Name: Key Per: ____ Date: _____

Final Exam Review—Semester 1 (Ch.1-8)

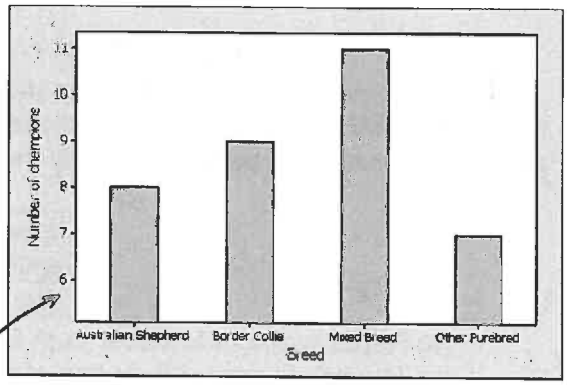
AP Statistics

Ch.1—Exploring Data

Part 1: Multiple Choice. *Circle the letter corresponding to the best answer.*

d **1.** The bar graph at right shows the distribution of breeds for all the champions of the annual World Canine Disc Championships from 1975 to 2009. Which statement can be made on the basis of this graph?

(a) Mixed breed dogs have won the championship about twice as often as Australian Shepherds.
(b) Most of the mixed breed dogs were at least half Border Collie.
(c) None of the champion dogs were Labrador Retrievers.
(d) The graph exaggerates the difference between the number of champions of each breed category.
(e) Border Collies are larger dogs than Australian Shepherds.



Starts @ 6, but the range is really only 4 (11-7).

C **2.** You measure the age (years), weight (pounds), and marital status (single, married, divorced, or widowed) of 1400 women. How many variables did you measure?

(a) 1403
(b) 1400
(c) 3
(d) 2
(e) 1

age = x
weight = y
marital status = z

d **3.** The population of the United States is aging, though less rapidly than in other developed countries. To the right is a stemplot of the percent of residents aged 65 and older in each of the 50 states, according to the 2000 census. There are two outliers: Alaska has the lowest percent of older residents, and Florida has the highest. What is the percent for Florida?

(a) 13.8%
(b) 57%
(c) 176%
(d) 17.6%
(e) 5.7%

17|6 = 17.6%

```
 5 | 7
 6 |
 7 |
 8 | 5
 9 | 679
10 | 6
11 | 02233667
12 | 00111134457|89
13 | 00012233345568
14 | 034579
15 | 36
16 |
17 | 6
```

Key: 9|7 = 9.7%

Q_1 Q_3 median

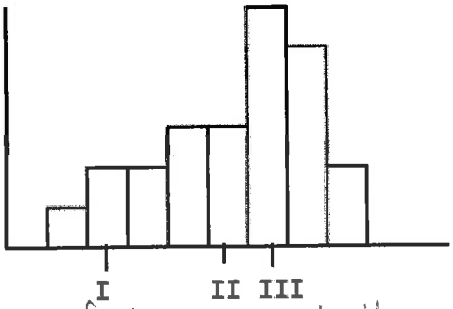
C **4.** The Interquartile range for the distribution of 50 states in the previous question is:

(a) 11.7% to 13.5%
(b) 117% to 135%
(c) 1.8%
(d) 2.7%
(e) 18%

$IQR = Q_3 - Q_1$
$= 13.5 - 11.7$
$= 1.8$

e **5.** For the histogram below, what is the proper ordering of the mean and median? Note that the graph is NOT numerically precise—only the relative positions are important.

(a) I is the mean and II is the median.
(b) II is the median and III is the mean.
(c) I is the median and II is the mean.
(d) I is the mean and III is the median.
(e) II is the mean and III is the median.



The graph is skewed left, so the mean is pulled away from the median in the direction of the long tail.

d **6.** A review of voter registration records in a small town yielded the following table of the number of males and females registered as Democrat, Republican, or some other affiliation. Which of the following conclusions seems to be supported by the data?

	Male	Female	
Democrat	300 30%	600 60%	900
Republican	500 50%	300 30%	800
Other	200 20%	100 10%	300
	1000	1000	2000

(a) Republicans outnumber both Democrats and "Other."
(b) The conditional distribution of party affiliation for males is 1100.
(c) The marginal distribution of party affiliation is 1000, 1000.
(d) There is an obvious association between gender and political party registration.
(e) It is unclear whether there is an association between gender and political party registration.

c **7.** A sample of 99 distances has a mean of 24 feet and a median of 24.5 feet. Unfortunately, it has just been discovered that the maximum value in the distribution, which was erroneously recorded as 40, actually had a value of 50. If we make this correction to the data, then

(a) the mean remains the same, but the median is increased.
(b) the mean and median remain the same.
(c) the median remains the same, but the mean is increased.
(d) the mean and median are both increased.
(e) we do not know how the mean and median are affected without further calculations, but the variance is increased.

b **8.** Mr. Yates picked up a dozen items in the grocery store with a mean cost of $3.25. Then he added an apple pie for $6.50. The new mean for all 13 items is

(a) $3.00
(b) $3.50
(c) $3.75
(d) $4.88
(e) None of the above

$3.25 \times 12 = 39$ (sum of 12 items)

$39 + 6.50 = 45.5$ (sum of 13 items)

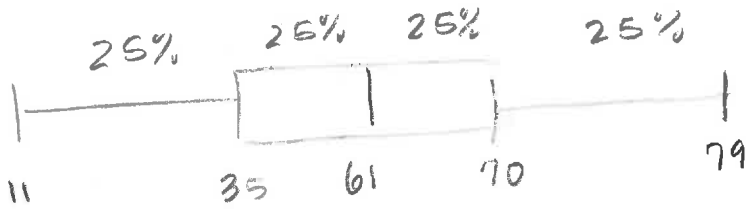
$\frac{45.5}{13} = 3.50$

d **9.** A small company estimating its photocopying expenses finds that the mean number of copies made per day for the past 12 months is 258 copies per day with a standard deviation of 24 copies per day. Which of the following is a correct interpretation of standard deviation?

(a) The number of copies made per day was always between 234 and 282.
(b) About 95% of the time, the number of copies made per day was between 234 and 282.
(c) The difference between the mean number of copies made per day and the median number of copies made per day was 24.
(d) On average, the number of copies made each day was about 24 copies per day away from the mean, 258.
(e) 1.5 times the interquartile range of copies made per day is 24.

C **10.** The five-number summary for scores on a statistics exam is 11, 35, 61, 70, 79. In all, 380 students took the test. About how many had scores between 35 and 61?

(a) 26
(b) 76
(c) 95
(d) 190
(e) None of these

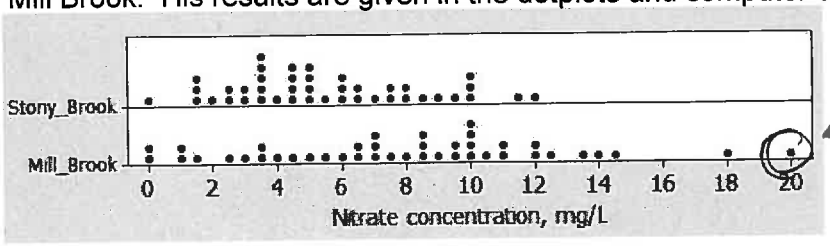


25% of 380
= 95

Part 2: Free Response

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

11. Nitrates are organic compounds that are a substantial component of agricultural fertilizers. When those fertilizers run off into streams, the nitrates can have a toxic effect on animals that live in those streams. An ecologist studying nitrate pollution in two streams collects data on nitrate concentrations at 42 places on Stony Brook and 42 places on Mill Brook. His results are given in the dotplots and computer output below.



Variable	n	Mean	SE Mean	StdDev	Min	Q1	Median	Q3	Max
Stony Brook	42	5.524	0.451	2.922	0.000	3.500	5.000	7.500	12.000
Mill Brook	42	7.929	0.710	4.607	0.000	4.500	8.250	10.500	20.000

a.) Determine if there are any outliers in each distribution. Show your work.

Stony Brook: $Q_1 - 1.5IQR = 3.5 - 1.5(4) = -2.5$ (lower limit), min = 0 → no low outliers

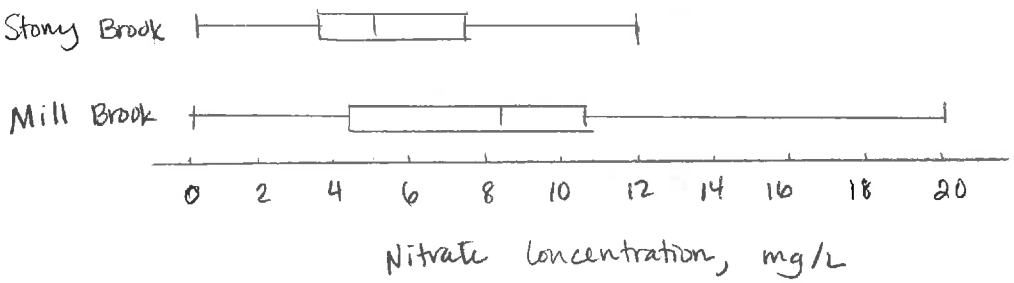
$Q_3 + 1.5IQR = 7.5 + 1.5(4) = 13.5$ (upper limit), max = 12 → no high outliers

Mill Brook: $Q_1 - 1.5IQR = 4.5 - 1.5(6) = -4.5$ (lower limit), min = 0 → no low outliers

$Q_3 + 1.5IQR = 10.5 + 1.5(6) = 19.5$ (upper limit), max = 20 → 1 high outlier

There are no outliers for Stony Brook, but one high outlier for Mill Brook, 20.

b.) Draw parallel boxplots of these two distributions. Be sure to label the plots and provide a scale.



c.) Write a few sentences comparing the nitrate concentrations in Stony Brook and Mill Brook.

