

Exam 1

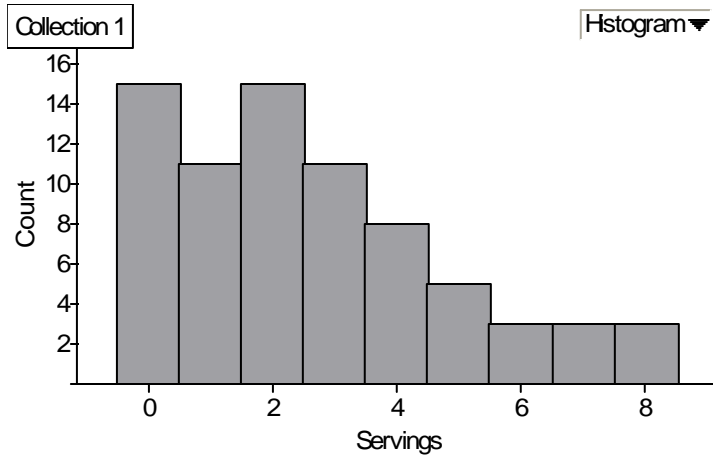
STA 2023, Fall 2007

Name _____

Instructions: Please show your work and clearly indicate your answers. For full credit, your work and/or explanation must support your answer. Always specify units. Provide explanations where requested!

I expect you to exhibit a level of individual academic integrity that is commensurate with being a part of the Honors College. Please acknowledge that integrity by signing the honor statement at the end of the test.

1. (14 points) The figure below shows the number of servings of fruit per day claimed by 74 seventeen-year-old girls in a study in Pennsylvania.



- a. Describe this distribution in words. Address the features of distributions discussed in class and in your text.
- b. What percent of the girls ate fewer than two servings of fruit per day?
- c. Compute the 5 number summary for these data.
2. (6 points) Each of the following statements contains a blunder. Explain in each case what is wrong.
- a. "There is a high correlation between the gender of American workers and their income."
- b. "We found high correlation ($r = 1.09$) between students' ratings of faculty teaching and ratings made by other faculty members."
- c. "The correlation between planting and yield of corn was found to be $r = 0.23$ bushel."

3. (15 points) Mechanical measurements on supposedly identical objects usually vary. The variation often follows a normal distribution. The stress required to break a type of bolt varies Normally with mean 75 kilopounds per square inch (ksi) and standard deviation 8.3 ksi.

a. Find the z-score for a bolt that breaks at a stress of 90 ksi.

b. What proportion of these bolts will withstand a stress of 90 ksi without breaking?

c. What range covers the middle 50% of breaking strengths for these bolts?

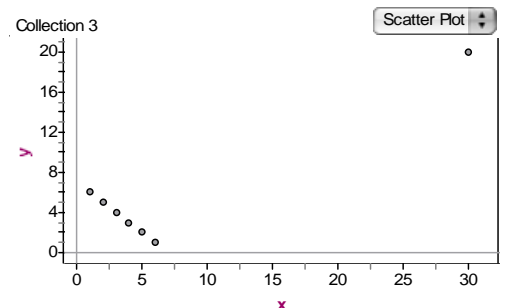
4. (6 points) We expect that students who do well on the midterm exam in a course will usually also do well on the final exam. Professor Smith looked at the exam scores of all 346 students who took his statistics class over a 10 year period. The least-square line for predicting the final exam score from midterm-exam score was $\hat{y} = 46.6 + 0.41x$.

Octavio scores 10 points above the class mean on the midterm. How many points above the class mean do you predict that he will score on the final? (Hint: Use the fact that the LSR line passes through the point (\bar{x}, \bar{y}) and the fact that Octavio's midterms score is $\bar{x} + 10$.)

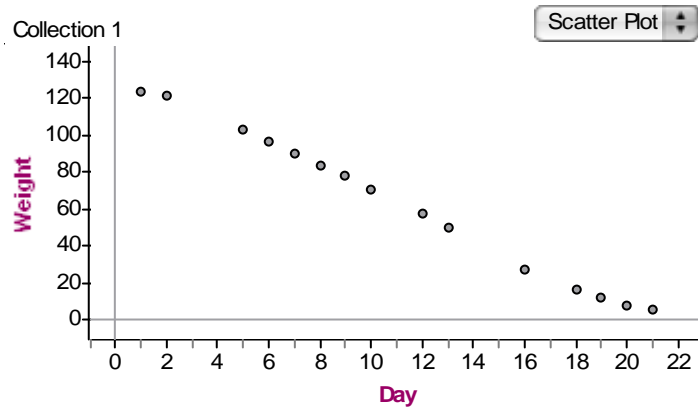
5. (5 points) A scatterplot is shown below.

