Montogue

AP Statistics

40+6 Practice Problems

Lucas Monteiro Nogueira

SECTION I

1. The director of a forest conservation non-profit has just commissioned a data scientist to gather information on the donations the organization received in the past few months. The scientist says that in the month of July, of the 100 donations received, the average donation was \$180 with a standard deviation of \$36. Which of the following is the most appropriate statement?

(A) This July, the average donation is \$180.

(B) This July, 50% of all donations were more than \$180.

(C) We are 95% confident that the average donation in July was approximately between \$174 and \$186.

(D) We are 95% confident that the average donation in July was approximately between \$108 and \$252.

(E) This July, about 95% of donations were between \$108 and \$252.

2. A team of market forecasters working for a car manufacturer evaluates the revenue potential of two new vehicle designs. The team feels that Design A will yield \$500 million in revenue with a probability of 0.6 or \$400 million with a probability of 0.4. Likewise, they estimate that Design B will generate \$600 million in revenue with a probability of 0.2, \$500 million with a probability of 0.3, or \$400 million with a probability of 0.5. With reference to these data, which design should the car maker choose to produce?

(A) Design A is a better choice because its expected revenue value is greater than that of Design B.

(B) Design B is a better choice because its expected revenue value is greater than that of Design A.

(C) Design A is a better choice because the standard deviation of its revenue is greater than that of Design B.

(D) Design B is a better choice because the standard deviation of its revenue is greater than that of Design A.

(E) The expected revenue is the same regardless of which design the company chooses.

3. Which of the following is a discrete random variable?

(A) The daytime temperature in a city.

(B) The body mass of newborn children.

(C) The height of pine trees in a boreal forest.

(D) The distribution of wind velocities in a coastal area.

(E) The number of times a student guesses the answers to questions on a test.

4. A teacher's union and a school district are negotiating salaries for the coming year. The teachers want more money, and the school district administrators want to pay as little as possible. Before debate can begin, the salaries of all teachers are gleaned to figure out, on average, how much they currently earn. Which of the following measurements will the teachers and the school administrators likely quote during negotiations?
(A) Teachers will most likely use the arithmetic mean of salaries, while school administrators will most likely use the harmonic mean of salaries.
(B) Teachers will most likely use the arithmetic mean of salaries, while school administrators will most likely use the geometric mean of salaries.
(C) Teachers will most likely use the harmonic mean of salaries, while school administrators will most likely use the arithmetic mean of salaries.
(D) Teachers will most likely use the harmonic mean of salaries, while school administrators will most likely use the geometric mean of salaries.
(E) Teachers will most likely use the geometric mean of salaries, while school administrators will most likely use the arithmetic mean of salaries.

5. One advantage to using observational studies as opposed to experiments is that
(A) Observational studies involve use of randomization.
(B) Concluding cause and effect is generally easier from observational studies.
(C) Observational studies are generally not subject to bias.
(D) Observational studies are cheaper to conduct.
(E) Observational studies can make use of stratification.

6. Suppose X and Y are random variables with $E(X)$ = 26, $var(X)$ = 11, $E(Y)$ = 24, and $var(Y)$ = 14, where E denotes expected value and var denotes variance. What are the expected value and variance of the random variable $X + Y$?
(A) $E(X + Y) = 25$, $var(X + Y) = 25$.
(B) $E(X + Y) = 25$, $var(X + Y) = 317$
(C) $E(X + Y) = 50$, $var(X + Y) = 25$.
(D) $E(X + Y) = 50$, $var(X + Y) = 317$.
(E) There is insufficient information to answer this question.

7. A New York company provides financial and logistic aid to American graduate students who want to attend one of the UK's two "Oxbridge" universities. They estimate that, in each semester, converting costs in pounds to dollars, students will expend \$1800 in books, with a standard deviation of \$300; \$5000 in room and board, with a standard deviation of \$500; and \$600 in personal expenditures, with a standard deviation of \$90. Assuming independence among categories, what is the standard deviation of the total student costs?
(A) \$30
(B) \$450
(C) \$590
(D) \$700
(E) \$5350