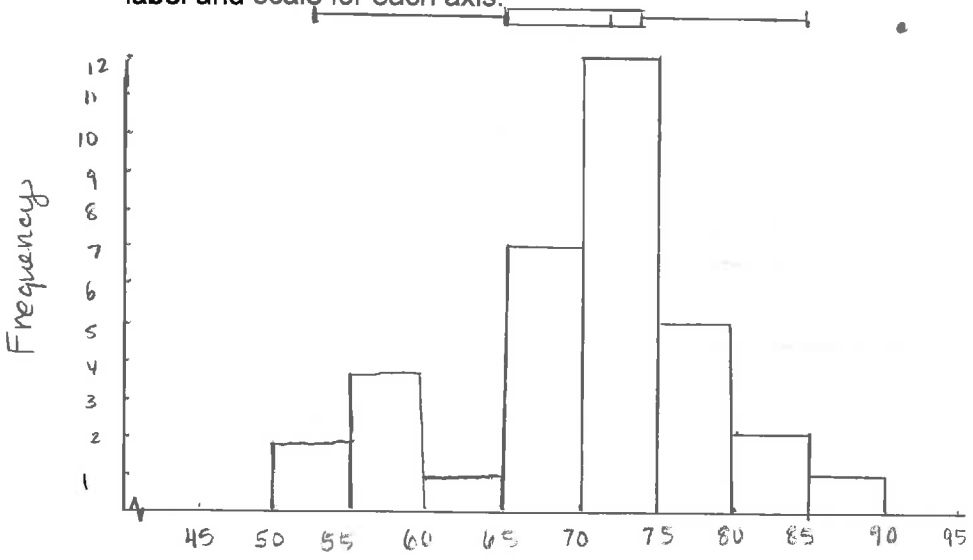
(Since both distributions are skewed right, use median-based summaries.)

The median for Stony is less than the median for Mill, 5.524 vs. 7.929. Both distributions are skewed right, Mill more than Stony. This is apparent because the IQRs (9 vs. 6), ranges (20 vs 12), and St. deviations (4.6 vs. 2.92) are all larger for Mill compared to Brook. Mill also has one high outlier. Overall, Mill appears to have higher concentrations because all numbers in the descriptive statistics, except min, is larger for Mill than Stony.

12. The following data represent scores of 34 students on a calculus test.

72	59	73	72	74	74	61	67	67
72	78	80	57	56	76	72	54	65
88	74	57	83	68	79	73	71	
70	65	67	76	67	72	76	53	

a.) Construct a histogram for this distribution. Choose an appropriate bin width, and be sure to provide a label and scale for each axis.



$\bar{x} = 69.64$ $s_x = 8.35$

{53, 65, 72, 74, 88}

Outliers:

$Q_1 - 1.5IQR$

$65 - 1.5(9) = 51.5$ ✓

no low outliers

$Q_3 + 1.5IQR$

$74 + 1.5(9) = 87.5$ ×

one high outlier, 88.

b.) Based on your histogram, what numerical measures of center and spread would be best to use for this distribution? Explain your choice.

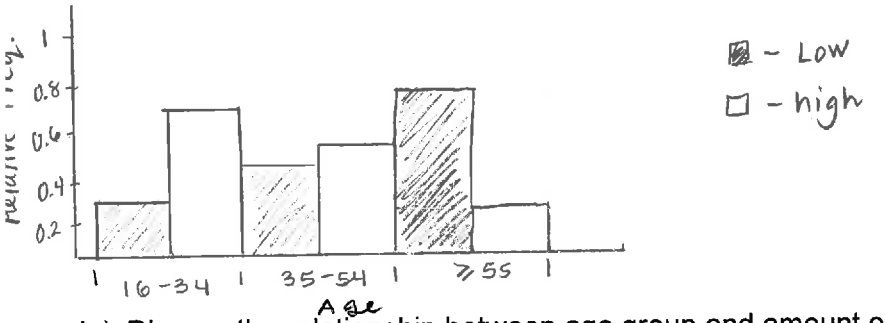
The distribution of test scores is skewed right with one high outlier, 88. Because there is at least one extreme value and the graph is skewed, the mean and median are not approximately equal ($\bar{x} = 69.6$ and $M = 72$). The median and 5-number summary will be best for describing the data because they are resistant to outliers and skew.

13. In a study of the relationship between the amount of violence a person watches on TV and the viewer's age, 81 regular TV watchers were randomly selected and classified according to their age group and whether they were a "low-violence" or "high-violence" viewer. Here is a two-way table of the results.

	Age Group			
Violence Watched	16-34	35-54	55 and over	Total
Low	8	12	21	41
High	18	15	7	40
Total	26	27	28	81

a.) Calculate three conditional distributions for violence watched among each age group. You may present your results in either a table or a graph.

	Age		
	16 - 34	35-54	≥ 55
LOW	8/26 = 31%	12/27 = 44%	21/28 = 75%
HIGH	18/26 = 69%	15/27 = 56%	7/28 = 25%



b.) Discuss the relationship between age group and amount of violence watched in two or three sentences.

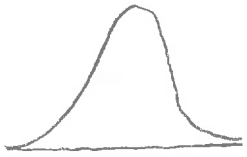
For the 16-34 age group, 69% were "high-violence" viewers compared to 31% being "low-violence" viewers. For the 35-54 age group, the percentage gap narrows but the majority is still "high-violence" viewers; 56%. 44% are considered "low-violence" viewers. For the 55 and older age group, the rolls have reversed. The majority prefer "low-violence" tv, at 75%, and 25% are considered "high-violence" viewers. Overall, it seems as though as age increases, the amount of violence watched on tv decreases.

CH.2—DESCRIBING LOCATION IN DISTRIBUTIONS

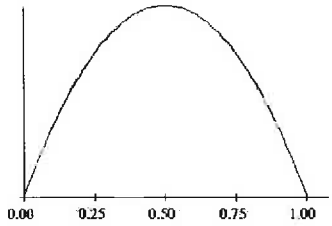
Part 1: Multiple Choice. *Circle the letter corresponding to the best answer.*

d **1.** For the density curve shown, which statement is true?

(a) The density curve is Normal.
(b) The density curve is skewed right.
(c) The density curve is skewed left.
(d) The density curve is symmetric
(e) None of the above is correct.



Normal



Symmetric

C **2.** For the density curve shown in Question 1, which statement is true?

(a) The mean is greater than the median.
(b) The mean is less than the median.
(c) The mean and median are equal.
(d) The mean could be either greater than or less than the median.
(e) None of the above is correct.

C **3.** Suppose that 16-ounce bags of chocolate chip cookies are produced with weights that follow a Normal distribution with mean weight 16.1 ounces and standard deviation 0.1 ounce. The percent of bags that will contain between 16.0 and 16.1 ounces is about

(a) 10 (b) 16 (c) 34 (d) 68 (e) None are correct. $N(16.1, 0.1)$

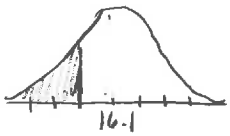
normalcdf or 68-95-99.7 Rule

b **4.** For the distribution of cookie bags described in Question 3, approximately what percent of the bags will likely be underweight (that is, less than 16 ounces)?

(a) 10 (b) 16 (c) 32 (d) 64
(e) none of the above

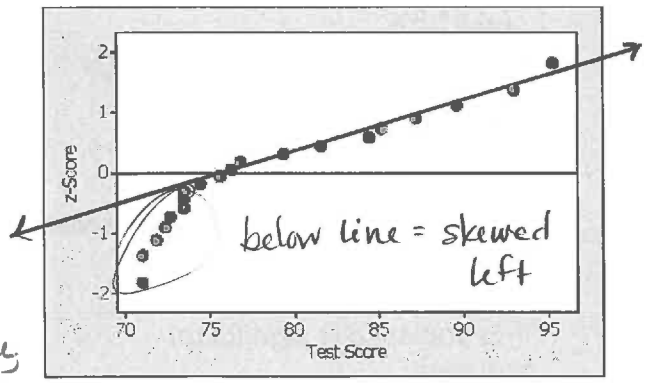
$N(16.1, 0.1)$

$P(X < 16)$ normalcdf



C **5.** The plot shown at the right is a Normal probability plot for a set of test scores. Which statement is true for these data?

(a) The data are clearly Normally distributed.
(b) The data are approximately Normally distributed.
(c) The data are clearly skewed to the left.
(d) The data are clearly skewed to the right.
(e) There is insufficient information to determine the shape of the distribution.



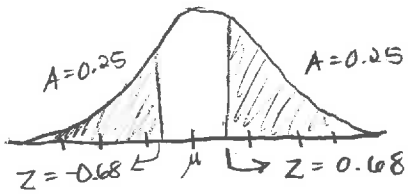
Normal prob. plots should be a straight line which shows a dist. is ≈ Normal.

_____ **6.** Which of the following statements are true?

I. The area under a Normal curve is always 1, regardless of the mean and standard deviation. ✓
II. The mean is always equal to the median for any Normal distribution. ✓
III. The interquartile range for any Normal curve extends from $\mu - \sigma$ to $\mu + \sigma$. x $\mu - 0.68\sigma$ to $\mu + 0.68\sigma$

(a) I and II
(b) I and III
(c) II and III
(d) I, II, and III
(e) None of the above gives the correct set of true statements.

