

6:45 7:15

## ST 305: Exam 2

By handing in this completed exam, I state that I have neither given nor received assistance from another person during the exam period. I have used no resources other than the exam itself and the basic mathematical functions of a calculator (ie, no notes, electronic communication, notes stored in calculator memory, etc.) I have not copied from another person's paper. I understand that the penalty if I am found guilty of any such cheating will include failure of the course and a report to the NCSU Office of Student Conduct. **I understand that I must show all work/calculations, even if they seem trivial, to get credit for my answers.**

Using your calculator for values from probability distributions like the normal, binomial, or t is OK; however, if you are doing that type of calculation, show your work all the way to the point of plugging the final entries into your calculator.

Name: \_\_\_\_\_

ID#: \_\_\_\_\_

$\bar{x} = \frac{1}{n} \sum x_i$ $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$ $Z = \frac{X - \mu}{\sigma}$ $r = \frac{\sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)}{n-1}$ $b_1 = r \frac{s_y}{s_x}$ $b_0 = \bar{y} - b_1 \bar{x}$ $\text{residual} = y - \hat{y}$ $P(A \text{ or } B) = P(A) + P(B)$ $P(A^c) = 1 - P(A)$ $P(A \text{ and } B) = P(A) \times P(B)$	$\mu_X = \sum x_i p_i$ $\mu_{a+bX} = a + b\mu_X$ $\mu_{X+Y} = \mu_X + \mu_Y$ $\sigma_X^2 = \sum (x_i - \mu_X)^2 p_i$ $\sigma_{a+bX}^2 = b^2 \sigma_X^2$ $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$ $\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2$ $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y$ $\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y$	$\mu_X = np$ $\sigma_X = \sqrt{np(1-p)}$ $\hat{p} = X/n$ $\mu_{\hat{p}} = p$ $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ $P(X=k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$ $\mu_{\bar{X}} = \mu$ $\sigma_{\bar{X}} = \sigma/\sqrt{n}$
---	---	---

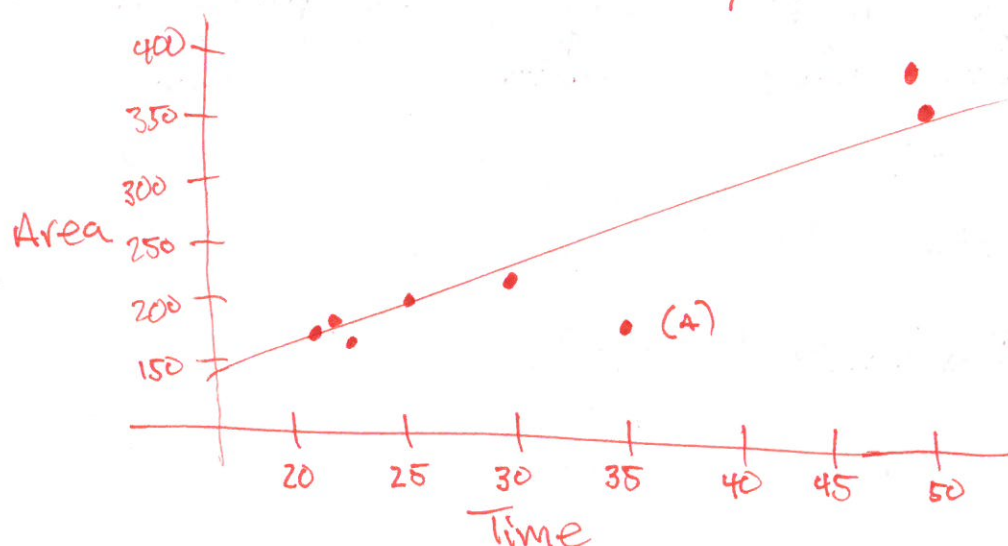
12. The following data shows the amount of time (in minutes) my robot vacuum cleaner (Rosie the Robot) was turned on and the amount of area (in square feet) it cleaned for 8 recent days:

