DALITE Q7 - Two Sample Inference. Solutions.

EPIB607 - Inferential Statistics[a]

[a]Fall 2018, McGill University

This version was compiled on November 8, 2018

This DALITE quiz will cover two sample inference. This builds on the one sample inference we have seen so far.

Two sample means | Two sample proportions

1. A researcher who wished to test the difference of two (independent) y-bars with reported SEMs, SE0 and SE1 , did so by computing which test statistic?

a. (ybar1 - ybar0) / (SE1 + SE0)
b. (ybar1 - ybar0) / (SE1 - SE0)
c. **(ybar1 - ybar0) / sqrt((SE1)^2 + (SE0)^2) (Correct)**
d. (ybar1 - ybar0) / sqrt((SE1)^2 - (SE0)^2)

1.1. Correct rationales.

- Two-sample t-test statistic - A isn't right because we aren't taking the squares of the sum of the SEs and then taking the sqrt in the denominator - B isn't right because it is taking the difference in the denominator - D isn't right because it is taking the difference in the denominator

1.2. Incorrect rationales.

- We compute the two-sample t-statistic which is given by t = (ybar1 - ybar0) / SE, and SE of the difference in the sampling means = sqrt(s1^2/n1 + s2^2 / n2), which can be simplified to (sqrt(s1^2) / sqrt (n1)) + (sqrt(s2^2) / sqrt(n2)) = (s1/sqrt(n1)) + (s2/sqrt(n2)) = SE1 + SE0
- Since the SEMs are given, they are simply added to obtain the denominator of the t statistic formula.
- sqrt(s1^2/n1 + s2^2/n2)= s1/sqrt(n1) + s2/sqrt(n2) = SE1 +SE2
- Because SE= sigma / sqrtn
- This is the formula for two-sample inference.

2. Which distribution is closer to Normal ?

a. Body weights of (unrelated) male air passengers
b. Body weights of (unrelated) female air passengers
c. **Sum of weights of a (random male, random female) pair (Correct)**
d. **Difference of weights of a (random male, random female) pair (Correct)**

2.1. Correct rationales.

- when two random variables are normally distributed, the new variable "difference in sample means" also follows a normal distribution, centered on the difference in two variables' means and with variance equal to the sum of the two variables variances. The same applies for the sum but centered on the sum of the means. C and D are correct.
- When 2 random variables are Normally distributed, the new variable "difference in sample means" is also Normally distributed and is centered on the difference in 2 variables' means

2.2. Incorrect rationales.

- difference has less variance, so more normal
- males show more variation
- We can reasonable assume that unrelated air passengers are a SRS and body weight follows a normal distribution.
- Adding weights would increase n, which could make the distribution more normal

3. Refer to the Table on "Postoperative Effect on Plasma Ascorbic Acid for 105 Cases; Readings on the Same Individuals". You would get contradictory answers if you tested the differences in the means using i. the SE's in columns 1 and 2, and ii. the SE in column 3. TRUE or FALSE? Explain why in your rationale and say which one is correct.

a. **TRUE (Correct)**
b. FALSE

Postoperative Effect on Plasma Ascorbic Acid for 105 Cases; Readings on the Same Individuals

All Cases No. = 105	(1) PreopValue mg/100 ml	(2) PostopValue mg/100 ml	(3) Difference (postop -preop) mg/100 ml
Mean	0.43	0.36	-0.07
SE	0.036	0.028	0.015
95% CI	0.36 to 0.50	0.30 to 0.41	-0.10 to -0.04
			t = |4.93|
			P ≤ 0.01(sig)

3.1. Correct rationales.

- The SE in column 3 is the SE for a new variable which is the di fference between pre and post and it is different than the SE in columns 1 and 2 which correspond to the Se for each individual group.
- You would get contradictory answers since the proper way to calculate the standard error for a two-sample t-test is to take the square root of the sum of the standard errors of each column. Therefore, the correct way is to use the SE's in columns 1 and 2

3.2. Incorrect rationales.

- For two-sample test: SE = sqrt (SE1^2 + SE2^2) SE in column3 comes from SE in columns 1 and 2
- The SE in column 3 is the SE of the sampling distribution of the mean difference and 1 and 2 are SE of the respective mean. When computing the mean difference, SE calculation accounts for both SEs in 1 and 2.

4. How many calories does a fruit juice popsicle labeled "no sugar added" pack? A random sample of 10 popsicles of a particular brand is selected, and the calories per serving are measured. To get a confidence interval for the mean calories per serving of all no-sugar-added popsicles from this brand, you would use

a. **the one-sample t interval (Correct)**
b. the matched pairs t interval
c. the two sample t interval
d. the one sample z interval

4.1. Correct rationales.

- I would use the one-sample t interval because we are only comparing one mean and the population standard deviation is unknown.

4.2. Incorrect rationales.

- Since we are only inferring about one sample you will do a one sample z interval

5. A study is designed to compare the amount of vitamin C (in milligrams per serving) in oranges that reach the stores either 1 day after being picked or 3 days after being picked. A random sample of 15 oranges is taken for each group. To test whether there is a difference in the mean vitamin C amount of all oranges reaching stores either 1 day or 3 days after being picked, you would use

a. the one-sample t test
b. the one-sample z test
c. the one-sample t interval
d. the matched pairs t test
e. **the two-sample t test (Correct)**

5.1. Correct rationales.

- It is a two-sample t test because we are comparing means of two different samples. It is not the matched pairs t-test because the two samples are not paired, the oranges reaching the stores one day after picking do not have an impact on the oranges reaching the stores three days after picking, they are independent of each other.
- There are two samples and they are not matched, because the samples are not from the same oranges.
- You have two samples, so only D or E should be considered. The way the question is worded, we have a random sample for EACH of the groups and so it's not a matched test. We measure oranges that get to the store 1 day after picking and then measure a different group of oranges that arrive 3 days after picking.

5.2. Incorrect rationales.

- This is a before-after test and so the two samples are not independent of each other.
- The matched pairs t test is concerned with the difference between two measures in one population. In this case, we wish to compare the mean vitamin C amount in the same population (of oranges) at two time points post-harvesting: 1 day or 3 days. We are using the t-test because the sample size is small (n<30). Thus we use the matched pairs t-test.

6. There are two common methods for measuring the concentration of a pollutant in fish tissue. Do the results of the two methods differ on the average? You apply both methods to one sample of 18 carp and use

a. **the one-sample t test (Correct)**
b. the one-sample z test
c. **the matched pairs t test (Correct)**
d. the two-sample t test

6.1. Correct rationales.

- A can be correct since can do a t-test on the difference
- This is a matched pairs t test because you are applying the two difference methods to the sample sample of fish. Therefore the samples are not independent of each other and are paired.

6.2. Incorrect rationales.