Standard Normal Curve

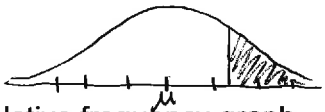


→ z-Scores

a **7.** The proportion of scores in a standard Normal distribution that are greater than 1.25 is closest to:

(a) 0.1056
(b) 0.1151
(c) 0.1600
(d) 0.8849
(e) 0.8944

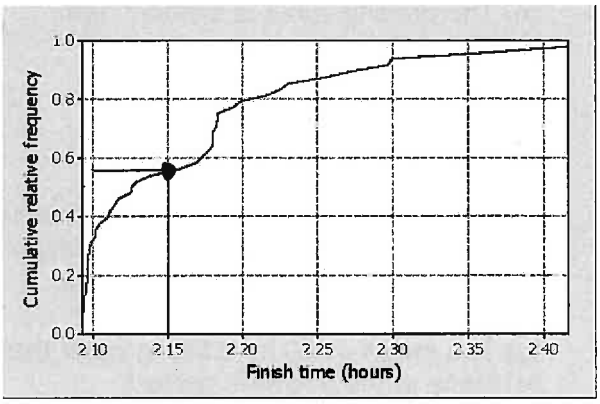
$P(z > 1.25)$
normalcdf (lower: 1.25, upper: 10^{10}, $\mu: 0$, $\sigma: 1$)



c **8.** At right is a cumulative relative frequency graph for the 48 racers who finished the grueling 50km cross-country ski race at the 2010 Vancouver Olympics. Approximately what proportion of the racers finished the race in *more* than 2.15 hours?

(a) 0.17
(b) 0.40
(c) 0.45
(d) 0.50
(e) 0.55

100% − 55%
≈45%



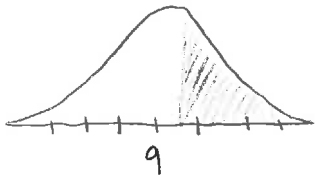
d **9.** In the previous question, the mean finish time is 2.164 hours and the standard deviation is 0.85 hours. The distribution is skewed right. What are the mean, standard deviation, and shape of the distribution of *z*-scores of the same data?

(a) Mean = 2.164, Standard deviation = 0.85, skewed right
(b) Mean = 2.164, Standard deviation = 0.85, skewed left
(c) Mean = 2.164, Standard deviation = 0.85, approximately normal
(d) Mean = 0, Standard deviation = 1, skewed right
(e) Mean = 0, Standard deviation = 1, approximately normal

b **10.** Kitchen appliances don't last forever. The lifespan of all microwave ovens sold in the United States is approximately Normally distributed with a mean of 9 years and a standard deviation of 2.5 years. What percentage of the ovens last more than 10 years?

(a) 11.5%
(b) 34.5%
(c) 65.5%
(d) 69%
(e) 84.5%

$N(9, 2.5)$
$P(x > 10)$
normalcdf



c **11.** In an experiment, an observed effect so large that it would rarely occur by chance is called

(a) an outlier
(b) influential
(c) statistically significant
(d) bias
(e) replication

b **12.** Jack and Jill are both enthusiastic players of a certain computer game. Over the past year, Jack's mean score when playing the game is 12,400 with a standard deviation of 1500. During the same period, Jill's mean score is 14,200, with a standard deviation of 2000. They devise a fair contest: each one will play the game once, and they will compare *z*-scores. Jack gets a score of 14,000, and Jill gets a score of 16,000. Who won the contest, and what were each of their *z*-scores?

(a) Jack's $z = 1.07$; Jill's $z = 1.11$; Jill wins the contest
(b) Jack's $z = 1.07$; Jill's $z = 0.90$; Jack wins the contest
(c) Jack's $z = 0.94$; Jill's $z = 1.11$; Jill wins the contest
(d) Jack's $z = 0.94$; Jill's $z = 0.90$; Jack wins the contest
(e) Jack's $z = 0.81$; Jill's $z = 0.99$; Jill wins the contest

Jack: $z = \frac{14000 - 12400}{1500} = 1.07$

Jill: $z = \frac{16000 - 14200}{2000} = 0.9$

b **13.** A company produces ceramic floor tiles that are supposed to have a surface area of 16.0 square inches. Due to variability in the manufacturing process, the actual surface area has a Normal distribution with a mean of 16.1 square inches and a standard deviation of 0.2 square inches. The proportion of tiles produced by the process with surface area less than 16.0 square inches is

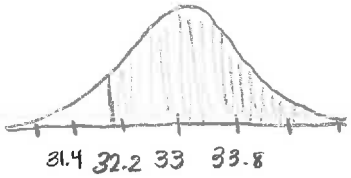
(a) 0.1915 (b) 0.3085 (c) 0.3173 (d) 0.4115 (e) 0.6915 $N(16.1, 0.2)$ $P(x < 16)$ normal cdf

a **14.** A company produces packets of soap powder labeled "Giant Size 32 Ounces." The actual weight of soap powder in a box has a Normal distribution with a mean of 33 oz. and a standard deviation of 0.8 oz. What proportion of packets are underweight (i.e., weigh less than 32 oz.)?

(a) 0.106 (b) 0.115 (c) 0.159 (d) 0.212 (e) 0.841

$N(33, 0.8)$

$P(x < 32)$ normal cdf



Part 2: Free Response

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

15. Lamar is shopping for a used car, and he's interested in determining the typical mileage on cars that are three or four years old. He looks at an online car-buying site and compares the number of miles on 30 cars that are three years old to 30 cars that are four years old. His results are summarized by Minitab below.
All values are in thousands of miles.

Descriptive Statistics: Mileage on Four year old cars and Three year old cars

Variable	N	Mean	Std dev	Min	Q1	M	Q3	Max
Four year old cars	30	56.68	17.82	23.60	47.80	54.70	64.50	100.30
Three year old cars	30	33.33	12.70	14.10	22.33	32.10	39.23	66.40

Both distributions are approximately Normally distributed.

a.) One car that Lamar is interested in is four years old and has been driven 60 thousand miles. Another one is three years old and has 40 thousand miles on it. How does the number of miles on these cars compare, relative to other cars of the same age? Provide appropriate statistical calculations to support your answer.

3-yr-old car: $z = \frac{40 - 33.33}{12.70} = 0.525$

4-yr car: $z = \frac{60 - 56.68}{17.82} = 0.1863$

Although the 4-yr-old car has more miles, when we compare z-scores, we can see the 3-yr-old car has more miles relative to its age. It is 0.52 st. deviations above the mean compared to the 4-yr-old car which is 0.19 st. dev. above the mean.

b.) Based on the information above, estimate the number of three year old cars Lamar looked at that had been driven more than 20 thousand miles.

$P(X > 20)$

$z = \frac{20 - 33.33}{12.70}$

$z = -1.05$

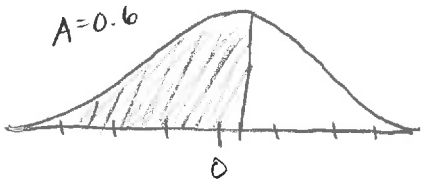
$P(z > -1.05)$

normalcdf(lower: −1.05, upper: 10^{10}, μ: 0, σ: 1) = 0.853

$n = 30$, then 85% of 30 = 25.5

Approximately 26 of the 30 3-yr-old cars Lamar looked at had been driven more than 20k miles.

c.) Estimate the 60th percentile for mileage on the cars Lamar found that were three years old.



invNorm(area: 0.6, μ: 0, σ: 1)

$z = 0.253$

$0.253 = \frac{x - 33.33}{12.7}$ $x \approx 36.5$

60% of 3-yr-old vehicles of this type have approximately 36.5 thousand miles or less.

16. A researcher wishes to calculate the average height of patients suffering from a particular disease. From patient records, the mean was computed to be 156 cm, with a standard deviation of 5 cm. Further investigation reveals that the scale was misaligned, and that all readings are 2 cm too large, for example, a patient whose height is really 180 cm was measured as 182 cm. Furthermore, the researcher would like to work with statistics based on meters (1 meter = 100 centimeters). What would be the revised values for the mean and standard deviation of the patients' heights?

$Y = \frac{X-2}{100}$ or $Y = \frac{1}{100}X - \frac{1}{50}$

$\mu_Y = \mu_{\frac{1}{100}X - \frac{1}{50}} = \frac{1}{100}\mu_X - \frac{1}{50} = \frac{1}{100}(156) - \frac{1}{50} = 1.54\text{m}$

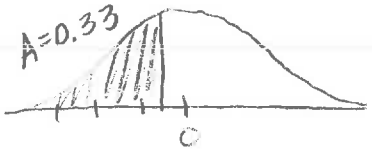
$\sigma_Y = \sigma_{\frac{1}{100}X - \frac{1}{50}} = \frac{1}{100}\sigma_X = \frac{1}{100}(5) = 0.05\text{m}$

The mean and st. dev would be 1.54 m and 0.05m, respectively.

17. During the 2009-2010 basketball season, the number of points scored in each game by the Boston Celtics was approximately Normally distributed with a mean of 99.2 points and a standard deviation of 10.5 points.

a.) What is the 33rd percentile of points scored by the Celtics?

$N(99.2, 10.5)$ invNorm(area:0.33, μ:0, σ:1) = -0.44



$-0.44 = \frac{X-99.2}{10.5}$

$X = 94.6$

Less than 94.6 points were scored is 33% of the Boston Celtics games for 2009-2010.

b.) The mean number of points scored by Los Angeles Lakers was 101.7. In what proportion of their games did the Celtics score more than the Lakers' mean score?

$P(X > 101.7)$

$Z = \frac{101.7 - 99.2}{10.5}$

$Z = 0.238$

$P(z > 0.238)$

normalcdf(lower: 0.238, upper: 10^{10}, μ: 0, σ:1)

$= 0.406$

The Celtics scored more than the Laker's mean score, 101.7 points, approximately 41% of the time during the 2009-2010 season.

