $21,500 | $29,600 | $46,300 |
| Men | $19,580 | $17,500 | $21,400 | $32,300 | $17,720 | $20,300 | $28,300 | $45,630 |
| Difference | $520 | –$100 | $900 | $200 | –$10 | $1,200 | $1,300 | $670 |

| | Mean | Standard Deviation |
|---|---|---|
| Women | $25,926.25 | $9,846.61 |
| Men | $25,341.25 | $9,728.60 |
| Difference | $585.00 | $530.71 |

Dotplots of the data and the differences are shown below.




Do the data provide convincing evidence that, on average, women pay more than men in the county for the same car model?

**Q#8) AP 2013**

Psychologists interested in the relationship between meditation and health conducted a study with a random sample of 28 men who live in a large retirement community. Of the men in the sample, 11 reported that they participate in daily meditation and 17 reported that they do not participate in daily meditation.

The researchers wanted to perform a hypothesis test of

$$H_0 : p_m - p_c = 0$$
$$H_a : p_m - p_c < 0,$$

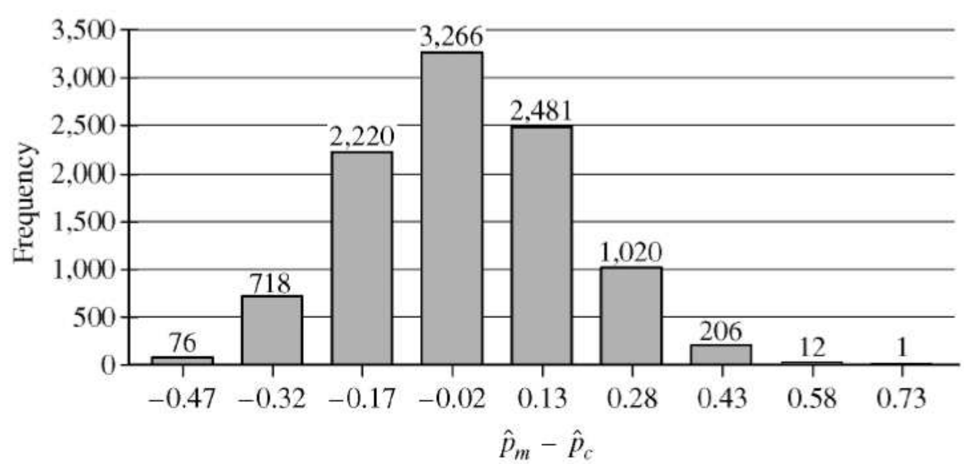
where $p_m$ is the proportion of men with high blood pressure among all the men in the retirement community who participate in daily meditation and $p_c$ is the proportion of men with high blood pressure among all the men in the retirement community who do not participate in daily meditation.

(a) If the study were to provide significant evidence against $H_0$ in favor of $H_a$, would it be reasonable for the psychologists to conclude that daily meditation causes a reduction in blood pressure for men in the retirement community? Explain why or why not.

The psychologists found that of the 11 men in the study who participate in daily meditation, 0 had high blood pressure. Of the 17 men who do not participate in daily meditation, 8 had high blood pressure.

(b) Let $\hat{p}_m$ represent the proportion of men with high blood pressure among those in a random sample of 11 who meditate daily, and let $\hat{p}_c$ represent the proportion of men with high blood pressure among those in a random sample of 17 who do not meditate daily. Why is it not reasonable to use a normal approximation for the sampling distribution of $\hat{p}_m - \hat{p}_c$?

Although a normal approximation cannot be used, it is possible to simulate the distribution of $\hat{p}_m - \hat{p}_c$. Under the assumption that the null hypothesis is true, 10,000 values of $\hat{p}_m - \hat{p}_c$ were simulated. The histogram below shows the results of the simulation.




(c) Based on the results of the simulation, what can be concluded about the relationship between blood pressure and meditation among men in the retirement community?

ANSwer:

Q#6) AP 2015

**Intent of Question**

The primary goal of this question was to assess a student's ability to identify, set up, perform, and interpret the results of an appropriate hypothesis test to address a particular question. More specific goals were to assess a student's ability to (1) state appropriate hypotheses; (2) identify the appropriate statistical test procedure and check appropriate conditions for inference; (3) calculate the appropriate test statistic and $p$-value; and (4) draw an appropriate conclusion, with justification, in the context of the study.

**Solution**

Step 1: States a correct pair of hypotheses.

Let $p_{\text{asp}}$ represent the population proportion of adults similar to those in the study who would have developed colon cancer within the six years of the study if they had taken a low-dose aspirin each day. Similarly, let $p_{\text{plac}}$ represent the population proportion of adults similar to those in the study who would have developed colon cancer within the six years of the study if they had taken a placebo each day.

The hypotheses to be tested are $H_0 : p_{\text{asp}} = p_{\text{plac}}$ versus $H_a : p_{\text{asp}} < p_{\text{plac}}$ or equivalently, $H_0 : p_{\text{asp}} - p_{\text{plac}} = 0$ versus $H_a : p_{\text{asp}} - p_{\text{plac}} < 0$.

