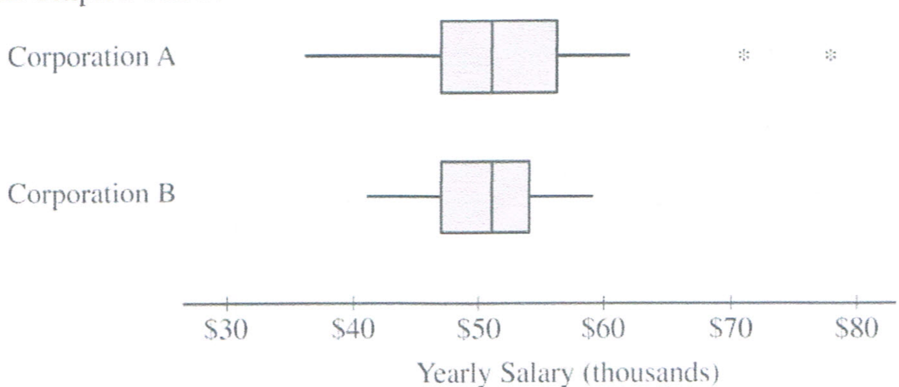
STATISTICS
SECTION II
Part A
Questions 1-5

Spend about 65 minutes on this part of the exam.

Percent of Section II score—75

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

1. Two large corporations, A and B, hire many new college graduates as accountants at entry-level positions. In 2009 the starting salary for an entry-level accountant position was $36,000 a year at both corporations. At each corporation, data were collected from 30 employees who were hired in 2009 as entry-level accountants and were still employed at the corporation five years later. The yearly salaries of the 60 employees in 2014 are summarized in the boxplots below.



(a) Write a few sentences comparing the distributions of the yearly salaries at the two corporations.

(b) Suppose both corporations offered you a job for $36,000 a year as an entry-level accountant.

(i) Based on the boxplots, give one reason why you might choose to accept the job at corporation A.

(ii) Based on the boxplots, give one reason why you might choose to accept the job at corporation B.

a) The yearly salary distributions for both corporations are roughly symmetric with similar medians. The salaries at corporation A are more variable, since corporation A has a greater range and IQR than B, and also has two outliers whereas B has none.

bi) I might choose A because I would have greater opportunity to earn a very high salary there. At least 3 of the 30 employees surveyed there make over $60k/yr, while none of the 30 at B make that much.

bii) I might choose B because I would be more likely to earn at least a modest raise. All 30 of those surveyed at B earn over $40k, which is not true at A.

GO ON TO THE NEXT PAGE.

2. To increase business, the owner of a restaurant is running a promotion in which a customer's bill can be randomly selected to receive a discount. When a customer's bill is printed, a program in the cash register randomly determines whether the customer will receive a discount on the bill. The program was written to generate a discount with a probability of 0.2, that is, giving 20 percent of the bills a discount in the long run. However, the owner is concerned that the program has a mistake that results in the program not generating the intended long-run proportion of 0.2.

The owner selected a random sample of bills and found that only 15 percent of them received discounts. A confidence interval for p, the proportion of bills that will receive a discount in the long run, is 0.15 ± 0.06. All conditions for inference were met.

(a) Consider the confidence interval 0.15 ± 0.06.

(i) Does the confidence interval provide convincing statistical evidence that the program is not working as intended? Justify your answer.

(ii) Does the confidence interval provide convincing statistical evidence that the program generates the discount with a probability of 0.2 ? Justify your answer.

A second random sample of bills was taken that was four times the size of the original sample. In the second sample 15 percent of the bills received the discount.

(b) Determine the value of the margin of error based on the second sample of bills that would be used to compute an interval for p with the same confidence level as that of the original interval.

(c) Based on the margin of error in part (b) that was obtained from the second sample, what do you conclude about whether the program is working as intended? Justify your answer.

a i) No. The confidence interval is a set of plausible values for p. Since 0.2 is in the interval, it is plausible that the machine is working as intended.

ii) No. The confidence interval includes many values other than 0.2. for example, 0.1 is a plausible value, meaning it is plausible that the machine prints only half as many discounts as intended.

b) $\frac{0.06}{\sqrt{4}} = 0.03$

c) The interval 0.15 ± 0.03 does not contain 0.2. Therefore, since 0.2 is not a plausible value for p, it appears that the program is not working as intended.