CH.3—EXPLORING RELATIONSHIPS

Part 1: Multiple Choice. *Circle the letter corresponding to the best answer.*

Questions **1** and **2** refer to the following information:
For children between the ages of 18 months and 29 months, there is an approximately linear relationship between height and age. The relationship can be represented by $\hat{y} = 64.93 + 0.63x$, where *y* represents height (in centimeters) and *x* represents age (in months).

d **1.** Joseph is 22.5 months old. What is his predicted height?
(a) 50.80 (b) 64.96 (c) 65.96 (d) 79.11 (e) 87.40

$\hat{y} = 64.93 + 0.63(22.5)$
$x = 79.11$

b **2.** Loretta is 20 months old and is 80 centimeters tall. What is her residual? $= y - \hat{y}$
(a) - 2.47 (b) 2.47 (c) –12.60 (d) 12.60 (e) 77.53

$= 80 - (64.93 + 0.63(20))$

b **3.** You have data for many families on the parents' income and the years of education their eldest child completes. Your initial examination of the data indicates that children from wealthier families tend to go to school for longer. When you make a scatterplot,
(a) the explanatory variable is parents' income, and you expect to see a negative association.
(b) the explanatory variable is parents' income, and you expect to see a positive association.
(c) the explanatory variable is parents' income, and you expect to see very little association.
(d) the explanatory variable is years of education, and you expect to see a negative association.
(e) the explanatory variable is years of education, and you expect to see a positive association.

e **4.** A community college announces that the correlation between college entrance exam grades and scholastic achievement was found to be –1.08. On the basis of this you would tell the college that
(a) the entrance exam is a good predictor of success.
(b) the exam is a poor predictor of success.
(c) students who do best on this exam will be poor students.
(d) students at this school are underachieving.
(e) the college should hire a new statistician.

$-1 \le r \le 1$

r cannot be less than -1.

a **5.** An agricultural economist says that the correlation between corn prices and soybean prices is r = 0.7. This means that
(a) when corn prices are above average, soybean prices also tend to be above average.
(b) there is almost no relation between corn prices and soybean prices.
(c) when corn prices are above average, soybean prices tend to be below average.
(d) when soybean prices go up by 1 dollar, corn prices go up by 70 cents.
(e) the economist is confused, because correlation makes no sense in this situation.

$r = 0.7 \Rightarrow$ slope is positive

c **6.** Which of the following statements is/are true?
I. Correlation and regression require that there are clearly-identified explanatory and response variables.
II. Scatterplots require that both variables be quantitative.
III. Every least-squares regression line passes through $(\bar{x}, \bar{y})$.

(a) I and II only
(b) I and III only
(c) II and III only
(d) I, II, and III
(e) None of the above

An explanatory-response relationship is required for regression but not for correlation.

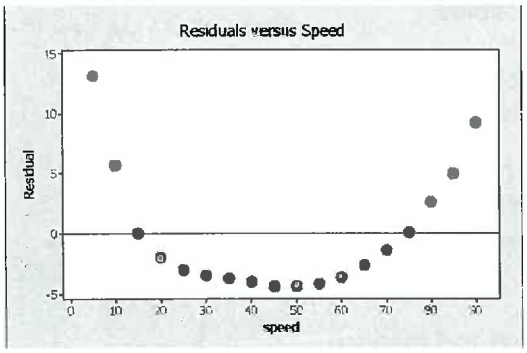
C **7.** There is an approximate linear relationship between the height of females and their age (from 5 to 18 years) described by predicted height = 50.3 + 6.01(age) where height is measured in centimeters and age in years. Which of the following is *not* correct?

(a) The estimated slope is 6.01, which implies that female children between the ages off 5 and 18 increase in height by about 6 cm for each year they grow older.
(b) The estimated height of a female child who is 10 years old is about 110 cm.
(c) The estimated intercept is 50.3 cm. We can conclude from this that the typical height of female children at birth is 50.3 cm. This is for girls 5-18yrs, so we cannot predict height at birth.
(d) The average height of female children when they are 5 years old is about 50% of the average height when they are 18 years old.
(e) My niece is about 8 years old and is about 115 cm tall. She is taller than average for girls her age.

e **8.** You are interested in predicting the cost of heating houses on the basis of how many rooms the house has. A scatterplot of 25 houses reveals a strong linear relationship between these variables, so you calculate a least-squares regression line. "Least-squares" refers to

(a) Minimizing the sum of the squares of the 25 houses' heating costs.
(b) Minimizing the sum of the squares of the number of rooms in each of the 25 houses.
(c) Minimizing the sum of the products of each house's actual heating costs and the predicted heating cost based on the regression equation.
(d) Minimizing the sum of the squares of the difference between each house's heating costs and number of rooms.
(e) Minimizing the sum of the squares of the residuals. $y-\hat{y}$ = residual (the distance from each point to the LSRL). We want to minimize each residual.

d **9.** A study of the fuel economy for various automobiles plotted the fuel consumption (in liters of gasoline used per 100 kilometers traveled) vs. speed (in kilometers per hour). A least-squares line was fit to the data. Here is the residual plot from this least-squares fit.



What does the residual plot tell you about the linear model?

(a) The residual plot confirms the linearity of the fuel economy data.
(b) The residual plot does not confirm *nor* rule out the linearity of the data.
(c) The residual plot suggests that the model may be linear, but more data points are needed to confirm this.
(d) The residual plot clearly indicates that the data isn't linear.
(e) A residual plot is not an appropriate means for evaluating a linear model.

residual plots should show no clear pattern, which would suggest a linear model is appropriate for prediction.

Leonardo da Vinci, the renowned painter, speculated that an ideal human would have an armspan (distance from the outstretched fingertip of the left hand to the outstretched fingertip of the right hand) that was equal to his height. The following computer regression printout shows the results of a least-squares regression of armspan on height, both in inches, for a sample of 18 high school students.

da Vince:
armspan = height

descriptive statistics:

Predictor	Coef.	SE Coef	T	P
Constant	11.5474	5.6	2.06	0.0558
Height	0.84024	0.08091	10.4	0.000
R-sq = 87.1%		R-sq(adj.) = 86.3%		

armspan = 11.5474 + 0.84024 (height)

C **10.** Which of the following statements is *false*?
(a) This least-squares regression model would make a prediction that is 1.64 inches higher than da Vinci projected for a 62-inch tall student.
(b) If one of the students in the sample had a height of 70.5 inches and an armspan of 68 inches, then the residual for this student would be –2.78 inches.
(c) The least-squares regression line has a steeper slope than the equation for da Vinci's relationship between armspan and height. 0.84024 < 1
(d) For every one-inch increase in height, the regression model predicts about a 0.84-inch increase in armspan.
(e) For a student 66 inches tall, our model would predict an armspan of about 67 inches.

e **11.** The correlation coefficient measures
(a) whether there is a relationship between two variables.
(b) the strength of the relationship between two quantitative variables.
(c) whether or not a scatterplot shows an interesting pattern.
(d) whether a cause and effect relation exists between two variables.
(e) the strength of the linear relationship between two quantitative variables.

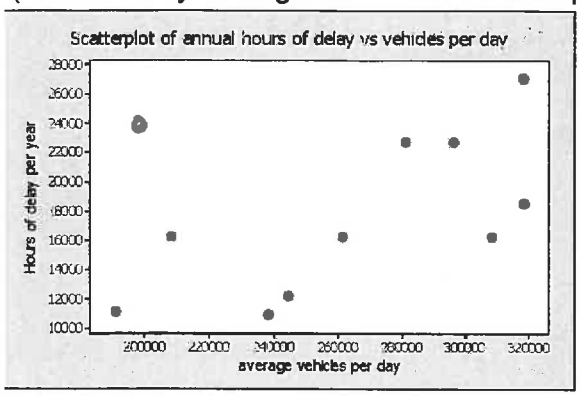
e **12.** Which of the following are most likely to be negatively correlated?
(a) The total floor space and the price of an apartment in New York.
(b) The percentage of body fat and the time it takes to run a mile for male college students.
(c) The heights and yearly earnings of 35-year-old U.S. adults.
(d) Gender and yearly earnings among 35-year-old U.S. adults.
(e) The prices and the weights of all racing bicycles sold last year in Chicago.

lighter bikes cost more, heavier bikes cost less.
as $x \to \infty$, $f(x) \to -\infty$

Part 2: Free Response

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

11. How are traffic delays related to the number of cars on the road? Below is data on the total number of hours of delay per year at 10 major highway intersections in the western United States *versus* traffic volume (measured by average number of vehicles per day that pass through the intersection).



a.) Describe what the scatterplot reveals about the relationship between traffic delays and number of cars on the road.

The scatterplot shows a moderately positive, linear relationship between average number of vehicles and yearly delays.

b.) Suppose another data point at (200000, 24000), that is 200,000 vehicles per day and 24,000 hours of delay per year, were added to the plot. What effect, if any, will this new point have on the correlation between these two variables? Explain.

This point would decrease correlation r, because it does not follow the general pattern. It would also decrease the y-int and increase slope.

Below is the computer output for the regression of hours of delay versus number of vehicles per day.

Predictor	Coef	SE Coef	T	P
Constant	-3629	7367	-0.49	0.634
vehicles per day	0.07822	0.02684	2.91	0.017

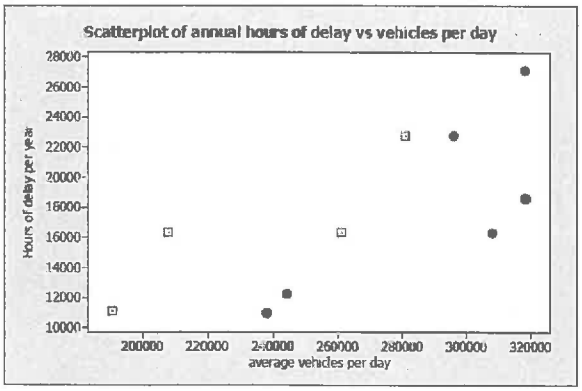
S = 3899.57 R-Sq = 48.6% R-Sq(adj) = 42.8%

c.) What is the slope of the regression line? Interpret the slope in the context of this problem.

