DALITE Q5 - Bootstrap, Tests of Significance, Binomial Distribution, Inference for Means and Proportions. Solutions.

EPIB607 - Inferential Statistics[a]

[a] Fall 2019, McGill University

This version was compiled on October 15, 2019

This DALITE quiz will cover the bootstrap, an introduction to significance testing, and inference for a single mean using the t distribution.

Hypothesis testing | Bootstrap | t distribution | One sample mean | Normal calculations | Confidence intervals | Central Limit Theorem (CLT)

1. Hypothesis tests 1

The average human gestation time is 266 days from conception. A researcher suspects that proper nutrition plays an important role and that poor women with inadequate food intake would have shorter gestation times even when given vitamin supplements. A random sample of 20 poor women given vitamin supplements throughout the pregnancy has a mean gestation time from conception of $\bar{y}$=256 days. The null hypothesis for the researcher's test is

a. **$\mathbf{H_0}$: $\mu = 266$ (Correct)**
b. H_0: $\mu = 256$
c. H_0: $\mu < 266$
d. H_0: $\bar{y} = 266$

1.1. Correct rationales.

- The null hypothesis is usually a statement of “no effect” or “no difference.”
- The null hypothesis is the statement of no effect, so if there was no effect, the average gestation time would be the population mean.
- The null hypothesis of this study is that the average gestation period of women with proper nutrition and poor women with inadequate food intake is the same that is 266 days. Hence, the population mean (all poor women) have the gestation period of 266 days.

1.2. Incorrect rationales.

- The null hypothesis is “the accepted fact”, which in this case is that poorer nutrion will yield results lower than the mean (<266) or H null: u<266
- The researcher is testing that the hypothesis that poor women with inadequate food intake will have lower gestation time

2. Hypothesis tests 2

The average human gestation time is 266 days from conception. A researcher suspects that proper nutrition plays an important role and that poor women with inadequate food intake would have shorter gestation times even when given vitamin supplements. A random sample of 20 poor women given vitamin supplements throughout the pregnancy has a mean gestation time from conception of $\bar{y}$=256 days. The researcher's alternative hypothesis for the test is

a. H_a: $\mu \neq 256$
b. **$\mathbf{H_a}$:** $\mu < 266$ **(Correct)**
c. H_a: $\mu < 256$

2.1. Correct rationales.

- The alternative hypothesis is the point of view of the researchers. In this scenario, the researchers are suspecting that there will be a shorter gestation time for poor women in comparison to an average woman. The alternative hypothesis should be one-sided because the investigators only want to know if gestation time will be shorter and not if it is overall different (shorter or greater) from the average gestation time.
- Researcher has specified a specific direction - that the sample would generate a mean gestation time less than 266. Therefore the alternative hypothesis is that μ is less than 266.

2.2. Incorrect rationales.

3. Hypothesis test 3

The average human gestation time is 266 days from conception. A researcher suspects that proper nutrition plays an important role and that poor women with inadequate food intake would have shorter gestation times even when given vitamin supplements. A random sample of 20 poor women given vitamin supplements throughout the pregnancy has a mean gestation time from conception of $\bar{y}$=256 days. Human gestation times are approximately Normal with standard deviation $\sigma = 16$ days. The p-value for the researcher's test is (provide your calculation in the rationale)

a. more than 0.1
b. **less than 0.01 (Correct)**
c. less than 0.001
d. less than 0.05
e. less than 0.025

3.1. Correct rationales.

- $Z = (y - \mu)/(\sigma/sqrt(n))$ = -2.795 → `pnorm(-2.795, mean = 0 , sd=1 ) = 0.00026`
- `mosaic::xpnorm(q= (256 - 266)/(16/sqrt(20)))`
- `stats::pnorm(q = 256, mean = 266, sd = 16/(sqrt(20)), lower.tail = T)`

3.2. Incorrect rationales.

- $t* = (256 - 266)/(16/\sqrt(20))$ = -2.795. degrees of freedom = n-1 = 19. p-value is between 0.01 and 0.005. This is a one tailed t-test so divide by 2 and p-value is between 0.005 and 0.0025. Therefore, less than 0.01.
- $(256 - 266)/16 = -0.625$ → `pnorm(-0.625)` corresponds to a p value of 0.27
- $t = (\bar{y} - \mu)/(\sigma/\sqrt{n})$ then use t distribution with df=19

