Midterm 2 Review

Elaine Chiu

10/24/2021

1. Concepts and Keywords

- Expected value and Standard Error
- Expected Sum and Expected percentage
- P-value
- Null Hypothesis and Alternative Hypothesis
- standard Normal Distribution and t distribution
- z score
- Confidence Interval
- Law of Large Number and Central Limit Theorem
- one-tailed and two tailed tests
- Population, Parameter, Sample, Statistic
- Chance error
- Systematic Bias
- Type I Error and Type II error

2. selected Problems

Problem 1

One hundred draws would be made at random from the box $[1, 6, 7, 9, 9, 10]$. How large can the sum of the draws be? How small? What is the chance that the sum is between 650 to 750?
Answer:
The smallest sum would be $1 \times 100 = 100$, The largest sum would be $10 \times 100 = 1000$. To know the change of the sum between 650 to 750 we need to find the expected value and the standard error. The expected value for 100 draws is $100\times$ (the expected value of one single draw). The expected value of one single draw is $1 \times \frac{1}{6} + 6 \times \frac{1}{6} + 7 \times \frac{1}{6} + 9 \times \frac{2}{6} + 10 \times \frac{1}{6} = 7$, thus the expected value of 100 draws is $7 \times 100 = 700$. Now compute the standard error of 100 draws. We have SE(100 draws) = SE(one single draw) $\times\sqrt{100}$. The standard error for one single draw is $\sqrt{(1-7)^2 \times \frac{1}{6} + (6-7)^2 \times \frac{1}{6} + (7-7)^2 \times \frac{1}{6} + (9-7)^2 \times \frac{2}{6} + (10-7)^2 \times \frac{1}{6}} = 3$, thus the SE for 100 draws is $3 \times \sqrt{100} = 30$. Based on central limit theorem, we could compute the z score of 650 and 750 and find the area. $\frac{650-700}{30} = -1.67$, $\frac{750-700}{30} = 1.67$, find the area between -1.67 to 1.67 to get the probability is 90%

Problem 2

According to the simplest genetic model, the sex of a child is determined at random, as if drawing a ticket at random from the box [male, female]. what is the chance of next 2500 births, more than 1275 would be female?
Answer:

Since we only have two possible outcomes, female or male, this is a binomial distribution. The expected value of getting a female baby in one birth is $\frac{1}{2}$, thus the expected births of female babies out of 2500 births is $2500 \times \frac{1}{2} = 1250$. SE(2500 births) $= \sqrt{2500} \times$ SE(one single birth). SE(one single birth) = $\sqrt{\frac{1}{2} \times \frac{1}{2}} = \frac{1}{2}$, thus SE(2500 births) = $\sqrt{2500} \times \frac{1}{2} = 25$. Use these numbers to get the z score of 1275. $\frac{1275-1250}{25} = 1$. Find the area more than 1 z score on the z table, it is about 16%.

Problem 3

Find the expected value of the sum of 100 draws from the box $[0, 1, 1, 6]$.
Answer:
The expected value of 1 single draw is $0 \times \frac{1}{4} + 1 \times \frac{2}{4} + 6 \times \frac{1}{4} = 2$, thus the expected value of 100 draws is $2 \times 100 = 200$.

Problem 4

Suppose there is a box of $100,000$ tickets, each marked 0 or 1. Suppose that in fact, 20% of tickets are 1's. Calculate the standard error of percentage of 1's in 400 draws.
Answer:
The standard error for one single draw is $\sqrt{20\% \times 80\%} = 0.4$, the standard error of 400 draws is $\sqrt{400} \times 0.4 = 8$. Convert this into percentage, is $\frac{8}{400} = 2\%$.