The slope of the regression line is 0.07822. This means, as the average number of vehicles traveling through said intersections increases by 1, the amount of yearly delays increases by 0.078 hrs.

d.) Explain what the quantity S = 3899.57 measures in the context of this problem.

S is the standard deviation of the residuals. That means, the average prediction error will be 3,899.57 hours of traffic delays per year.



12. Below is the same scatterplot, but with the six intersections in California plotted as circles and the four in other western states plotted as squares.

Comment on how the relationship between average number of vehicles per day and hours of delay per year differs between the California intersections and the intersections in other western states.

When the intersections are split up, we can see the relationship between average numbr of vehicles per year and yearly delays grows much stronger. It's apparent there is less traffic volume in other western states which yields less delay, and more traffic volume for California with more delays. The slope of the LSRL for California would be larger compared to other western States, and the y-int. would smaller. Both would have similar

13.) An ecologist studying breeding habits of the common crossbill in different years finds that there is a linear relationship between the number of breeding pairs of crossbills and the abundance of the spruce cones. Below are statistics on eight years of measurements, where x = average number of cones per tree and y = number of breeding pairs of crossbills in a certain forest.

	Mean	Standard deviation
x = mean number of cones/tree	23.0 $\bar{x}$	16.2 s_x
y = number of crossbill pairs	18.0 $\bar{y}$	15.1 s_y

The correlation between x and y is $r = 0.968$

a.) Find the equation of the least-squares regression line (with y as the response variable).

$\hat{y} = a + bx$

$b = r\frac{s_y}{s_x} = (0.968)\left(\frac{15.1}{16.2}\right) = 0.9023$

$\bar{y} = a + b(\bar{x})$

$18.0 = a + 0.9023(23)$

$a = -2.7522$

(# crossbill pairs) $= -2.75 + 0.902$ (mean # of cones/tree)

b.) What percentage of the variation in numbers of breeding pairs of crossbills can be accounted for by this regression?

$r = 0.968 \Rightarrow r^2 = 0.937$ - coef. of determination

93.7% of the variation in the # of crossbill pairs can be accounted for by the linear model relating # of crossbill pairs to the mean number of cones per tree.

c.) Based on these data, can we conclude that the abundance of spruce cones is responsible for the number of breeding pairs of crossbills? Explain.

Correlation does not imply causation, but we can conclude there is a strong relationship between # of crossbill pairs observed and the mean number of cones per tree. This is evident because the coefficient of determination is 0.937. That means, 93.7% of the variation in the number of crossbill pairs can be accounted for by the LSRL relating Crossbill pairs to the mean # of cones observed. There is very little variation, 6.3%, that is due to other, unaccounted factors, e.g., temperature, predator pop., etc. Since $r = 0.968$, we know the slope is positive — as the mean # of cones/tree increases, so does the # of crossbill pairs. There is certainly a positive relationship between cones and pairs, but we cannot say "cones causes more crossbill pairs."

CH.4 & 5—Two-Way Tables & Producing Data

Part 1: Multiple Choice. *Circle the letter corresponding to the best answer.*

b **1.** What electrical changes occur in muscles as they get tired? Student subjects are instructed to hold their arms above their shoulders as long as they can. Meanwhile, the electrical activity in their arm muscles is measured. This is

(a) an observational study.
(b) an uncontrolled experiment.
(c) a randomized comparative experiment.
(d) a matched pairs design.
(e) impossible to describe unless more details of the study are provided.

Treatment: holding arms up
response: electrical activity
No control group or randomizing.

e **2.** Which of the following statements is false?

(a) Nonresponse can cause bias in surveys because nonrespondents often tend to behave differently than people who respond.
(b) Non-sampling errors can distort the results of a census.
(c) Slight changes in the wording of questions can make a measurable difference in survey results.
(d) People will sometimes answer a question differently for different interviewers.
(e) Sophisticated statistical methods can always correct the results if the population you are sampling from is different from the population of interest, for example, due to undercoverage.

c **3.** An experiment to measure the effect of giving growth hormones to girls affected by Turner's Syndrome was carried out recently in Vancouver. All 34 girls in the study were given the growth hormone and their heights were measured at the time the hormone was given and again one year later. No measurements were made on their final adult heights. Which of the following is not a problem with this experiment:

(a) There was no blinding.
(b) There was no control group.
(c) Nonresponse bias – pertaining to surveys; not applicable to experiments
(d) There was insufficient attention to the placebo effect.
(e) Because final heights were not measured, it is impossible to tell if the hormone affected final height or only accelerated growth and made no difference to final height.

e **4.** The following numbers appear in a table of random digits:

38683 50279 36224 09844 13578 28251 12708 24684 38, 35, 02, 22, 40

A scientist will be measuring the total amount of leaf litter in a random sample ($n = 5$) of forest sites selected without replacement from a population of 45 sites. The sites are labeled 01, 02, . . . , 45 and she starts at the beginning of the line of random digits and takes consecutive pairs of digits. Which of the following is correct?

(a) Her sample is 38, 25, 02, 38, 22
(b) Her sample is 38, 68, 35, 02, 22
(c) Her sample is 38, 35, 27, 28, 08
(d) Her sample is 38, 65, 35, 02, 79
(e) Her sample is 38, 35, 02, 22, 40

Choose 2-digit #s, throw out repeats, 00, and 46-99. Stop after you have 5 distinct 2-digit #s.

b **5.** To test the effects of a new fertilizer, 100 plots were divided in half. Fertilizer A is randomly applied to one half, and B to the other. This is

(a) an observational study.
(b) a matched pairs experiment.
(c) a completely randomized experiment.
(d) a block design, but not a matched pairs experiment.
(e) impossible to classify unless more details of the study are provided.

All 100 plots are assumed to be similar. The two halves act as pairs. Compare A with B after new fertilizer is applied.

a **6.** A civil engineer is testing the reliability of traffic signal controllers produced by two different companies. He has 20 sets of controllers from each company, and he has been given clearance to install them at 40 different intersections in the city. He randomly assigns the controllers from company A to 20 intersections and the controllers from company B to the other intersections. The most important reason for this random assignment is that → reduces

(a) randomization is a good way to create two groups of 20 intersections that are as similar as possible, so that comparisons can be made between the two groups.
(b) randomization eliminates the impact of any confounding variables.
(c) randomization makes the analysis easier since the data can be collected and entered into the computer in any order.
(d) randomization ensures that the study is double-blind.
(e) randomization reduces the impact of outliers.

c **7.** An airline that wants to assess customer satisfaction chooses a random sample of 10 of its flights during a single month and asks all of the passengers on those flights to fill out a survey. This is an example of a

(a) multistage sample.
(b) stratified sample.
(c) cluster sample. → surveyed everyone in the selected groups
(d) simple random sample.
(e) convenience sample.

c **8.** You work for an advertising agency that is preparing a new television commercial to appeal to women. You have been asked to design an experiment to compare the effectiveness of three versions of the commercial. Each subject will be shown one of the three versions and then asked her attitude toward the product. You think there may be large differences between women who are employed outside the home and those who are not. Because of these differences, you should use

(a) a completely randomized design.
(b) a categorical variable.
(c) a block design. split subjects into homogeneous groups: employed & unemployed
(d) a stratified design.
(e) a multistage sample. → not an experimental design method but a sampling method.

c **9.** According to the 1990 census, those states with an above-average number of people *X*, who fail to complete high school tend to have an above average number of infant deaths, *Y*. In other words, there is a positive association between *X* and *Y*. The most plausible explanation for this is

(a) *X* causes *Y*. Programs to keep teens in school will help reduce the number of infant deaths.
(b) *Y* causes *X*. Programs that reduce infant deaths will ultimately reduce high school dropouts.
(c) Lurking variables are probably present. For example, states with large populations may have both larger numbers of people who don't complete high school and more infant deaths.
(d) Both of these variables are directly affected by the higher incidence of cancer in certain states.
(e) The association between *X* and *Y* is purely coincidental.

correlation does not imply causation

b **10.** Eighty volunteers who currently use a certain brand of over-the-counter allergy medication have been recruited to participate in a trial of a new allergy medication. The volunteers are randomly assigned to one of two groups. One group continues to take their current medication, the other group switches to the new experimental medication. Each is asked after two weeks if their allergy symptoms are worse, better, or about the same as they were at the start of the study. Which of the follow best describes a conclusion that can be drawn from this study?

(a) We can determine whether the new drug reduces symptoms more than the old drug for anyone who suffers from allergies.
(b) We can determine whether the new drug reduces symptoms more than the old drug for the subjects in the study. need more control to generalize to entire population
(c) We can determine whether the allergies sufferers' symptoms improved more with the new drug than with the old drug, but we can't establish cause and effect.
(d) We cannot draw any conclusions, since the all the volunteers were already taking the old drug when the experiment started.
(e) We cannot draw any conclusions, because there was no control group.

Use the following information to answer Questions 11-13.

A review of voter registration records in a small town yielded the following table of the number of males and females registered as Democrat, Republican, or some other affiliation.