3. A shopping mall has three automated teller machines (ATMs). Because the machines receive heavy use, they sometimes stop working and need to be repaired. Let the random variable X represent the number of ATMs that are working when the mall opens on a randomly selected day. The table shows the probability distribution of X.

Number of ATMs working when the mall opens	0	1	2	3
Probability	0.15	0.21	0.40	0.24

(a) What is the probability that at least one ATM is working when the mall opens?

(b) What is the expected value of the number of ATMs that are working when the mall opens?

(c) What is the probability that all three ATMs are working when the mall opens, given that at least one ATM is working?

(d) Given that at least one ATM is working when the mall opens, would the expected value of the number of ATMs that are working be less than, equal to, or greater than the expected value from part (b) ? Explain.

a) $1 - P(0) = 1 - 0.15 = 0.85$

b) $0.15(0) + 0.21(1) + 0.40(2) + 0.24(3) = 1.73$

c) $\frac{0.24}{0.85} \approx 0.2824$

d) Greater. The new expected value calculation would use $P(0) = 0$ and the probabilities of $x = 1, 2,$ and 3 would increase proportionately (by a factor of $1/0.85$).

4. A researcher conducted a medical study to investigate whether taking a low-dose aspirin reduces the chance of developing colon cancer. As part of the study, 1,000 adult volunteers were randomly assigned to one of two groups. Half of the volunteers were assigned to the experimental group that took a low-dose aspirin each day, and the other half were assigned to the control group that took a placebo each day. At the end of six years, 15 of the people who took the low-dose aspirin had developed colon cancer and 26 of the people who took the placebo had developed colon cancer. At the significance level $\alpha = 0.05$, do the data provide convincing statistical evidence that taking a low-dose aspirin each day would reduce the chance of developing colon cancer among all people similar to the volunteers?

$H_0: P_{aspirin} - P_{no\ aspirin} = 0 \quad H_a: P_{aspirin} - P_{no\ aspirin} < 0$

2-proportion z test using $\alpha = 0.05$

subjects were randomly assigned to treatment groups

observations and treatment groups were independent of each other

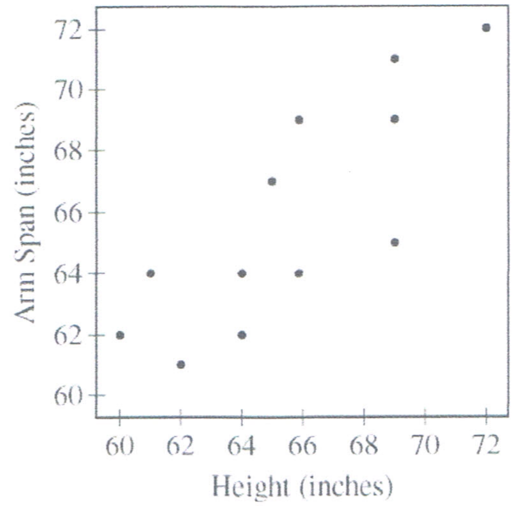
15, 500-15, 26, 500-26 all greater than 10

$z = -1.754 \qquad p = 0.0397$

Since $p < \alpha$, we have evidence that taking a low-dose aspirin each day would reduce the chance of developing colon cancer among all people similar to the volunteers.

GO ON TO THE NEXT PAGE.

5. A student measured the heights and the arm spans, rounded to the nearest inch, of each person in a random sample of 12 seniors at a high school. A scatterplot of arm span versus height for the 12 seniors is shown.

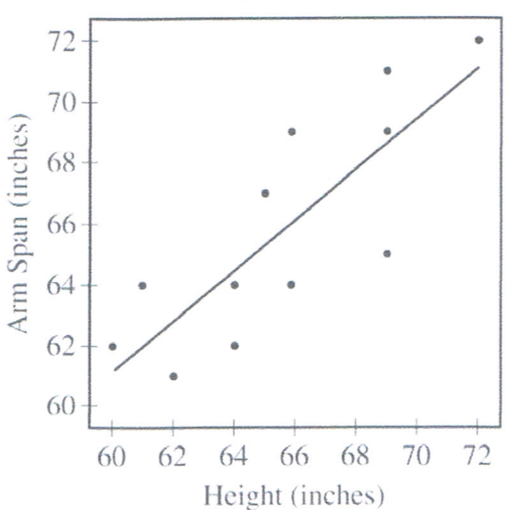


(a) Based on the scatterplot, describe the relationship between arm span and height for the sample of 12 seniors.