4. Hypothesis tests 4

Average human gestation time is 266 days, when counted from conception. A hospital gives a 90% confidence interval for the mean gestation time from conception among its patients. That interval is 264 ± 5 days. Is the mean gestation time in that hospital significantly different from 266 days?

a. **It is not significantly different at the 10% level and therefore is also not significantly different at the 5% level. (Correct)**
b. It is not significantly different at the 10% level but might be significantly different at the 5% level. c. It is significantly different at the 10% level

4.1. Correct rationales.

- The average human gestation time falls within the interval of the hospital which is 259-269 days, therefore the mean gestation time is not significantly different at the 10% level and since the 5% level is even more specific, then it is not significantly different at the 5% level either.
- 266 is included in the 259-269 range, therefore the mean gestation time in the hospital is not significantly different from 266 days at the 10% level. Looking at the significance at the 5% level would make the confidence interval bigger, meaning there's not significant different there either

4.2. Incorrect rationales.

- Because 266 falls at the upper tail of the 90% confidence interval, therefore it is significant different at 10% significance level
- It is significantly different at the 10% level, since the CI does not contain the value zero.
- The 90% CI includes 266, but the 95% CI may not include 266, in which case a value like 266 would only be observed less than or equal to 5% of the time.
- To be significantly different, p-value should be less than 5%.
- As 266 falls within 264+/- 5 days, it is not significant at the 10% level, but it may not fall under this range because when the level of confidence is increased, the confidence interval becomes smaller.

5. One sample mean

We prefer the t procedures to the z procedures for inference about a population mean because

a. z can be used only for large samples
b. z **requires that you know the population standard deviation σ (Correct)**
c. z requires that you can regard your data as an SRS from the population

5.1. Correct rationales.

- The reason we can use z on large samples is because when we have large samples, we can use the sample sd as an estimate for the population sd. The core reason for using t is because we don't know the population sd and we can't estimate the sd with the small sample size.
- because you don't know population sd, only sample sd for t procedures
- Z test needs to have a known population standard deviation sigma and either a normal distribution or the sample size n is large. If the population standard deviation is unknown and the sample size is small then t test statistic can be used.

5.2. Incorrect rationales.

- df are smaller for t-test, and you do not need a SD value

6. One sample mean 2

Because t procedures are robust, the most important condition for their safe use is that

a. the population standard deviation σ is known
b. the population distribution is exactly Normal
c. **the data can be regarded as an SRS from the population (Correct)**
d. the CLT hasn't kicked in yet
e. the sample size is small

6.1. Correct rationales.

- The point of the t test is that sigma is not known (rules out A) The CLT says that the population distribution doesn't matter (rules out B) We want the CLT to kick in (rules out D) The larger the sample size the better(rules out E).

6.2. Incorrect rationales.

- We should only use t procedures if the CLT is not applicable.
- In order to use the T procedure comfortably, the sample size must be large and the population distribution must be normal.

7. Bootstrap

Which of the following statements *best* describes the utility of the bootstrap

a. The bootstrap frees us from the requirement of using simple formulas to derive confidence intervals
b. The bootstrap allows us to simulate a sampling distribution
c. **The bootstrap frees us from the assumption of a Gaussian sampling distribution for the mean (as per the CLT) (Correct)**
d. The bootstrap tells us if the sampling distribution is asymmetric

7.1. Correct rationales.

- Computer intensive methods can solve most problems without assuming that the data have a Gaussian distribution.
- The bootstrap frees us from the assumption that the data conform to a bell-shaped normal distribution.

7.2. Incorrect rationales.

- due to the nature of bootstrap, you can resample things MANY times, which means that CLT will always kick in giving a gaussian distribution instead of assuming one exists
- Bootstrap is useful in because it does not require the CLT and **the population** does not have to be normally distributed
- The bootstrap doesn't rely on CLT, and gives us freedom from having to assume the normal distribution of **the population**.
- The bootstrap allows us to derive estimates when often-used theories do not apply. This is especially useful when n is small (e.g. CLT requires n > 30 to assume normality).