- (a) The correlation is most likely (circle one):
- (i) negative, because most of the points lie on a downward sloping line
 - (ii) positive, because of the outlier
 - (iii) 0, because the outlier and the negative sloping points cancel each other
 - (iv) not defined in this situation

(b) Do you think the outlier is influential? Explain briefly.



6. (20 points) For three weeks, Rex Boggs weighed the bar of soap in his shower before showering almost every morning. A scatterplot of the data appears below, with weight in grams.



a. Describe the relationship between day and weight of the bar of soap. Address the features of scatterplots discussed in class and in your text.

b. The LSR line for weight on day is given by **Predicted Weight = -6.31 * Day + 133**. Write a sentence that carefully interprets the slope of the line in terms of the bar of soap. Give units!

c. Mr. Boggs did not weigh the soap on Day 17. What weight does the regression line predict on Day 17?

d. On Day 7, the weight Mr. Boggs' soap was 90 grams. Compute the residual for Day 7. What does the residual say about the relationship between the actual and predicted weights of the soap?

e. What does the regression equation predict for the weight of the soap on Day 30? Is the prediction reasonable? Explain briefly. What's the problem with using the LSR equation to predict the weight of the soap on Day 30?

f. The correlation between day and weight is -0.998 , the mean of the Day variable is 11.133 , and the standard deviation of the days is 6.523 . Using this information and what you know about the equation for the LSR line, what is the mean of the given weights? What is the standard deviation of the weights? Include units.

$$s_y = \underline{\hspace{2cm}}$$

$$\bar{y} = \underline{\hspace{2cm}}$$

h. Report the value of r^2 and write a sentence that carefully interprets this value in the context of Mr. Boggs' soap.

7. (12 points) Wabash Tech has two professional schools, business and law. The tables below show applicants to both schools, categorized by gender and admission decision, as well as the combined totals

Business School			Law School			Totals		
	Admit	Deny		Admit	Deny		Admit	Deny
Male	480	120	Male	10	90	Male	490	210
Female	180	20	Female	100	200	Female	280	220

- a. Calculate the percent of all male applicants that are admitted and the percent of all female applicants that are admitted.

males _____ females _____

- b. Now compute separately the percents of male and female applicants admitted by the business school and by the law school.

Business males _____ females _____

Law males _____ females _____

- c. What is interesting about your answers to the previous parts? This phenomenon has a name... what is it?

- d. Explain carefully, as if speaking to a skeptical reporter, how it can happen that Wabash appears to favor males when each school individually favors females. (In particular, what is going on behind the scenes?)

8. (8 points) Here are data on the numbers of degrees (in thousands) earned in 2005-2006.

	Female	Male
Associate	431	244
Bachelor	813	584
Master	298	215
Professional	42	47
Doctor	21	24

Write a few sentences describing the relationship between gender and level of degree earned. [Hint: Use the four step process of state, formulate, solve and conclude. Computing some marginal and conditional distributions may be useful!]

9. (10 points) The table to the right gives the volume of water discharged from the Mississippi River into the Gulf of Mexico for each year from 1954 to 2001. The units are cubic kilometers. The data have been reordered by discharge instead of year to make your computations easier.

a. Create a stemplot of discharges. Be sure to give a key that explains what your stems and leaves are.

b. Describe the distribution of discharges. Be sure to address the features of distributions discussed in class and in your text.

c. Would the mean and standard deviation or the five number summary be more appropriate summary statistics for these data? Explain briefly.