Day	1	2	3	4	5	6	7	8
Time	35	48	26	22	49	29	21	23
Area	187	396	201	181	367	232	179	176

Not waiting for your answer to part (c) below, we applied simple linear regression to this data set and found the following results:

- $\hat{y} = 8.26 + 7.32x$
- $r^2 = 0.86$

- a. Display the relationship between Time and Area. (3 points)



- b. Describe the relationship between Time and Area? (5 points)

moderately ~~weak~~ strong, positive, linear relationship  
point A might be an outlier

- c. Predict the amount of area cleaned if Rosie is turned on for 45 minutes. (3 points)

$$\hat{y} = 8.26 + 7.32(45) = \boxed{337.66 \text{ sq ft}}$$

- d. What is the correlation between Time and Area? (2 points)

$r^2 = .86 \Rightarrow r = .93$  (graph & LS line show us it is positive)

13. An experiment is conducted to study the effect of sleep on stress level. It is conducted as follows:

- A SRS of 200 men and 150 women are selected
- Of those 350 people, 100 random men and 75 random women are restricted to only 6 hours of sleep per night (Sleep Deprived). The others sleep as much or as little as they like (Free Choice).
- At the end of 3 months, the individuals' stress levels are measured and the average values between Sleep Deprived and Free Choice individuals are compared.

Answer the following questions about this experiment (2 points each):

- How many subjects were there? **350**
- Clearly identify the levels and factors in this experiment?  

<b>Sex (M/F)</b> ↑ Factor	*	↑ Levels	<b>Sleep (Sleep Deprived / Free Choice)</b> ↑ Factor	↑ Levels
---------------------------------	---	-------------	--	-------------
- What was the explanatory variable?  
**Sleep**
- What was the response variable?  
**Stress Level**
- What type of experimental design was this? (Just give the name)  
**Randomized Block**

14. The number of days it takes tomato plants to grow to full height has a distribution that is skewed to the right. The distribution has mean 48 days and standard deviation 6 days.

For each of the following, find the probability if possible. If you can't compute the probability, explain why (or explain what information is missing).

a. Find the probability that a single randomly chosen plant takes more than 52 days to reach full height. (5 points)

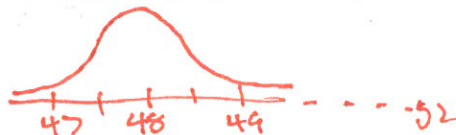
**Not enough info. We only have the mean & std dev of the dist'n, we need to know the type of dist'n (e.g., is it normal?)**

b. Suppose we take a SRS of 144 plants and compute the sample mean number of days to full height for those 144 plants. What is the probability that the value of this sample mean is greater than 52 days? (5 points)

$$\bar{X} \sim N(48, \frac{6}{\sqrt{144}})$$

$$\bar{X} \sim N(48, .5)$$

$$P(\bar{X} > 52) = P(Z > \frac{52 - 48}{.5}) = P(Z > 8) \approx 0$$



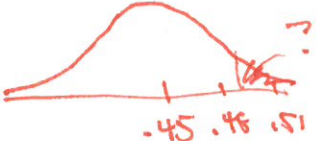




15. A recent study suggested that roughly 45% of calls to cell phone numbers in 2019 will be "spam" calls falling into one of two categories: 40% will be robocalls, 5% will be phishing calls. Suppose we plan to take a SRS of 225 calls.

a. If the values above are actually correct, what is the probability that at least 50% of the 225 calls in our sample will be spam? (10 points)

$P(\text{spam}) = .45$       Need to find:  $P(\hat{p} \geq .50)$   
 $\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$        $P(\hat{p} \geq .50)$   
 $\hat{p} \sim N(.45, .033)$        $= P(Z \geq \frac{.50 - .45}{.033})$   
 $= P(Z \geq 1.52)$   
 $=$



b. We are planning to estimate the proportion of all calls that are spam by using the sample proportion,  $\hat{p}$ . Prove that  $\hat{p}$  is an unbiased estimate of the true proportion of all calls that are spam. (HINT: set up an indicator function and use its properties to derive the result.) (10 points)