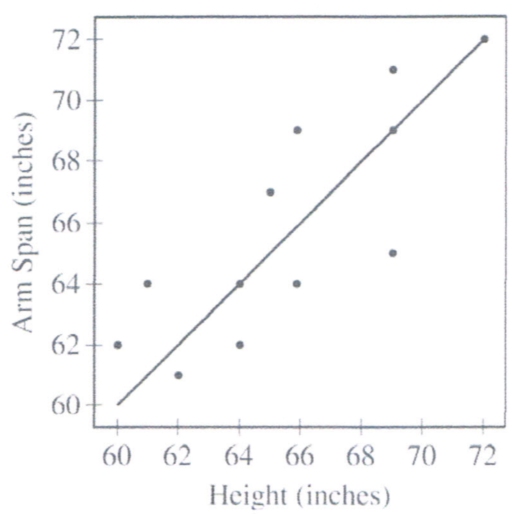
There is a moderate, positive, linear relationship between arm span & height for the 12 seniors in the sample.

Let x represent height, in inches, and let y represent arm span, in inches. Two scatterplots of the same data are shown below. Graph 1 shows the data with the least squares regression line $\hat{y} = 11.74 + 0.8247x$, and graph 2 shows the data with the line $y = x$.

GRAPH 1



GRAPH 2



(b) The criteria described in the table below can be used to classify people into one of three body shape categories: square, tall rectangle, or short rectangle.

Square	Tall Rectangle	Short Rectangle
Arm span is equal to height.	Arm span is less than height.	Arm span is greater than height.

(i) For which graph, 1 or 2, is the line helpful in classifying a student's body shape as square, tall rectangle, or short rectangle? Explain. Graph 2. Points above the line have a greater y-coordinate (arm span) than x-coordinate (height). The opposite is true for points below the line, whereas points on the line represent equal arm span + height.

(ii) Complete the table of classifications for the 12 seniors.

Classification	Square	Tall Rectangle	Short Rectangle
Frequency	3	4	5

(c) Using the best model for prediction, calculate the predicted arm span for a senior with height 61 inches.

$\hat{y} = 11.74 + 0.8247(61) \approx 62.0467$ inches

STATISTICS

SECTION II

Part B

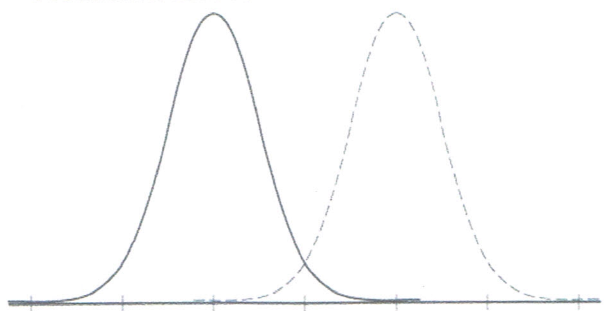
Question 6

Spend about 25 minutes on this part of the exam.

Percent of Section II score—25

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

6. Corn tortillas are made at a large facility that produces 100,000 tortillas per day on each of its two production lines. The distribution of the diameters of the tortillas produced on production line A is approximately normal with mean 5.9 inches, and the distribution of the diameters of the tortillas produced on production line B is approximately normal with mean 6.1 inches. The figure below shows the distributions of diameters for the two production lines.



The tortillas produced at the factory are advertised as having a diameter of 6 inches. For the purpose of quality control, a sample of 200 tortillas is selected and the diameters are measured. From the sample of 200 tortillas, the manager of the facility wants to estimate the mean diameter, in inches, of the 200,000 tortillas produced on a given day. Two sampling methods have been proposed.