10. (4 points) In recent years, the mean SAT score of all high school seniors has increased. But if the seniors are divided into groups based on their grades (A students, B students, C students, and so on), the mean SAT score has decreased for each group. Explain how the lurking variable of grade inflation in high school can account for this pattern.

ye
195
196
195
196
200
196
195
197
198
198
197
195
198
196
196
197
196
198
196
195
197
198
198
196
198
198
199
197
199
198
199
199
197
198
198
197
197
199

Exam 2
STA 2023, Fall 2007

Name _____

Show your work on all problems. Answers with no work will receive no credit. When you are asked for conclusions or interpretations of your results, your response should say something about the original data. Include units whenever appropriate. If you use your calculator for something other than arithmetic, say what you did with your calculator and indicate what function(s) you used.

I expect you to exhibit a level of individual academic integrity that is commensurate with being a part of the Honors College. Please acknowledge that integrity by signing the honor statement at the end of the test.

1. (8 points) Suppose we must choose 4 addresses from a list of 100 addresses. We choose our sample as follows: Since we need a sample of size 4, we think of our list of 100 as four lists of size 25 ($=100/4$). Choose one address from the first list of 25 at random. For your sample, take the chosen address, together with the addresses 25, 50 and 75 places further down the list from it. For instance, if the random selection from the first list was the 13th address, the sample would be the 13th, 38th, 63rd and 88th addresses. This is called a *systematic random sample*.
 - (a) Use the random number table (Table B) starting on line 115 to choose a systematic random sample of 5 addresses from a list of 200. [Think of _____ lists of size _____!]
 - (b) Like a SRS, a systematic sample gives all individuals the same chance to be chosen. Explain why this is true (what's the probability in the situation of (a)?), then explain carefully why a systematic sample is nonetheless *not* a SRS.

2. (8 points) The Ministry of Health in the Canadian province of Ontario wants to know whether the national health care system is achieving its goals in the province. Much information about health care comes from patient records, but that source doesn't allow us to compare people who use health services with those who don't. So the Ministry of Health conducted the Ontario Health Survey, which interviewed a random sample of 61,239 people who live in Ontario.
 - (a) What is the population for this survey? What is the sample?

population:

sample:
 - (b) The survey found that 76% of males and 86% of females in the sample had visited a general practitioner at least once in the past year. These values are (circle one): parameters or statistics.
 - (c) Do you think these sample survey results are close to the truth about the entire population? Why?

3. (8 points) The most common treatment for breast cancer discovered in its early stages used to be removal of the breast. It is now usual to remove only the tumor and nearby lymph nodes, followed by radiation. To study whether these treatments differ in their effectiveness, a medical team examines the records of 25 large hospitals and compares the survival times after surgery of all women who have had either treatment.
- (a) Identify the explanatory variable:
the response variable:
- (b) Explain carefully why this study is not an experiment.
- (c) Explain why confounding will prevent this study from discovering which treatment is more effective. (The current treatment was in fact recommended after several large randomized comparative experiments.)
4. (10 points) You are designing an experiment to determine the effects of repeated exposure to an advertising message. A group of 36 undergraduate students was recruited to take part in the study. All subjects viewed a 40 minute program that included ads for a digital camera. Some subjects saw a 30 second commercial; others, a 90 second commercial. The same commercial was shown either 1, 3 or 5 times during the program. After viewing, all of the subjects answered questions about their attitude towards the camera and their intentions to purchase one.
- (a) What are the two factors in this experiment?
- (b) What are the levels for each factor?
- (c) How many treatment groups are there?
- (d) Outline your randomized comparative experiment using a diagram. Be sure to say where and how randomization is used.
- (e) Suppose you now realize that men and women often react differently to advertising, so you decide to use a block design with the two genders as blocks. The 36 subjects include 24 women and 12 men. Explain carefully how you will adjust your experimental design. As usual, be sure to say where randomization is used. You may use a diagram if you wish.