Step 2: Identifies a correct test procedure (by name or by formula) and checks appropriate conditions.

The appropriate procedure is a two-sample $z$-test for comparing proportions.

Because this is a randomized experiment, the first condition is that the volunteers were randomly assigned to one treatment group or the other. The condition is satisfied because we are told that the volunteers were randomly assigned to take a low-dose aspirin or a placebo.

The second condition is that the sample sizes are large, relative to the proportions involved. The condition is satisfied because all sample counts are large enough; that is, 15 with colon cancer in aspirin group, 26 with colon cancer in placebo group, $500 - 15 = 485$ cancer-free in aspirin group, and $500 - 26 = 474$ cancer-free in placebo group.

Step 3: Calculates the appropriate test statistic and $p$-value.

The sample proportions who developed colon cancer are $\hat{p}_{\text{asp}} = \frac{15}{500} = 0.030$ and $\hat{p}_{\text{plac}} = \frac{26}{500} = 0.052$.

The combined sample proportion who developed colon cancer is $\hat{p}_{\text{combined}} = \frac{15 + 26}{500 + 500} = 0.041$.

The test statistic is $z = \dfrac{0.030 - 0.052}{\sqrt{0.041(1 - 0.041)\left(\frac{1}{500} + \frac{1}{500}\right)}} \approx -1.75$ ($-1.7542$ from calculator).

The $p$-value is $P(Z \leq -1.75) = 0.0401$ (0.0397 from calculator), where $Z$ has a standard normal distribution.

Step 4: States a correct conclusion in the context of the study, using the result of the statistical test.

Because the $p$-value is less than the given significance level of $\alpha = 0.05$, we reject the null hypothesis. The data provide convincing statistical evidence that the proportion of all adults similar to the volunteers who would develop colon cancer if they had taken a low-dose aspirin every day is less than the proportion of all adults similar to the volunteers who would develop colon cancer if they had not taken a low-dose aspirin every day.

## Scoring

Steps 1, 2, 3, and 4 are scored as essentially correct (E), partially correct (P), or incorrect (I).

**Step 1** is scored as follows:

Essentially correct (E) if the response identifies correct parameters *AND* both hypotheses are labeled and state the correct relationship between the parameters.

Partially correct (P) if the response identifies correct parameters *OR* states correct relationships, but not both.

Incorrect (I) if the response does not meet the criteria for E or P.

*Note*: Either defining the parameters in context, or simply using common parameter notation with subscripts clearly relevant to the context, such as $p_{\text{asp}}$ and $p_{\text{plac}}$, is sufficient.

**Step 2** is scored as follows:

Essentially correct (E) if the response correctly includes the following three components:

1. Identifies the correct test procedure (by name or by formula).
2. Notes that the use of random assignment satisfies the randomness condition.
3. Checks for approximate normality of the test statistic by citing that all four counts are larger than some standard criterion such as 5 or 10.

Partially correct (P) if the response correctly includes only two of the three components.

Incorrect (I) if the response correctly includes at most one of the three components.

*Notes*:

- For the randomness component, it is (minimally) acceptable to say "random assignment — check" but not acceptable to say "random — check" or "SRS — check." The important concept here is that it is random assignment, and not random sampling, that is required. If the response implies that the study used a random sample, the randomness component is not satisfied, regardless of whether random assignment is correctly addressed.
- The normality check may use the expected counts under the null hypothesis in place of observed counts.

**Step 3** is scored as follows:

Essentially correct (E) if the response correctly calculates both the test statistic and a $p$-value that is consistent with the stated alternative hypothesis.

Partially correct (P) if the response correctly calculates the test statistic but not the $p$-value;

*OR*

if the response calculates the test statistic incorrectly but then calculates the correct $p$-value for the computed test statistic;

*OR*

if the response reports the correct $p$-value but no calculations or test statistic are shown.

Incorrect (I) if the response fails to meet the criteria for E or P.

*Note*: The $p$-value is considered correct if it is consistent with the alternative stated in the response and the calculated test statistic, even if those are incorrect.

**Step 4** is scored as follows:

Essentially correct (E) if the response provides a correct conclusion in context, with justification based on linkage between the $p$-value and the given $\alpha = 0.05$.

Partially correct (P) if the response provides a correct conclusion, with linkage to the $p$-value, but not in context;

*OR*

if the response provides a correct conclusion in context, but without justification based on linkage to the $p$-value.

Incorrect (I) if the response does not meet the criteria for E or P.

*Notes*:

- The conclusion must be related to the alternative hypothesis.
- If the $p$-value is incorrect, then step 4 is scored as E if the response includes proper linkage and a conclusion in context consistent with that $p$-value.
- If the $p$-value is less than 0.05, wording that states or implies that the alternative hypothesis is *proven* lowers the score one level (that is, from E to P or P to I) in step 4.
- If the $p$-value is incorrect and greater than 0.05, wording that states or implies that the null hypothesis is *accepted* lowers the score one level (that is, from E to P or P to I) in step 4.