	Male	Female	
Democrat	300	600	900
Republican	500	300	800
Other	200	100	300
	1000	1000	2000

d **11.** The proportion of males that are registered as Democrat is

(a) 300
(b) 30
(c) 0.33
(d) 0.30
(e) 0.15

$\frac{300}{1000}$ Democrats among males

d **12.** Your percentage from Question 11 is part of

(a) The marginal distribution of political party registration.
(b) The marginal distribution of gender.
(c) The conditional distribution of gender among Democrats.
(d) The conditional distribution of political party registration among males.
(e) The conditional distribution of males within gender.

c **13.** The proportion of registered Democrats that are male is

(a) 300
(b) 33
(c) 0.33
(d) 0.30
(e) 0.15

$\frac{300}{900}$ males among Democrats

a **14.** The principle reason for the use of *controls* in designing experiments is that it

(a) distinguishes a treatment effect from the effects of confounding variables.
(b) allows double-blinding
(c) reduces sampling variability
(d) creates approximately equal groups for comparison
(e) eliminates the placebo effect

Part 2: Free Response

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

15. Read the following article about the connection between vitamin E and heart bypass surgery.

Vitamin E may have special health benefits

Large doses of vitamin E apparently can reduce harmful side effects of bypass surgery in heart patients. A study involving 28 bypass patients found that the 14 randomly-assigned patients who took vitamin E for two weeks before their operations had significantly better heart function after the procedure than the 14 randomly-assigned patients who took placebos.

The vitamins apparently prevent damage to the heart muscle by destroying the toxic chemicals, called free radicals, that form when blood is cut off during the surgery, said Dr. Terrance Yau of the University of Toronto.

(a) Explain why this is an experiment and not an observational study.

A clear treatment was imposed on all patients: vitamin E or a placebo.

(b) Identify the explanatory and response variables.

The explanatory variable is the treatment of vitamin E, and the response is the patient's heart function post-surgery, after the treatment.

(c) Identify the type of experimental design used in this study. Justify your answer.

The experimental design was completely randomized. 28 people were split into one of two groups, vitamin E or placebo, randomly. Then heart function was compared after surgery.

(d) In the second sentence above is the phrase, "...the 14 patients who took vitamin E for two weeks before their operations had significantly better heart function after the procedure ..." What is the statistical meaning of the word "significantly" in the context of this study?

Statistically significant means the difference in the heart function of the vitamin E group versus the placebo group was so large that it would rarely happen by chance.

(e) This was a controlled experiment. Describe how it was controlled and explain the purpose of doing so.

The experiment had a control group, the placebo group, which helps experimenters compare the treatment, not other lurking variables. Randomizing helps eliminate potential confounding variables. Both techniques help experimenters determine a more clear cause-and-effect relationship between vitamin E treatment and heart function after bypass surgery.

16. As a researcher for a pharmaceutical company, you are designing a study to test the effectiveness of a new treatment for migraine headaches. You have been given a list of 126 people willing to participate in the trial. The first 70 people are female; the remaining 56 are male.

(a) Preliminary research suggests that men and women respond differently to this new treatment. What sort of experimental design would you choose for this study, and why?

Due to gender differences, a blocking design should be used where females are one block and males are another block. The treatment effects will be compared among gender groups.

(b) Explain why an experiment involving 70 women and 56 men is preferable to one involving 10 women and 8 men.

Larger samples reduce the random variation that can occur in small samples, which will ultimately increase the ability to determine a cause-and-effect relationship of the treatment.

(c) Describe a design for this experiment. Be sure to include a description of how you assign individuals to the treatment groups.

Put the 70 females in alphabetical order and number them 1-70. Assign 1-35 to the "new" treatment and 36-70 to the "existing" treatment. For males, put the 56 males in alphabetical order and number 1-56. 1-28 will be assigned to the "new" treatment and 29-56 to the "existing" treatment. Compare new with existing among gender.

17. Bias is present in each of the following sample designs. In each case, identify the type of bias involved and state whether you think the sample proportion obtained is higher or lower than the true population proportion.

(a) A political pollster is seeking information on public attitudes toward funding of pornographic art by the National Endowment for the Arts (NEA). He asks an SRS of 2000 U.S. adults, "Rather than support government censorship of artistic expression, are you in favor of continuing federal funding of artists whose work may be controversial?" 85% of those surveyed answer "yes." Biased is in the wording of the question, specifically the first portion, "Rather than support... artistic expression." The question is causing respondents to question their beliefs and what they might represent rather that answering the question truthfully. The question is most likely causing an overestimate and should simply be phrased "Are you in favor..." Respondents will be less focused on their response and the possible external meaning.

(b) In 2003, the AARP conducted a survey of their members (people over age 50 who pay membership dues) on proposed Medicare legislation. One of the questions was: "Even if this plan won't affect you personally either way, do you think it should be passed so that people with low incomes or people with high drug costs can be helped?" 75% of respondents answered yes.

Biased existed in the sampling method. Specifically, the survey appears to be targeting the wrong population, those who are not affected personally. This is called undercoverage. Those that need to respond are not given an equal chance. This will most likely result in an underestimate of the true populations response, because the audience responding has no vested interest.

CH.6—Probability and Simulation

Part 1: Multiple Choice. *Circle the letter corresponding to the best answer.*

a **1.** When rolling a fair die, you roll a 6 four times in a row. Given that each roll is independent, what is the probability that the next roll yields a six also?

(a) $\frac{1}{6}$ (b) $\left(\frac{1}{6}\right)^5$ (c) $1-\left(\frac{1}{6}\right)^5$ (d) $\frac{5}{6}$ (e) $\left(\frac{1}{6}\right)^6$

b **2.** Suppose there are three cards in a deck, one marked with a 1, one marked with a 2, and one marked with a 5. You draw *two* cards at random and without replacement from the deck of three cards. The sample space $S = \{(1, 2), (1, 5), (2, 5)\}$ consists of these three equally likely outcomes. Let X be the **sum** of the numbers on the two cards drawn. Which of the following is the correct set of probabilities for X?

(a)

X	P(X)
1	1/3
2	1/3
5	1/3

(b)

X	P(X)
3	1/3
6	1/3
7	1/3

(c)

X	P(X)
3	3/16
6	6/16
7	7/16

(d)

X	P(X)
3	1/4
6	1/2
7	1/2

(e)

X	P(X)
1	1/4
2	1/2
5	1/2

Use the following table for questions 3 – 5.
The following table compares the hand dominance of 200 Canadian high-school students and what methods they prefer using to communicate with their friends. Suppose one student is chosen randomly from this group of 200.

	Cell phone/Text	In person	Online	**Total**
Left-handed	12	13	9	**34**
Right-handed	43	72	51	**166**
Total	**55**	**85**	**60**	**200**

c **3.** What is the probability that the student chosen prefers to communicate with friends in person?
(a) 0.065 (b) 0.153 (c) 0.425 (d) 0.382 (e) 0.595 $\frac{85}{200}$

d **4.** If you know the person that has been randomly selected is left-handed, what is the probability that they prefer to communicate with friends in person?
(a) 0.065 (b) 0.153 (c) 0.425 (d) 0.382 (e) 0.595 $\frac{13}{34}$

d **5.** You select a student from the group at random. Which of the following statements is true about the events "Left-Handed" and "Prefers to communicate with friends in person"?
(a) The events are mutually exclusive and independent.
(b) The events are not mutually exclusive but they are independent.
(c) The events are mutually exclusive, but they are not independent.
(d) The events are not mutually exclusive, nor are they independent.
(e) The events are independent, but we do not have enough information to determine if they are mutually exclusive.

A=left-handed
B=in person
If independent, then $P(A|B) = P(A)$
$P(A|B) = \frac{13}{85}$ $P(A) = \frac{34}{200} = \frac{17}{100}$
$P(A|B) \neq P(A)$
$\frac{13}{85} \neq \frac{17}{100}$

a **6.** Event *A* has probability 0.4. Event *B* has probability 0.5. If *A* and *B* are disjoint (mutually exclusive), then the probability that **both** events occur is
(a) 0 (b) 0.1 (c) 0.2 (d) 0.7 (e) 0.9

If mutually exclusive, $P(A \cap B) = 0$

e **7.** Event *A* has probability 0.4. Event *B* has probability 0.5. If *A* and *B* are disjoint (mutually exclusive), then the probability that **either** events occur is
(a) 0 (b) 0.1 (c) 0.2 (d) 0.7 (e) 0.9

$P(A \text{ or } B) = P(A) + P(B)$
$= 0.4 + 0.5$

c **8.** Event *A* has probability 0.4. Event *B* has probability 0.5. If *A* and *B* are independent, then the probability that **both** events occur is
(a) 0 (b) 0.1 (c) 0.2 (d) 0.7 (e) 0.9

$P(A \cap B) = P(A)P(B)$
$= 0.4(0.5)$

d **9.** Event *A* has probability 0.4. Event *B* has probability 0.5. If *A* and *B* are independent, then the probability that **either** events occur is
(a) 0 (b) 0.1 (c) 0.2 (d) 0.7 (e) 0.9

Mutually inclusive
$P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$
$= 0.4 + 0.5 - 0.2$

Use the following situation for questions 10 and 11.
Ignoring twins and other multiple births, assume that babies born at a hospital are independent random events with the probability that a baby is a boy and the probability that a baby is a girl both equal to 0.5.

e **10.** The probability that the next five babies are girls is
(a) 1 (b) 0.5 (c) 0.1 (d) 0.0625 (e) 0.03125

$(0.5)^5$

e **11.** The probability that at **least one** of the next three babies is a boy is
(a) 0.125 (b) 0.333 (c) 0.667 (d) 0.750 (e) 0.875

$P(x \geq 1) = 1 - P(x < 1) = 1 - P(x = 0) = 1 - (0.5)^3$

d **12.** In a cookie jar, there are 12 chocolate chip cookies, 5 oatmeal raison cookies, and 7 macadamia nut cookies. What is the probability that if two cookies were chosen (without replacement), that both cookies were the same type of cookie?
(a) 0.239 (b) 0.036 (c) 0.076 (d) 0.351 (e) 0.378

exclusive

$P(\text{2 chocolate or 2 oatmeal or 2 mac nut}) = \left(\frac{12}{24}\right)\left(\frac{11}{23}\right) + \left(\frac{5}{24}\right)\left(\frac{4}{23}\right) + \left(\frac{7}{24}\right)\left(\frac{6}{23}\right)$

c **13.** A die is loaded so that the number 3 comes up four times as often as any other number. What is the probability of rolling a 1, 3, or 5?
(a) $\frac{1}{2}$ (b) $\frac{1}{3}$ (c) $\frac{2}{3}$ (d) $\frac{4}{9}$ (e) $\frac{5}{9}$

exclusive

1 2 3 3 3 3 4 5 6

X	1	2	3	4	5	6
P(x)	1/9	1/9	4/9	1/9	1/9	1/9

$P(1 \text{ or } 3 \text{ or } 5) = P(1) + P(3) + P(5)$
$= 1/9 + 4/9 + 1/9$

Use the following for questions 14 and 15:
An event A will occur with probability 0.5. An event B will occur with probability 0.4. The probability that both A and B will occur is 0.2.

a **14.** The conditional probability of A, given B
(a) is 0.5
(b) is 0.4
(c) is 0.7
(d) is 0.2
(e) cannot be determined from the information given.