Problem 5 Confidence Interval Computation I

You want to rent an unfurnished one-bedroom apartment in Durham, NC next year. The mean monthly rent for a random sample of 60 apartments advertised on Craig's List (a website that lists apartments for rent) is 1000. Assume a population standard deviation of 200. Construct a 95% confidence interval.
Answer:
For the computation of confidence interval, remember that your center of the confidence interval is the sample mean. In this case is 100. And then you need to compute the standard error of the sample mean(!).

$$\text{standard deviation of sample mean} = \frac{\text{SE for a single box}}{\sqrt{n}}$$

In this case, would be $\frac{200}{\sqrt{60}}$. Also remember, if you get your standard deviation from the sample, you would need to use t distribution; if you get your standard deviation from the population, then you can use z distribution.
The z score corresponding to 95% confidence interval in this case is about 1.96, so the confidence interval is $[949.39, 1050.61)]$

Problem 6 Review of z score I

Suppose a population was normally distributed with a mean of 10 and standard deviation of 2. What proportion of the scores are below 12.5?
Answer:

With proportion, we can think of what percentage of the scores are below 12.5. To do this, first we need to calculate the Z score associated with 12.5. Using the formula $\left(z=\frac{(\text{observed}-\text{population mean})}{SE}\right)$, we plug in values: $\frac{12.5-10}{2}=1.25$ Then we look this up in the table. It's a positive value, and we want the scores below this which is 0.8944. For proportion, multiple this number by 100, and round to 2 decimal points. So 89.44% of the population is below this score.

Problem 7 Review of z score II

One year, many college-bound high school seniors in the U.S. took the Scholastic Aptitude Test (SAT).For the verbal portion of this test, the mean was 425 and the standard deviation was 110.
Based on this information what percentage of students would be expected to score between 350 and 550?
Answer:
The z score corresponding to 350 points is $\frac{350-425}{110}=-0.68$, The z score corresponding to 550 points is $\frac{550-425}{110}=1.14$, check the z table to find the area between these two z scores is $.6245$, so So 62.45% of the students would be expected to score between 350 and 550 on their verbal SAT.

Problem 8 Confidence Interval II t distribution version

A group of 10 foot surgery patients had a mean weight of 240 pounds. The sample standard deviation was 25 pounds. Find a confidence interval for a sample for the true mean weight of all foot surgery patients. Find a 95% CI.
Answer:
- Subtract 1 from your sample size. $10-1=9$. This gives you degrees of freedom.
- Subtract the confidence level from 1, then divide by two. $\alpha=(1-.95)/2=.025$
- Look up your t score in the t-distribution table. For 9 degrees of freedom (df) and $\alpha=0.025$, it is 2.262. - Divide your sample standard deviation by the square root of your sample size,this gives me the standard deviation of sample mean. $\frac{25}{\sqrt{10}}=7.9$. - confidence interval is then

$$[240-2.262*7.9, 240+2.262*7.9]=[222, 257.88]$$

Problem 9 - Hypothesis Testing Problem I

Marketers believe that 92% of adults in the United States own a cell phone. A cell phone manufacturer believes that number is actually lower. 200 American adults are surveyed, of which, 174 report having cell phones. Use a 5% level of significance. State the null and alternative hypothesis, find the p-value, state your conclusion.

Answer:

null hypothesis :$p=0.92$
Alternative Hypothesis:$p<0.92$
our sample percentage is $\frac{174}{200}=87\%$.
The SE of the sample mean under null hypothesis is $\sqrt{\frac{92\%\times 8\%}{200}}=0.0192=1.92\%$.
The z score is $\frac{87\%-92\%}{1.92\%}=-2.60$.
The p value is then the area less than z score, $0.0046=0.46\%$, since this is less than 5%, we reject the null hypothesis.
p-value=0.0046 Because p<0.05 , we reject the null hypothesis. There is sufficient evidence to conclude that fewer than 92% of American adults own cell phones.

Problem 10 - Concepts of errors, P-values, and Central Limit Theorem

- If you take a random sample of size 100 from a population with an unknown distribution, the distribution of that population becomes normal because of the Central Limit Theorem.

a. True. (X)
b. False (O)
c. You need to know whether you calculated averages, sums, or proportions in order to make that claim.(X)
Answer:
Population is fixed, so the distribution of population is also fixed. In other words, no matter what statistic you are calculating, any description related to the population is fixed. Central limit theorem is saying, no matter what your population distribution is, if the sample size is large, then the sample mean, sample sums or sample proportion would be normally distributed. Notice the description is associated with the sample! not the population, and only specific statistics can utilize central limit theorem. For example, the sample product (you multiply all values together) would not be normal.(recall the weird graph from one of the assignment problem).

