Math 10 Spring 2013

Statistics

Midterm Exam

Tuesday April 30, 5:00–7:00 PM

Your name (please print): ______________________________

Instructor (circle one): M^{c}NEW RUBINSTEIN-SALZEDO

Instructions: This is a closed book exam. You may refer to a $4'' \times 6''$ notecard that you have prepared, and you may use a calculator.

The Honor Principle requires that you neither give nor receive any aid on this exam.

Please sign below if you would like your exam to be returned to you in class. By signing, you acknowledge that you are aware of the possibility that your grade may be visible to other students.

____

For grader use only:

Problem	Points	Score
1	6	
2	8	
3	10	
4	10	
5	12	
6	10	
7	15	
8	15	
9	14	
Total	100	

1. Explain what a double-blind study is. What are the advantages of using them?

   A double-blind study is an experiment with a control group and a test group in which neither the subject nor the researcher knows which group the subject is in. Double-blind studies are advantageous because they eliminate the possibility of the researcher treating members of the two groups differently and thus potentially altering the result of the study.

2. What are the expected value and standard deviation for the sum of a roll of three dice?

   Let us call the result of the die rolls X_1, X_2, X_3. The expected value for one die is

$$\mathbb{E}(X_1) = \frac{1}{6}(1+2+3+4+5+6) = \frac{7}{2}.$$

   By linearity of expectation, we have

$$\mathbb{E}(X_1 + X_2 + X_3) = 3\mathbb{E}(X_1) = \frac{21}{2} = 10.5.$$

   Since the rolls are independent, we can use linearity of variance:

$$\mathrm{Var}(X_1 + X_2 + X_3) = 3\,\mathrm{Var}(X_1),$$

   and

$$\mathrm{Var}(X_1) = \frac{1}{6}((1-7/2)^2+(2-7/2)^2+(3-7/2)^2+(4-7/2)^2+(5-7/2)^2+(6-7/2)^2) = \frac{35}{12}.$$

   Hence $\mathrm{Var}(X_1 + X_2 + X_3) = \frac{35}{4}$, so the standard deviation is $\sqrt{\frac{35}{4}} \approx 2.958$.

3. Suppose that one person out of 500 has a certain disease. There is a test for this disease. 95% of the people who have the disease test positive, and 99% of people who do not have the disease test negative. What is the probability that a person who tests positive for the disease has the disease?

Define the events:
D: Person has the disease
+: Person tests positive

oWe would like to find the probability $P(\text{D}|+)$. Using Bayes Theorem, we have that

$$P(\text{D}|+) = \frac{P(+|\text{D})P(\text{D})}{P(+)}$$

By drawing a tree or by direct computation we find that

$$P(+) = P(+|D)P(D) + P(+|D^c) = 0.0019 + 0.00998.$$

Thus

$$P(\text{D}|+) = \frac{0.95 \times 0.002}{0.0019 + 0.00998} \approx 0.16$$

4. What types of distributions model the following random variables?

   (a) The number of times you have to shoot a basketball before you make a basket. Geometric

   (b) The number of car accidents per day in Los Angeles. Poisson

   (c) The number of times you roll a 3 among 100 rolls of a die. Binomial

   (d) The number of people you have to interview before you find the sixth left-handed person. Negative Binomial

   (e) The result of one coin flip. Bernoulli

5. Suppose you flip a coin 1000 times. Give the best estimate that you can for the probability that you get between 464 and 507 heads.

   We wish to use the normal approximation to the binomial distribution. In order to do that, we decrease the lower bound (464) by 0.5 to 463.5, and similarly, we increase the upper bound (507) by 0.5 to 507.5. The mean is $np = 500$, and the standard deviation is $\sqrt{np(1-p)} = \sqrt{250} \approx 15.811$. The Z-scores for the bounds are then

$$Z_{463.5} = \frac{463.5 - 500}{15.811} \approx -2.31,$$

   and

$$Z_{507.5} = \frac{507.5 - 500}{15.811} \approx 0.47.$$

   The probability of getting a Z-score between -2.31 and 0.47 can be obtained from looking at a Z-table: the probability of having a Z-value less than 0.47 is 0.6808, and the probability of having a Z-value less than -2.31 is 0.0104, so the probability of having a Z-value between -2.31 and 0.47 is $0.6808 - 0.0104 = 0.6704$.

6. Suppose you have sampled a population six times and obtained the following values:

$$21, 23, 24, 25, 22, 23$$

(a) Assuming these observations come from a normally distributed population, what would be your best guess for the mean μ, and standard deviation σ of the population?

The best guess for the population mean is the sample mean:

$$\bar{x} = \frac{21 + 23 + 24 + 25 + 22 + 23}{6} = 23.$$

The best guess for the population standard deviation is the *sample* standard deviation

$$s = \sqrt{\frac{(21-23)^2 + (23-23)^2 + (24-23)^2 + (25-23)^2 + (22-23)^2 + (23-23)^2}{6-1}}$$
$$= \sqrt{2} \approx 1.414.$$

(b) If we were to repeatedly sample this population, taking samples of the same size, what would we expect the standard deviation of the resulting sample distribution to be?