8. Binomial Distribution 1

In which of the following would Y not have a Binomial distribution? Provide your justification in the rationale.

a. **Y = Number, out of 60 occupants of 30 randomly chosen cars, wearing seatbelts. (Correct)**
b. Y = Number, out of 60 occupants of 60 randomly chosen cars, wearing seatbelts.
c. Y = Number, out of simple random sample of 100 individuals, that are left-handed.

8.1. Correct rationales.

- 60 occupants from 30 cars would mean more than one occupant per car - whether each occupant is wearing a seatbelt or not isn't independent if more than one comes from the same car.
- There is a chance that if someone in the car is not wearing a seatbelt, the other passenger in the car is not too. This means that it is not independent. #idiotswhodon'twearseatbeltsdrivetogether
- Not independent, idiots who don't wear seatbelts in cars drive together!!! VROOM VROOM
- If there are more than one occupant in a car, they may influence each other to wear or not wear a seat belt.
  - Richard: "hey Timmy, only losers wear seat belts, be cool like me and don't wear your seat belt"
  - Timothy: "oh snap, you're right Ricky, not wearing a seat belt is the fleekest"
  - Richard and Timothy: "we're going to live forever"
- The outcome does not fall into one of two categories as someone can be left-handed, right-handed, or ambidextrous.

8.2. Incorrect rationales.

- Each trial does not have an equal probability of success as there are more right-handed people than left-handed people.
- Because in this case the observations are not independent. If the first volunteer selected is left-handed, the second one is more likely to be right-handed because there are more right-handed individuals than left-handed in the remaining sample.

9. Binomial Distribution 2

In which of the following would Y not have a Binomial distribution? Provide your justification in the rationale.

a. **You want to know what percent of married people believe that mothers of young children should not be employed outside the home. You plan to interview 50 people, and for the sake of convenience you decide to interview both the husband and the wife in 25 married couples. The random variable Y is the number among the 50 persons interviewed who think mothers should not be employed. (Correct)**
b. You observe the sex of the next 50 children born at a local hospital; Y is the number of girls among them.
c. Y = number of occasions, out of a randomly selected sample of 100 occasions during the year, in which you were indoors.

9.1. Correct rationales.

- Husband's opinion may depend on wife's opinion, and vice versa - observations aren't independent.
- One of the conditions for the binomial distribution is that the trials must be independent. In A, they are selecting 25 married couples in order to gather data on 50 people, but in this scenario either the husband or the wife could influence the other, so the trials are not independent.

9.2. Incorrect rationales.

10. Binomial Distribution 3

The U.S. National Center for Health Statistics reports that approximately 12% of emergency department visits result in hospital admissions. Consider 20 randomly selected emergency department visits and assume that visits to emergency departments are independent. What is the approximate probability that at most 2 of the 20 visits would result in hospital admissions?

a. **0.5631 (Correct)**
b. 0.2740
c. 0.1344
d. 0.2891
e. 0.4369

10.1. Correct rationales.

- We have to callculate the probability at 0,1, and 2 and add them up: For 0: (1)(1)(0.88)ˆ20 For 1: 20(0.12)(0.88)ˆ19 For 2: 190(0.12)$^{2(0.88)}$18 Adding these all up equals 0.5631
- P(X< or=2) = P(X=2)+P(X=1)+P(X=0) = 0.5631
- pbinom(2, 20, 0.12, lower.tail = TRUE)

10.2. Incorrect rationales.

- 1- pbinom(2, size = 20, prob = 0.12)

11. Binomial DIstribution 4

The U.S. National Center for Health Statistics reports that approximately 12% of emergency department visits result in hospital admissions. Consider 20 randomly selected emergency department visits and assume that visits to emergency departments are independent. How many hospital admissions do we expect, on average, in a random sample of 20 emergency department visits?

a. 2
b. **2.4 (Correct)**
c. 12
d. 24
e. 1.4533

11.1. Correct rationales.

- mean of a binomial distribution = np = 20*0.12 = 2.4
- μ = np = (20)(0.12) = 2.4 hospital admissions

11.2. Incorrect rationales.

- 0.12 x 20 = 2.4 But we can't in a random sample get .4, so we choose 2.