5. (8 points) Many random number generators allow users to specify the range of numbers to be produced. Suppose that you specify that Y can take any value between 0 and 4. Then the density curve of the outcomes has constant height between 0 and 4, and height 0 elsewhere.
- Is the random variable Y discrete or continuous? Why?
 - What is the height of the density curve between 0 and 4? Sketch the density curve.
 - Find the probability that Y is greater than 1.
6. (6 points) The numbers racket is a well-trenched illegal gambling operation in most large cities. One version works as follows: you choose one of the 1000 three-digit numbers 000 to 999 and pay your local numbers runner a dollar to enter your bet. Each day, one three-digit number is chosen at random and pays of \$600. (If your number is not chosen, you lose your \$1.) The mean payoff for the population of all daily bets is $\mu = 60$ cents.
- Is it possible to win 60 cents on any given day?
 - Joe makes one bet every day for many years. Explain what the law of large numbers says about Joe's results as he keeps on betting. (Will he make money or lose money in the long run?)
7. (10 points) Sheila's doctor is concerned that she may suffer from gestational diabetes (high blood glucose levels during pregnancy). There is variation both in the actual glucose level and in the blood test that measures the level. A patient is classified as having gestational diabetes if the glucose level is above 140 milligrams per deciliter (mg/dl) on hour after a sugary drink. Sheila's measured glucose level one hour after a sugary drink varies according to a Normal distribution with mean 125 mg/dl and standard deviation 10 mg/dl.
- If a single glucose measurement is made, what is the probability that Sheila is diagnosed as having gestational diabetes?
 - If measurements are made on 4 separate days and the mean result is compared with the 140 mg/dl criterion, what is the probability Sheila is diagnosed as having gestational diabetes?
 - What is the level L such that there is a probability of only 0.05 that the mean glucose level of 4 tests will fall above L ?

8. (18 points) In a test for ESP (extrasensory perception), a subject is told that cards the experimenter can see but he cannot contain either a star, a circle, a wave, or a square. As the experimenter looks at each of 20 cards in turn, the subject names the shape on each card.

- (a) If subject is just guessing, what is the probability she will guess correctly on any given card?
- (b) The number of correct guesses in 20 cards has what sort of distribution? (Give its name and parameters.)
- (c) Find the probability of a subject correctly guessing the correct symbol on exactly 6 of the cards.

(d) Use your answer to the previous part and the table below (you do NOT need to complete the table after filling in the answer from (c)!) to find the probability of guessing the correct symbol on **at least 7** of the 20 cards.

x	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$P(X=x)$.003	.021	.067	.134	.190	.202															

- (e) What is the mean of the distribution of the correct number of guesses in 20 cards?
- (f) What is the standard deviation of the distribution of the correct number of guesses in 20 cards?
- (g) Use the normal approximation to find the probability of at least 7 correct guesses. Does it seem like this is a reasonable approximation to your answer in (d)?

(h) A subject claims to have ESP and you believe him. To test the claim, you will conduct a hypothesis test. What are the null and alternative hypotheses for the test (they should relate to this subject's average number of correct identifications from 20 cards in many repetitions of the experiment)?

H_0 :

H_a :

(i) If the subject correctly identified the symbols on 7 of 20 cards correctly, would you believe his claim of having ESP? Explain briefly how your answer is related to your computation in (d). (Is it likely or unlikely that he would get 7 or more correct by guessing alone?)

9. (16 points) A class survey in a large class for first-year college students asked, “About how many minutes do you study on a typical weeknight?” The mean response of the 269 students was 137. Suppose that we know that the study time follows a Normal distribution with standard deviation $\sigma = 65$ minutes in the population of all first-year students at this university.
- (a) Give a 95% confidence interval for the mean study time of all first year students.
- (b) Write a complete English sentence that gives a practical interpretation of your confidence interval in terms of students and study times.
- (c) There were actually 270 responses to the class survey, but one student claimed to study 30,000 minutes per night. We know this student was joking, so we left out this value. However, if we left it in, we would get a sample mean of 248 minutes for all 270 students. Now compute a 95% confidence interval for the population mean (assuming $\sigma = 65$ minutes, as before).
- (d) Which of our three assumptions are we violating when we do the computation in (c)?
- (e) If you wanted to compute a 97% confidence interval, what critical value (z^*) would you use? (You need not compute the confidence interval.)
- (f) Would your 97% confidence interval be wider or narrower than your 95% confidence interval? Explain briefly.
- (g) Suppose you wanted a 97% confidence interval with the same margin of error as you found in (a). How big would your sample need to be?