Define  $Y_i = 1$  if the  $i$ th call is spam, 0 if not  
 $\hat{p} = \frac{X}{n}$ , where  $X =$  total # of spam calls in sample  
 $X = \sum_{i=1}^n Y_i$   
 $E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{1}{n} E\left(\sum Y_i\right) = \frac{1}{n} \sum_{i=1}^n E(Y_i)$   
 $E(Y_i) = 1(p) + 0(1-p) = p$   
 $\rightarrow = \frac{1}{n} \sum_{i=1}^n p = \frac{1}{n} (np) = p$   
 $\Rightarrow$  since  $E(\hat{p}) = p$ ,  $\hat{p}$  is unbiased

**Definitions. (5 points each)** Clearly define each of the following terms.

1. Random variable:

numerical value from a random phenomenon

2. Statistic:

numerical value computed from a sample

3. Simple Random Sample:

sample taken from a pop'n in such a way that every combination of  $n$  individuals has an equal chance of being selected

**Multiple Choice. (3 points each)** Select the one best answer.

4. Undercoverage occurs when

- B
- a. a portion of the individuals in a sample refuse to provide information
  - ☒ b. a portion of the population is not included in the sampling plan
  - c. a portion of the individuals in a sample give incorrect or misleading information

5. If a statistic is shown to be a *consistent* estimate of the population mean, then

- B
- a. that statistic must be unbiased
  - ☒ b. it is possible for that statistic to be biased
  - c. that statistic can only be the sample mean,  $\bar{X}$ .

6. If two events A and B are independent, then

- B
- a. A and B are definitely disjoint
  - ☒ b. A and B are definitely not disjoint
  - c. A and B may or may not be disjoint; it depends on the setting

7. Least squares regression should **not** be used if

- A
- ☒ a. the residual plot shows a positive linear trend
  - b. both explanatory and response variables are quantitative
  - c. it is unclear which variable is explanatory and which is response

8. A scatterplot with correlation near 1 will lead to a least squares line with

- C
- a. a positive, very steep slope
  - b. a positive, very shallow slope
  - ☒ c. a positive slope, but there is not enough information to tell how steep it will be



**For the remaining problems, SHOW YOUR WORK. Numerical answers with no supporting work or explanation will receive zero credit, even if the calculations are trivial.**

9. Explain the difference between a simple random sample and a stratified random sample. Make up an example to help illustrate your answer (5 points)

In a stratified sample, we take a SRS from each of several groups (strata). For example, we often stratify by sex & take samples from men & women separately. In a SRS, the numbers of men & women in the sample would not be pre-selected, as they are in a stratified sample.

10. Explain the difference between an experiment and an observational study. Make up an example to help illustrate your answer (5 points)

In an experiment we assign treatments to individuals; in an observational study they "choose" their own.

We can imagine 2 ways of studying the effectiveness of the flu vaccine:

(1) randomly assign some subjects to get the vaccine, while the others do not [Experiment]

(2) Compare flu rates b/w individuals who chose to get the vaccine & those who didn't [Obs. Study]

11. What does the Central Limit Theorem tell us? Briefly explain why it is so important for the study of statistics. (5 points)

If  $\bar{X}$  is the sample mean computed from a SRS of size  $n$  from a pop'n w/ mean  $\mu$  & std dev  $\sigma$ , then

$\bar{X} \sim N(\mu, \sigma/\sqrt{n})$  if  $n$  is large.

Among many important impacts, the CLT tells us that if we take a <sup>large</sup> SRS from virtually any sample, we can plan to treat  $\bar{X}$  as having a normal dist'n - even if we know nothing about that pop'n's properties