Q#7 AP 2014

**Intent of Question**

The primary goal of this question was to assess students' ability to identify, set up, perform, and interpret the results of an appropriate hypothesis test to address a particular question. More specific goals were to assess students' ability to (1) state appropriate hypotheses; (2) identify the appropriate statistical test procedure and check appropriate conditions for inference; (3) calculate the appropriate test statistic and *p*-value; and (4) draw an appropriate conclusion, with justification, in the context of the study.

**Solution**

Step 1: States a correct pair of hypotheses.

Let $\mu_{\text{diff}}$ represent the population mean difference in purchase price (woman – man) for identically equipped cars of the same model, sold to both men and women by the same dealer, in the county.

The hypotheses to be tested are $\text{H}_0 : \mu_{\text{diff}} = 0$ versus $\text{H}_\text{a} : \mu_{\text{diff}} > 0$.

Step 2: Identifies a correct test procedure (by name or by formula) and checks appropriate conditions.

The appropriate procedure is a paired *t*-test.

The conditions for the paired *t*-test are:

1. The sample is randomly selected from the population.
2. The population of price differences (woman – man) is normally distributed, or the sample size is large.

The first condition is met because the car models and the individuals were randomly selected. The sample size ($n = 8$) is not large, so we need to investigate whether it is reasonable to assume that the population of price differences is normally distributed. The dotplot of sample price differences reveals a fairly symmetric distribution, so we will consider the second condition to be met.

Step 3: Correct mechanics, including the value of the test statistic and *p*-value (or rejection region).

The test statistic is $t = \dfrac{585 - 0}{\dfrac{530.71}{\sqrt{8}}} \approx 3.12.$

The *p*-value, based on a *t*-distribution with $8 - 1 = 7$ degrees of freedom, is 0.008.

Step 4: States a correct conclusion in the context of the study, using the result of the statistical test.

Because the *p*-value is very small (for instance, smaller than $\alpha = 0.05$), we reject the null hypothesis. The data provide convincing evidence that, on average, women pay more than men in the county for the same car model.

## Scoring

Each of steps 1, 2, 3, and 4 were scored as essentially correct (E), partially correct (P), or incorrect (I).

**Step 1** is scored as follows:

Essentially correct (E) if the response identifies the correct parameter *AND* states correct hypotheses.

Partially correct (P) if the response identifies the correct parameter *OR* states correct hypotheses, but not both.

Incorrect (I) if the response does not meet the criteria for E or P.

*Note*: Defining the parameter symbol in context or simply using common parameter notation is sufficient.

**Step 2** is scored as follows:

Essentially correct (E) if the response identifies the correct test procedure (by name or by formula) *AND* checks <u>both</u> conditions correctly.

Partially correct (P) if the response correctly completes two of the three components (identification of procedure, check of randomness condition, check of normality condition).

Incorrect (I) if the response does not meet the criteria for E or P.

*Note*: The random sampling condition can be verified by referring to the random selection of car models or to the random selection of male and female car buyers.

**Step 3** is scored as follows:

Essentially correct (E) if the response correctly calculates both the test statistic and the *p*-value.

Partially correct (P) if the response correctly calculates the test statistic but not the *p*-value;

*OR*

if the response calculates the test statistic incorrectly but then calculates the correct *p*-value for the computed test statistic.

Incorrect (I) if the response does not meet the criteria for E or P.

*Note:* If the response identifies a *z*-test for a mean as the correct procedure in step 2, then the response can earn a P in step 3 if both the test statistic and the *p*-value are calculated correctly.

**Step 4** is scored as follows:

Essentially correct (E) if the response provides a correct conclusion in context, also providing justification based on linkage between the *p*-value and the conclusion.

Partially correct (P) if the response provides a correct conclusion with linkage to the *p*-value, but not in context;

*OR*

if the response provides a correct conclusion in context, but without justification based on linkage to the *p*-value.

Incorrect (I) if the response does not meet the criteria for E or P.

*Notes*:

- If the conclusion is consistent with an incorrect *p*-value from step 3 and also in context with justification based on linkage to the *p*-value, step 4 is scored as E.
- A response that performs a two-sample *t*-test with correct calculations should fail to reject $H_0$. A conclusion that is equivalent to "accept $H_0$" (such as "we conclude that women pay the same amount as men, on average"), either as a stated decision or as a conclusion in context, cannot be scored as E . Such a response will be scored as P provided that the conclusion is in context with linkage. Such a response will be scored as I if it lacks either context or linkage.

Each essentially correct (E) step counts as 1 point. Each partially correct (P) step counts as ½ point.

**4 Complete Response**

**3 Substantial Response**

**2 Developing Response**

**1 Minimal Response**

If a response is between two scores (for example, 2½ points), use a holistic approach to decide whether to score up or down, depending on the overall strength of the response and communication.