10. (8 points) A poker player holds a flush when all 5 cards in the hand belong to the same suit (clubs, diamonds, hearts, or spades). We will compute the probability of being dealt a flush in several steps. Assume we have a full deck of 52 cards (13 in each suit) that have been randomly shuffled.
- (a) What is the probability that the first card dealt is a spade?
 - (b) What is the probability that the second card dealt is a spade, given that the first one was a spade?
 - (c) What is the probability that the first five cards dealt are all spades? (Hint: You should multiply together 5 probabilities, two of which were answers to the previous parts.)
 - (d) The probability of being dealt 5 hearts or 5 diamonds or 5 clubs is the same as the probability of being dealt 5 spades. What is the probability of being dealt a flush?

BONUS (6 points) The unique colors of cashmere sweaters your firm makes result from heating undyed yarn in a kettle with a dye liquor. The pH (acidity) of the liquor is critical for regulating dye uptake and hence the final color. There are 5 kettles, all of which receive dye liquor from a common source. Twice each day, the pH of the liquor in each kettle is measured, giving a sample of size 5. The process has been operating in control with $\mu = 4.22$ and $\sigma = 0.127$.

- (a) Give the center line and control limits for the \bar{x} chart.
- (b) What are the natural tolerances for the individual pH measurements?
- (c) Explain what an \bar{x} control chart is used for.

BONUS (4 points) Choose a point at random in the square with sides $0 < x < 1$ and $0 < y < 1$. The probability that the point falls in any region within the square is equal to the area of that region. Let X be the x coordinate and Y the y coordinate of a randomly chosen point. Find the conditional probability $P(Y < 0.5 \mid Y > X)$. [Hint: Draw a diagram of the square and the events $Y < 0.5$ and $Y > X$.

I have adhered to the principles of the Honor Code in completing this test. _____

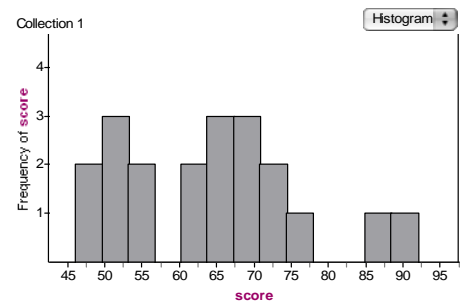
Exam 3
STA 2023, Fall 2007

Name _____

Instructions: For questions that call for calculations, present your method of solution in a clear, well-labeled manner and show the details of your calculations. For questions that ask for interpretations and explanations, explain your answers fully unless instructed otherwise and always explain your answers *in the context of the original data*. Include units whenever appropriate. For problems with multiple parts, be aware that you can usually complete later parts successfully regardless of whether or not you have correctly answered earlier parts.

I expect you to exhibit a level of individual academic integrity that is commensurate with being a part of the Honors College. Please acknowledge that integrity by signing the honor statement at the end of the test.

1. A SRS of 20 third grade children is selected in Chicago and each is given a test to measure his or her reading ability. In the sample, the mean score is 64 points and the standard deviation is 12 points. A histogram of the scores is shown below.



(a) (2 points) What is the population in this study?

(b) (3 points) What is the parameter of interest in this study?

(c) (5 points) What conditions must be satisfied if we want to construct a confidence interval for the parameter you identified above? Comment on the degree to which each is satisfied.

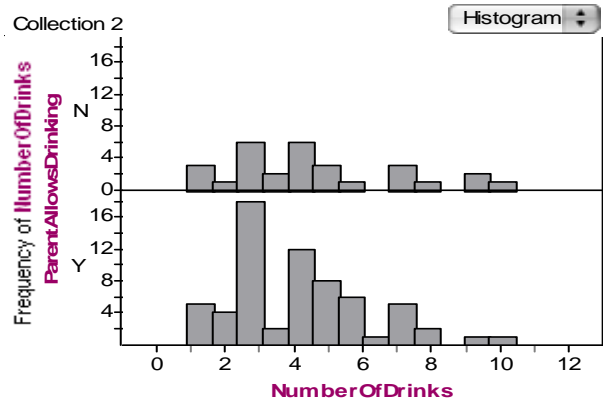
(d) (6 points) Construct a 90% confidence interval for parameter you identified above.

(e) (4 points) Write a sentence that interprets your confidence interval in the context of grade school children.