12. Binomial Distribution 5

The U.S. National Center for Health Statistics reports that approximately 12% of emergency department visits result in hospital admissions. Consider 20 randomly selected emergency department visits and assume that visits to emergency departments are independent. If Y is the number of emergency department visits that result in hospital admissions in random samples of 20 visits, what is approximately the standard deviation of Y?

a. **1.45 (Correct)**
b. 2.11
c. 4.46

12.1. Correct rationales.

- The standard deviation of U can be solved using the equation sigma = sqrt(np(1-p)) = sqrt(20*.12*(1-0.12)) = 1.45

12.2. Incorrect rationales.

13. Binomial Distribution 6

According to the 2015 U.S. census update, approximately 13% of Americans are black. Let Y be the number of blacks in a random sample of 1500 Americans. What is the probability that the sample will contain 200 or more blacks?

a. pbinom(q = 200, size = 1500, prob = 0.13)
b. 1-dbinom(x = 200, size = 1500, prob = 0.13)
c. dbinom(x = 200, size = 1500, prob = 0.13)
d. `pnorm(200, mean = 195, sd = 13.025, lower.tail = FALSE)` **(Correct)**
e. `1 - pbinom(q = 200, size = 1500, prob = 0.13, lower.tail = FALSE)`

13.1. Correct rationales.

- a) is not correct because it is looking at all values below 200 b) is not correct because it is looking at the probability of getting anything but 200 c) is not correct because it is looking at getting exactly 200 e) is not correct because it is looking at all values below **201** d) since sample size is large, we can use normal approximation for binomial distribution

- In this case, we can use the normal approximation because n is large enough, n*p = 195 > 10, n(1-p) = 1305 > 10. Therefore, we can use pnorm function with q=200, lower.tail false. A doesn't specify a lower.tail=FALSE, so R gives lower.tail = TRUE by default. dbinom are wrong.

13.2. Incorrect rationales.

- because you are calculating the sample that will contain 200 or more, lower.tail is false.

- E) would be right if it didn't have the 1 - before the code.

- Enough ppl not to do a t-test

14. One Sample Proportion 1

A CDC report on secondhand smoke at home gives the following 95% confidence interval for the percent of California households that are free of secondhand smoke: (90.8, 92.2). The correct interpretation for this confidence interval is that we can be 95% confident that

a. the proportion of households free of secondhand smoke in another sample of California households would be between 0.908 and 0.922
b. the population mean number of households in California that are free of secondhand smoke is between 90.8 and 92.2
c. **the true proportion of all California households that are free of secondhand smoke is between 0.908 and 0.922 (Correct)**

14.1. Correct rationales.

- The answer is C because the confidence interval looks at where the true population proportion will fall 95% of the time.
- A is not right because the CI is only for that trial, not for all trials. B is not right because they are not talking about a mean value but a proportion.

14.2. Incorrect rationales.

15. One Sample Proportion 2

A CDC report on secondhand smoke at home gives the following 95% confidence interval for the percent of California households that are free of secondhand smoke: (90.8, 92.2). What is the margin of error for this interval?

a. **0.007 (Correct)**
b. 0.014
c. 0.028

15.1. Correct rationales.

- The margin of error is half the width of the confidence interval.
- The margin of error is half of the confidence interval. 92.2 - 90.8 / 2 = 0.007

15.2. Incorrect rationales.

16. One Sample Proportion 3

How many observations must be recorded to estimate a population with unknown proportion p to within +/- 0.02 with 95% confidence?

a. n=25
b. n=1225
c. **n=2401 (Correct)**
d. n=2350
e. n=1691

16.1. Correct rationales.

- We would have to guess p in this case because it is not given. It is best to assume p is 0.5 when it is unknown because the margin of error is largest when p is 0.5 so by doing so we are giving a more conservation estimation for how many people we will need. We then use the formula n=(z/m)ˆ2(p)(1-p) where p=0.5, z=1.96, m=0.02.

16.2. Incorrect rationales.

- 1691 = (1.645/0.02)ˆ2 * (0.5 * (1-0.5))
- Using the formula: n = (z*/m)(p)(1-p) = (1.96/0.02)(0.5)(1-0.5) = 24.5 = 25. Here, test is conservative and p is set to 0.5.