Method 1: Take a random sample of 200 tortillas from the 200,000 tortillas produced on a given day. Measure the diameter of each selected tortilla.

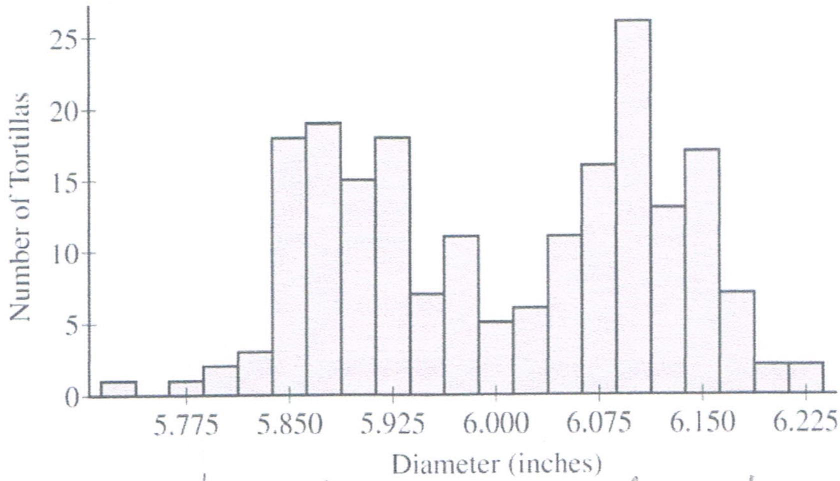
Method 2: Randomly select one of the two production lines on a given day. Take a random sample of 200 tortillas from the 100,000 tortillas produced by the selected production line. Measure the diameter of each selected tortilla.

(a) Will a sample obtained using Method 2 be representative of the population of all tortillas made that day, with respect to the diameters of the tortillas? Explain why or why not.

No. All would come from one line, and the tortillas from one line are generally not the same size as those from the other.

GO ON TO THE NEXT PAGE.

(b) The figure below is a histogram of 200 diameters obtained by using one of the two sampling methods described. Considering the shape of the histogram, explain which method, Method 1 or Method 2, was most likely used to obtain a such a sample.



Method 1. Since a large majority of the tortillas from line A are under 6", and a large majority from line B are over 6", this histogram almost certainly has tortillas from both lines.

(c) Which of the two sampling methods, Method 1 or Method 2, will result in less variability in the diameters of the 200 tortillas in the sample on a given day? Explain.

Method 2. All tortillas would be close to the mean of the selected line, and not spread over both means as in method 1.

Each day, the distribution of the 200,000 tortillas made that day has mean diameter 6 inches with standard deviation 0.11 inch.

(d) For samples of size 200 taken from one day's production, describe the sampling distribution of the sample mean diameter for samples that are obtained using Method 1.

approximately normal, $\mu_{\bar{x}} = 6$, $\sigma_{\bar{x}} = \frac{0.11}{\sqrt{200}}$

(e) Suppose that one of the two sampling methods will be selected and used every day for one year (365 days). The sample mean of the 200 diameters will be recorded each day. Which of the two methods will result in less variability in the distribution of the 365 sample means? Explain.

Method 1. Nearly all will be within $3\sigma_{\bar{x}} \approx 0.0233$ of 6.0, while with method 2, the range would almost certainly be greater than $6.1 - 5.9 = 0.2$.

(f) A government inspector will visit the facility on June 22 to observe the sampling and to determine if the factory is in compliance with the advertised mean diameter of 6 inches. The manager knows that, with both sampling methods, the sample mean is an unbiased estimator of the population mean. However, the manager is unsure which method is more likely to produce a sample mean that is close to 6 inches on the day of sampling. Based on your previous answers, which of the two sampling methods, Method 1 or Method 2, is more likely to produce a sample mean close to 6 inches? Explain.

Method 1. As stated in (e), nearly all sample means using that method will lie within $3\sigma_{\bar{x}} \approx 0.0233$ inches of 6. Method 2, on the other hand, would give a sample mean either close to 5.9 or close to 6.1. Furthermore, the distance from such a sample mean to the population mean for the selected line would likely be smaller than the distance from a mean using method 1 to 6.0, as explained in (c).

STOP

END OF EXAM