$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.2}{0.4} = \frac{1}{2}$

If independent $P(A|B) = P(A)$
$\frac{1}{2} = 0.5$ ✓

a **15.** We may conclude that
(a) events A and B are independent.
(b) events A and B are mutually exclusive.
(c) either A or B always occurs.
(d) events A and B are complementary.
(e) none of the above is correct.

All parts → A (60%) → defective (10%) = 0.06
A (60%) → ¬defective (90%) = 0.54
All parts → B (30%) → defective (30%) = 0.09
B (30%) → ¬defective (70%) = 0.21
All parts → C (10%) → defective (40%) = 0.04
C (10%) → ¬defective (60%) = 0.06

b **16.** Three machines – A, B, and C – are used to produce a large quantity of identical parts at a factory. Machine A produces 60% of the parts, while Machines B and C produce 30% and 10% of the parts respectively. Historical records indicate that 10% of the parts produced by Machine A are defective, compared with 30% for Machine B and 40% for Machine C. What is the probability that a randomly chosen part is defective?
(a) 0.8 (b) 0.19 (c) 0.06 (d) 0.04 (e) 0.09

exclusive $P(A\text{def or } B\text{def or } C\text{def}) = 0.06 + 0.09 + 0.04$

b **17.** You are told that your score on an exam is at the 85 percentile of the distribution of scores. This means that
(a) Your score was lower than approximately 85% of the people who took this exam.
(b) Your score was higher than approximately 85% of the people who took this exam.
(c) You answered 85% of the questions correctly.
(d) If you took this test (or one like it) again, you would score as well as you did this time 85% of the time.
(e) 85% of the people who took this test earned the same score you did.

Part 2: Free Response

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

24. Based on previous records, 17% of the vehicles passing through a tollbooth have out-of-state plates. This can be expressed as P(out-of-state plates) = 0.17. Assume each vehicle plate is independent of other vehicle plates.

(a) Describe what the Law of Large Numbers says in the context of this probability.

As the number of cars that pass through this particular tollbooth gets very large (approaches ∞), the proportion of vehicles with out-of-state plates approaches 0.17.

(b) What is the probability that none of the next four vehicles have out-of-state plates?

P(not out-of-state plate) = 1 − 0.17 = 0.83

Independent

$(0.83)^4 = 0.4746$

The probability none of the next 4 vehicles will have out-of-state plates is approximately 47%.

(c) You want to estimate the probability that exactly one of the next four vehicles have out of state plates. Describe the design of a simulation to estimate this probability. Explain clearly how you will use the partial table of random digits below to carry out your simulation.

What is the probability exactly one of the next four vehicles has out-of-state plates. To find out, Let the random numbers 00-16 = out-of-state plates. Let 17-99 be in-state plates. Choose 4 two-digit numbers, throwing out repeats. Determine the proportion of out-of-state plates for each trial. Repeat this process for several trials. Average each proportion to answer the question.

(d) Carry out 5 trials of your simulation. Mark on or above each line of the table so that someone can clearly follow your method.

	1	2	3	4	5
188	87370	88099	89695	87633	76987	88503	26257	51736
189	88296	95670	74932	65317	93848	43988	47597	83044
190	79485	92200	99401	54473	34336	82786	05457	60343
191	40830	24979	23333	37619	56227	95941	59494	86539
192	32006	76302	81221	00693	95197	75044	46596	11628

1: 87, 37, 08, 80 — 1/4

2: 99, 89, 69, 58 — 0/4

3: 76, 33, 76, 98 — 0/4

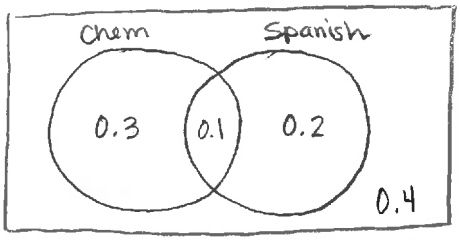
4: 78, 55, 03, 26 — 1/4

5: 25, 75, 17, 36 — 0/4

After 5 trials, exactly one out of 4 vehicles had out-of-state plates 2 times, so $\frac{2}{5}$ or 40%.

25 . A counselor analyzes student's course selection and calculates the following: The probability that a randomly-chosen student is taking Spanish is 0.4, that the student is taking Chemistry is 0.3, and that the student is taking BOTH Chemistry and Spanish is 0.1.

(a) Let S = Randomly-chosen student is taking Spanish, and C = Randomly-chosen student is taking Chemistry. Sketch a Venn diagram or two-way table that summarizes the probabilities above.



	Chem	¬Chem	
Spanish	0.1	0.2	0.3
¬Spanish	0.3	0.4	0.7
	0.4	0.6	1

(b) Find each of the following:

i. The probability that a randomly-selected student is taking Spanish OR Chemistry.

exclusive $P(S \text{ or } C) = P(S) + P(C) - P(S \cap C)$
$= 0.3 + 0.4 - 0.1 = 0.6$ 60%

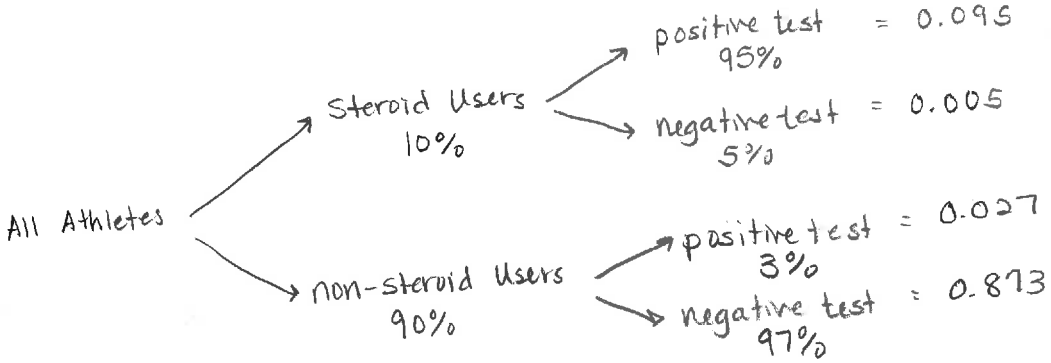
ii. The probability that a randomly-selected student is taking Spanish or <u>isn't</u> taking Chemistry.

exclusive $P(S \text{ or } \neg C) = P(S) + P(\neg C) - P(S \cap \neg C)$
$= 0.3 + 0.6 - 0.2 = 0.7$ 70%

iii. The probability that a randomly-selected student <u>doesn't</u> take Spanish and <u>doesn't</u> take Chemistry.

$P(\neg S \text{ and } \neg C) = 1 - P(S \text{ or } C)$
$= 1 - 0.6 = 0.4$ 40%

26. A company has developed a drug test to detect steroid use by athletes. The test is accurate 95% of the time when an athlete has taken steroids. It is 97% accurate when an athlete hasn't taken steroids. Suppose the drug test will be used in a population of athletes in which 10% have actually taken steroids. Suppose we know that the test was accurate, what is the probability that they didn't take steriods?



$P(\text{did not take steroids} \mid \text{accurate test}) = \dfrac{P(\text{non-steroid user} \cap \text{accurate test})}{P(\text{accurate test})}$

$= \dfrac{0.873}{(0.095 + 0.873)}$

$= 0.90 \approx 90\%$

CH.7 & 8—Random Variables, Binomial and Geometric Distributions

Part 1: Multiple Choice. *Circle the letter corresponding to the best answer.*

e **1.** In the town of Tower Hill, the number of cell phones in a household is a random variable *W* with the following distribution:

W	0	1	2	3	4	5
P(W)	0.1	0.1	0.25	0.3	0.2	0.05

The probability that a randomly-selected household has at least two cell phones is $P(X \geq 2)$

(a) 0.20. (b) 0.25. (c) 0.55. (d) 0.70. (e) 0.80.