The standard deviation of the sample distribution is the standard error:

$$SE = \frac{s}{\sqrt{n}} = \frac{1}{\sqrt{3}} \approx 0.577.$$

7. 191 households were surveyed as part of a study on water consumption. The mean water usage per day per household in this survey was 371.4 gallons, and a 95% confidence interval for the mean water usage per household is (365.7,377.1). Determine whether the following statements are true or false, and explain your reasoning: (In one sentence.)

   (a) The standard error for samples of size 191 from this population is approximately 5.7.

   False. $\bar{x}$ lies in the center of the inteval, so $ME = 377.1 - 371.4 = 5.7$, Since $ME = z * SE$ and $z* = 1.96$, $SE = 5.7/1.96 \approx 2.91$ not 5.7.

   (b) 5% of such random samples of this size would have sample means outside of the range (365.7,377.1).

   False. This is not what the 95%confidence interval is capturing, it only tries to capture the population average.

   (c) We are 95% confident that the actual mean value lies in the range (365.7,377.1).

   True. This is the definition of a confidence interval

   (d) We would be more confident that the average lies in the range (365,380).

   True. Since the interval is larger than and contains the orginal interval, this larger interval is more likely to capture the actual population mean.

   (e) If we perform an identical study of households that have dogs and found that they use an average of 387.2 gallons per day, with a 95% confidence interval of (382.7,391.7) we could conclude that having a dog causes households to use more water.

   False. While this would show that the two are correlated, this is an observational study and so we cannot conclude that there is a causal link between dog ownership and increased water usage.

8. Census data shows that the average income level in the vicinity of a mall is \$33,950. The owners of the mall are interested in determining whether mall shoppers have a higher income level, and are not interested in any result that shoppers have a lower income. They perform a simple random sample of 50 shoppers, who had an average income of \$34,076 with a standard deviation of \$474 and the observations were not too skewed.

   (a) Can they conclude that their shoppers have a higher income level at a 5% significance level?

   We set up the hypothesis test:
   $H_0 : \mu = 33950$
   $H_A : \mu > 33950$
   We compute the Standard Error: $SE = \frac{474}{\sqrt{50}} \approx 67.03$.
   Using this to compute the z score: $z = \frac{34076-33950}{67.03} \approx 1.88$ which corresponds to a p value (one sided) of 0.0301.
   We therefore reject the null hypothesis, and find that there is evidence that the shoppers do have a higher income level.

   (b) What about a 10% significance level, or a 1% significance level?

   Since our p-value computed in part a is .03, we reject the null hypothesis at the 10% significance level, but fail to reject it at the 1% level.

   (c) At the 5% significance level, what is the power (probability of rejecting the null hypothesis) of this test against the alternative hypothesis that the actual income of shoppers averages \$34,000?

   Since this is a one sided test, the z-score corresponding to our cutoff value is z = 1.65. Using this value, we first solve for the actual cutoff value.

$$1.65 = \frac{x_{cutoff} - 33950}{67.03} \rightarrow x_{cutoff} \approx= 34060$$

   Any observed mean value greater than 34060 will cause us to reject the null hypothesis in favor of an alternative. We now use this to predict the power against the specific alternative hypothesis $\mu = 34000$. The z score for 34060 in this case is:

$$z = \frac{34060 - 34000}{67.03} \approx .9041$$

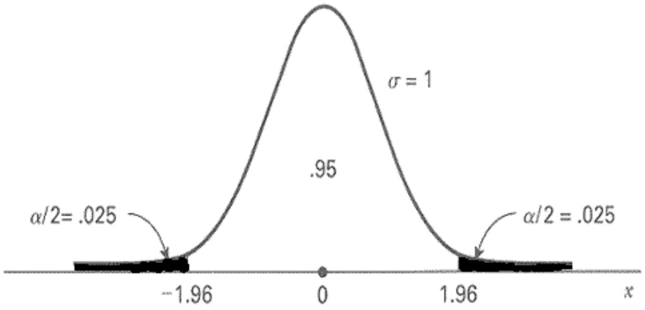
   Looking this score up in the table we find that the probability of observing a z score higher than this (the power) is 0.1841.

9. A psychology researcher wants to find out if exercising before taking a quiz affects a student's performance. To test this he randomly assigns students to either exercise for 10 minutes before taking a short quiz, or to take the quiz without exercising first. 37 students exercise first and average 84% on the quiz with a standard deviation of 7% while 32 students skip exercising and score 81% with a standard deviation of 6%. Neither sample was substantially skewed.

(a) Write down the null and alternative hypotheses in this test, and draw a picture of the sampling distribution assuming that the null hypothesis is true, shading in the region(s) that would result in a type 1 error (falsely rejecting the null hypothesis assuming that it's true.)

$H_0 : \mu_{exercise} = \mu_{no-exercise}$

$H_0 : \mu_{exercise} \neq \mu_{no-exercise}$



(b) Can he conclude at a 5% significance level that exercising has an effect?

We compute the standard error

$$SE = \sqrt{\frac{0.07^2}{37} + \frac{0.06^2}{32}} \approx 0.0157$$

and use this to compute a z score for the difference of these two means,

$$z = \frac{(.84 - .83) - 0}{.0157} \approx 1.91$$

. Looking this value up in our z table we find that this corresponds to a p-value of 0.0281. Since this is a 2 sided test, we need to multiply this value by two, giving us a p-value of .0562. Since this is greater than .05 we fail to reject the null hypothesis.