2. A professor asked her sophomore students, “Does either of your parents allow you to drink alcohol around him or her?” and “How many drinks do you typically have per session? (A drink is defined as one 12 oz beer, one 4 oz glass of wine, or one 1 oz shot of liquor.)” Summary statistics for the average number of drinks per session for the female students who were not abstainers are given below. We are willing to treat the students in the study as a SRS of sophomore students at this particular college.

Collection 2			
	ParentAllowsDrinking		Row Summary
	N	Y	
NumberOfDrinks	4.5517241	4.1769231	4.2925532
	2.4251098	2.0260622	2.1507808
	29	65	94

S1 = mean ()
 S2 = stdDev ()
 S3 = count ()



- (a) (18 points) Does the behavior of the parents make a significant difference in how many drinks students have on average? To answer this question, check the assumptions necessary for inference (say what you checked and how!), state your hypotheses in both symbols and words, compute the test-statistic and p-value, and state your conclusion in a complete sentence.

- (b) (6 points) We wonder what proportion of female students at this school have at least one parent who allows the student to drink around him or her. Give a 95% confidence interval for this proportion, based on the data.

(c) (6 points) How many female students would have to be surveyed in order to estimate the population proportion with 95% confidence and a margin of error of at most 0.03?

(d) (6 points) A 95% confidence interval for the difference in mean number of drinks per session between all sophomore **males** and all sophomore **females** at this college (men – women) is (0.2, 3.5). Write a sentence or two that clearly interprets this confidence interval for someone who does not know a lot of statistics.

3. (15 points) The water diet requires you to drink two cups of water every half-hour from when you get up until you go to bed, but otherwise allows you to eat whatever you like. Four adult volunteers agree to test the diet. They are weighed prior to beginning the diet and after six weeks on the diet. The weights (in pounds) are

Person	1	2	3	4
Weight before the diet	180	125	240	150
Weight after six weeks	170	130	215	152

For the population of all adults, assume that the weights both before and after the diet are normally distributed. Do these data provide evidence that the diet leads to weight loss? Assuming the necessary assumptions are satisfied, set up the appropriate hypothesis test, compute the test statistic and p-value, and clearly state your conclusion in a complete sentence (in the context of the diet).

4. (12 points) A company medical director tests the hypothesis that the mean blood pressure for the population of executives in his multinational company differs from the national mean of $\mu = 128$ is tested at a significance level of 0.01. The power of this test against the alternative hypothesis $\mu = 134$ is 0.81.
- Explain what a Type I error would be in this situation.
 - What is the probability of making a Type I error?
 - Explain what a Type II error would be in this situation.
 - What is the probability of a Type II error against the alternative hypothesis $\mu=134$?
5. (3 points) You plan to construct a confidence interval for the mean μ of a normal population with (known) standard deviation σ . Which of the following will reduce the size of the margin of error?
- Use a lower level of confidence.
 - Increase the sample size.
 - A and B
 - Neither A nor B
6. (3 points) A certain population follows a normal distribution with mean μ and standard deviation σ . You collect data and test the hypotheses $H_0: \mu = 1$, $H_a: \mu \neq 1$. You obtain a P -value of 0.022. Which of the following is true?
- A 95% confidence interval for μ will include the value 1.
 - A 95% confidence interval for μ will include the value 0.
 - A 99% confidence interval for μ will include the value 1.
 - A 99% confidence interval for μ will include the value 0.
7. (8 points) A radio talk show host with a large audience is interested in the proportion \hat{p} of adults in his listening area that think the drinking age should be lowered to 18. To find this out he poses the following question to his listeners: "Do you think that the drinking age should be reduced to 18 in light of the fact that 18-year-olds are eligible for military service?" He asks listeners to phone in and vote "yes" if they agree the drinking age should be lowered and "no" if they do not. The proportion \hat{p} of those who phoned in and answered yes is $\hat{p} = 0.70$ and the standard error $SE_{\hat{p}}$ of the proportion is 0.0459.
- 8.
- How many people phoned in?
 - 50
 - 100
 - 200
 - We cannot know how many people phoned in based on the information provided.
 - A large sample 95% confidence interval based on these data is approximately (0.60, 0.80). Are we 95% confident that the true proportion of adults in the listening area believe the drinking age should be lowered? Carefully explain your answer.
9. (3 points) What is the +4 method? When and why should you use it?