$= P(x=2) + P(x=3) + P(x=4) + P(x=5)$

$= 0.25 + 0.3 + 0.2 + 0.05$

a **2.** A random variable *Y* has the following distribution:

Y	-1	0	1	2
P(Y)	3C	2C	0.4	0.1

$= 1$

$3C + 2C + 0.4 + 0.1 = 1$

$5C + 0.5 = 1$

$5C = 0.5$

$C = 0.1$

The value of the constant *C* is:

(a) 0.10. (b) 0.15. (c) 0.20. (d) 0.25. (e) 0.75.

b **3.** A rock concert producer has scheduled an outdoor concert. If it is warm that day, she expects to make a $20,000 profit. If it is cool that day, she expects to make a $5000 profit. If it is very cold that day, she expects to suffer a $12,000 loss. Based upon historical records, the weather office has estimated the chances of a warm day to be 0.60; the chances of a cool day to be 0.25. What is the producer's expected profit?

(a) $5,000 (b) $11,450 (c) $13,000 (d) $13,250 (e) $15,050

Let x = Profit

x	$20000	$5000	-$12000
P(x)	0.6	0.25	0.15

$E(x) = \sum x_i p_i$

$= 20000(0.6) + 5000(0.25) + (-12000)(0.15)$

d **4.** Roll one 10-sided die 12 times. The probability of getting exactly 4 eights in those 12 rolls is given by

(a) $\binom{10}{4} \cdot \left(\frac{1}{10}\right)^4 \cdot \left(\frac{9}{10}\right)^8$

(b) $\binom{10}{4} \cdot \left(\frac{1}{10}\right)^4 \cdot \left(\frac{9}{10}\right)^6$

(c) $\binom{12}{4} \cdot \left(\frac{1}{10}\right)^4 \cdot \left(\frac{9}{10}\right)^6$

(d) $\binom{12}{4} \cdot \left(\frac{1}{10}\right)^4 \cdot \left(\frac{9}{10}\right)^8$

(e) $\binom{12}{4} \cdot \left(\frac{1}{10}\right)^8 \cdot \left(\frac{9}{10}\right)^4$

bin $(n=12, p=1/10)$

$P(X=4) = \binom{12}{4}\left(\frac{1}{10}\right)^4\left(\frac{9}{10}\right)^8$

d **5.** The variance of the sum of two random variables *X* and *Y* is

(a) $\sigma_X + \sigma_Y$

(b) $(\sigma_X)^2 + (\sigma_Y)^2$.

(c) $\sigma_X + \sigma_Y$, but only if *X* and *Y* are independent.

(d) $(\sigma_X)^2 + (\sigma_Y)^2$, but only if *X* and *Y* are independent.

(e) None of these.

a **6.** Let the random variable X represent the weight of male black bears before they begin hibernation. Research has shown that X is approximately Normally distributed with a mean of 250 pounds and a standard deviation of 50 pounds. What is $P(X > 325 \text{ pounds})$?

(a) 0.0668 (b) 0.2514 (c) 0.7486 (d) 0.8531 (e) 0.9332

N(250, 50)
normalcdf

e **7.** A set of 10 cards consists of 5 red cards and 5 black cards. The cards are shuffled thoroughly and you turn cards over, one at a time, beginning with the top card. Let Y be the number of cards you turn over until you observe the first red card. The random variable Y has which of the following probability distributions?

Not independent because it's sampling without replacement. It would be geometric if you were replacing the cards

(a) the Normal distribution with mean 5
(b) the binomial distribution with $p = 0.5$
(c) the geometric distribution with probability of success 0.5
(d) the uniform distribution that takes value 1 on the interval from 0 to 1
(e) none of the above

d **8.** A factory makes silicon chips for use in computers. It is known that about 90% of the chips meet specifications. Every hour a sample of 18 chips is selected at random for testing and the number of chips that meet specifications is recorded. What is the approximate mean and standard deviation of the number of chips meeting specifications?

$\mu_x = np = 18(0.9) = 16.2$

$\sigma_x = \sqrt{np(1-p)} = \sqrt{18(0.9)(0.1)} = 1.27$

(a) $\mu = 1.62$; $\sigma = 1.414$
(b) $\mu = 1.62$; $\sigma = 1.265$
(c) $\mu = 16.2$; $\sigma = 1.62$
(d) $\mu = 16.2$; $\sigma = 1.273$
(e) $\mu = 16.2$; $\sigma = 4.025$

a **9.** In order for the random variable X to have a geometric distribution, which of the following conditions must X satisfy?

I $p < 0.5$ ✗
II The number of trials is fixed. ✗
III Trials are independent. ✓
IV The probability of success has to be the same for each trial. ✓
V All outcomes in the sample space are equally likely. ✗

(a) III and IV (b) II, III, IV, and V (c) I and III
(d) I, III, and V (e) II and III

c **10.** If you buy one ticket in the Provincial Lottery, then the probability that you will win a prize is 0.11. Given the nature of lotteries, the probability of winning is independent from month to month. If you buy one ticket each month for five months, what is the probability that you will win at least one prize?

(a) 0.55
(b) 0.50
(c) 0.44
(d) 0.45
(e) 0.56

Bin ($n = 5$, $p = 0.11$)

$\binom{n}{k} p^k (1-p)^{n-k}$

$P(X \geq 1) = 1 - P(X < 1)$
$= 1 - P(X = 0)$
$= 1 - [\text{binompdf}(n = 5, p = 0.11, k = 0)]$

Part 2: Free Response

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

11. Picard Partners is planning a major investment. The amount of profit X is uncertain but a probabilistic estimate gives the following distribution (in millions of dollars):

X	1	2	4	10
P(X)	0.2	0.5	0.2	0.1

(a) Find an interpret the mean (expected value) of X.

$\mu_X = E(X) = \Sigma x_i p_i = 3$ Picard Partners is expected to earn a profit of approximately \$3 million on any given investment.

(b) Find and interpret the standard deviation of X.

$\sigma_X = \sqrt{\Sigma (x_i - \mu_X)^2 p_i} = 2.53$ A randomly chosen investment profit will typically vary from the mean profit by about \$2.5 million.

(c) Picard owes its source of capital a fee of \$200,000 plus 10% of the profits X. So the firm actually retains $Y = 0.9X - 0.2$ from the investment. Use a linear transformation of your results in (a) and (b) to find the mean and standard deviation for Y.

$\mu_Y = \mu_{0.9x-0.2} = 0.9\mu_x - 0.2 = 0.9(3) - 0.2 = 2.5$

$\sigma_Y = \sigma_{0.9x-0.2} = \sigma_{0.9x} = 0.9\sigma_x = 0.9(2.53) = 2.277$

12. The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and standard deviation 16 days. Choose two pregnancies independently and at random.

(a) What is the expected difference in the lengths of the two pregnancies? Show your work.

Let X_1 = pregnancy 1 and X_2 = pregnancy 2

$Y = X_1 - X_2$ $E(Y) = \mu_Y = \mu_{X_1 - X_2} = \mu_{X_1} - \mu_{X_2} = 266 - 266 = 0$

We expect there to be no difference in the duration of two pregnancies.

(b) What is the standard deviation of the difference in the lengths of the two pregnancies? Show your work

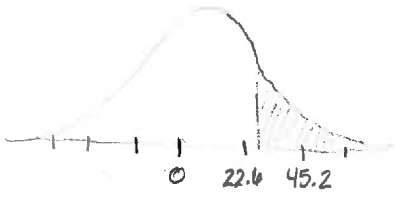
$\sigma_Y = \sigma_{X_1 - X_2} = \sqrt{\sigma_{X_1}^2 + \sigma_{X_2}^2} = \sqrt{16^2 + 16^2} = 22.6$

The typical difference in length of two pregnancies typically differs from 0 be about 22.6 days.

(c) Find the probability that the difference in the lengths of the two pregnancies is greater than 25 days. Show your work

Difference $N(0, 22.6)$

$P(Y > 25)$ normalcdf(lower: 25, upper: 10^{10}, μ: 0, σ: 22.6) = 0.1343



The probability the difference in lengths of two randomly chosen pregnancies is greater than 25 days is approximately 13%.

13. Witney Pete, a professional dart player, has a 70% chance of hitting the bull's eye on a dartboard with any throw. Assume that each throw of a dart is independent.

(a) Suppose Pete throws darts until he hits his first bull's eye. Find the probability that his first bull's eye occurs on the third throw.

Geometric $p=0.7$

$P(Y=3) = (1-0.7)^2(0.7)$ or geompdf$(p=0.7, K=3) = 0.063$

$= (0.3)^2(0.7)$

The probability it will take Pete 3 throws to hit his first bull's eye is about 6%.

(b) What is the probability that Pete hits 5 or fewer of his next 10 shots?

binomial $n=10$, $p=0.7$

$P(X \leq 5) = P(X=0) + P(X=1) + P(X=2) + \ldots + P(X=5)$

$\binom{n}{k} p^k (1-p)^{n-k}$ binomcdf$(n=10, p=0.7, k=5) = 0.15$

The probability Pete hits 5 or fewer shots on his next 10 throws is about 15%

(c) Pete forgets his eyeglasses one evening, but he's confident he's just as accurate without them. On his first 10 shots of the night, he hits the bull's eye only 5 times. Is this evidence that his glasses are important? Explain.

$P(X=5)$ with eyeglasses

$P(X \leq 5)$ with glasses is 15%.

bin$(n=10, p=0.7)$

$\binom{n}{k} p^k (1-p)^{n-k}$

binompdf$(n=10, p=0.7, k=5) = 0.10$

$\mu_x = E(x) = np = 10(0.7) = 7$

$\sigma_x = \sqrt{np(1-p)} = \sqrt{10(0.7)(0.3)} = 1.45$

The probability Pete hits the bull's eye exactly 5 times with his eye glasses is about 10%. We expect him to hit approximately 7 bull's eyes of of 10 shots with his probability of 70%. However, on a randomly selected evening, the number of bull's eyes he hits will typically vary from the mean by 1.45. Also, there is a 15% chance he will hit 5 or less out of 10 throws. This is convincing evidence that he probably does not perform as well without his glasses.