- Why is higher confidence in creating confidence intervals for population parameters not always better?

a. It can result in an interval that is too wide to be useful. (O)
b. It becomes worse at capturing the true population parameter.
c. This is a trick question, you should always use the interval that gives you the highest level of confidence.
d. Both A and B
e. None of the Above.

- Which statistical law tells us that with a large sample size of independent draws, the sampling distribution of an average, sum, count or percentage will follow an approximately normal distribution?
(a)Law of Large Numbers (X)
(b)Central Limit Theorem (O)
(c)Law of Averages (X) (d)Probability Theory (X)
(e)Normal Distribution “Rule of Thumb” (X)

- A sociology researcher is conducting a study to determine whether or not there is a difference in wages between male and female professors at UW-Madison. After sampling 160 female professors and comparing the sample average to the published average of all male professors, the researcher determines there is a difference between pay for the different sexes. (Assume this conclusion was reached by a hypothesis test)
What is the null hypothesis for the researcher’s study?
(a)There is no difference between male and female average wages (O)
(b)There is a difference between male and female average wages

(c)Males make higher average wages than females
(d)Males make lower average wages than females
(e)Hypothesis testing is inappropriate for this situation
Hint: We usually set new scientific discovery as the alternative hypothesis and the status quo as the null hypothesis.

- A sociology researcher is conducting a study to determine whether or not there is a difference in wages between male and female professors at UW-Madison. After sampling 160 female professors and comparing the sample average to the published average of all male professors, the researcher determines there is a difference between pay for the different sexes. (Assume this conclusion was reached by a hypothesis test)
  Suppose UW-Madison releases all information about the wages of all professors and finds out that there is actually no difference in the wages between male and female professors.
  What type of error did the researcher make with their study?

a. There was no error made in the study.
b. Type I Error – Rejected the null when the null was true (O)
c. Type II Error
d. Type III Error
e. There were unequal sample sizes
   Hint: Type I error is you reject the null when the null is true; Type II error is you do not reject the null when the null is false.

- When designing an experiment, a group of researchers decide that a type II error would be particularly bad, so they want to lower the probability of encountering type II error as much as possible. Which level of significance should they select holding sample size constant?

a. 0.1%
b. 1%
c. 5%
d. 10% (O)
e. Not enough information provided to answer.

Hint: Type II error is you do not reject the null when the null is false. If you just want to avoid type II error, you should try to reject the null. and for this question, one way to increase the probability to reject the null is you increase your significance level. Because remember! the decision rule is you reject the null when your p-value is less than significance level. Therefore, if you increase your significance level, you are more likely to reject.

- A diagnostic test is used to determine whether a patient has arthritis. A treatment will be prescribed only if the doctor thinks the test gives enough evidence to suggest the patient has arthritis. The hypotheses might be stated as follows:
  H_0: The patient does not have arthritis
  H_A: The patient has arthritis
  Label each of the following statements as ”Type 1 Error”, ”Type 2 Error” or ”No Error” as appropriate.
  • Diagnosing arthritis in a patient who has arthritis. No Error
  • Failing to diagnose arthritis in a patient who has arthritis. Type 2 Error

• Diagnosing arthritis in a patient who does not have arthritis. Type 1 Error
• Failing to diagnose arthritis in a patient who does not have arthritis. No Error

- What is the proper interpretation of a p-value?
  (a)This is the true difference between the sample value and the null hypothesis value.(X)
  (b)This is the probability the null hypothesis is true.(X)
  (C)This is the probability the alternative hypothesis is true.(X)
  (d)This is the probability of getting the sample statistic we drew in this experiment or a more extreme statistic assuming the null hypothesis is true.(O)
  (e)This is the percent chance we commit a Type II error.
  Hint: (b) and (c) are false because first, those are not the definitions of p-values, second, alternative hypothesis and null hypothesis are descriptions of the population. The population is fixed, so null hypothesis and alternative hypothesis are either true or false. No probability associated with them!