8. You are designing an experiment with one treatment and one control group. You are blocking for two different variables, gender (male, female) and blood type in the MN system (M, N, MN). If you want each group to contain 25 subjects, what is the total number of subjects needed for the experiment?
(A) 150
(B) 175
(C) 300
(D) 325
(E) 600

9. Two events M and N are such that $P(M) = 0.6$, $P(N) = 0.3$, and $P(M \cap N) = 0.18$. Which of the following is a correct conclusion?
(A) Events M and N are both independent and mutually exclusive.
(B) Events M and N are neither independent nor mutually exclusive.
(C) Events M and N are mutually exclusive but not independent.
(D) Events M and N are independent but not mutually exclusive.
(E) Events M and N are independent, but there is insufficient information to determine whether or not they are mutually exclusive.

10. It is estimated that 1 in every 20 people with healthy teeth will experience caries within the next year. In a group of 50 people with healthy teeth, what is the probability that at least 3 of them will experience caries within the next year?
(A) 0.28
(B) 0.46
(C) 0.54
(D) 0.58
(E) 0.72

11. Consider two six-sided dice, one of which is fair (i.e., such that probabilities $P(1) = P(2) = \ldots = P(6) = 1/6$), another of which is biased such that $P(1) = 1/2$ and $P(2) = P(3) = \ldots = P(6) = 1/10$. One of the dice is randomly selected and thrown twice, yielding 1 in both cases. What is the probability that the die chosen was the biased die?
(A) 65%
(B) 70%
(C) 80%
(D) 90%
(E) 95%

12. Cloud wants to send a letter to Tifa. The probability that Cloud will indeed write the letter is 0.8. The probability that the postal service will not lose it is 0.9. The probability that the mailman will deliver the letter is 0.9. Given that Tifa did not receive the letter, what is the probability that Cloud did not write a letter at all?
(A) 39.6%
(B) 48.5%
(C) 56.8%
(D) 64.5%
(E) 71.4%

13. According to the central limit theorem, the sample mean $E(X)$ of a random variable X is approximately normally distributed if
(A) The sample is small and X is normally distributed.
(B) The sample is small and X is represented by any type of distribution function.
(C) The sample is large and X is normally distributed.
(D) The sample is large and X is represented by any type of distribution function.
(E) Regardless of sample size and regardless of the PDF that represents X.

14. In a large population of military recruits, the time required to run a mile is skewed right with a mean of 360 seconds and a standard deviation of 30 seconds. If random samples of size 9 are repeatedly drawn from this population, which of the following appropriately describes the sampling distribution of these sample means?
(A) The shape is unknown with a mean of 330 and a standard deviation of 30.
(B) The shape is unknown with a mean of 360 and a standard deviation of 10.
(C) The shape is somewhat skewed right with a mean of 360 and a standard deviation of 30.
(D) The shape is somewhat skewed right with a mean of 360 and a standard deviation of 10.
(E) The shape is approximately normal with a mean of 360 and a standard deviation of 10.

15. The mean and standard deviation of a normally distributed dataset are 50 and 8, respectively. 50 is subtracted from every term in the dataset and then the result is divided by 8. Which of the following **best** describes the resulting distribution?
(A) It has a mean of 0, a standard deviation of 8, and its shape is unknown.
(B) It has a mean of 0, a standard deviation of 8, and its shape is normal.
(C) It has a mean of 1 and a standard deviation of 0.
(D) It has a mean of 0, a standard deviation of 1, and its shape is unknown.
(E) It has a mean of 0, a standard deviation of 1, and its shape is normal.

16. When all the values of a dataset are the same and different from zero, all but one of the following measurements equals 0. Which one is it?

(A) Range

(B) Mean

(C) Standard deviation

(D) Variance

(E) Interquartile range

17. Consider the following measurements.

I. Mean

II. Standard deviation

III. Skewness

Which of these measurements can be negative?

(A) I only.

(B) III only.

(C) I and II.

(D) I and III.

(E) II and III.

18. Which of the following samples does not suffer from undercoverage bias?

(A) A sample of adults from a city's phonebooks.

(B) A sample of motorists from a local traffic school.

(C) A sample of students drawn from biology classes.

(D) A sample of homeowners drawn from county property tax records.

(E) A sample of geologists from a list of employees of the US Geological Survey.

19. When a student newspaper in a local college ran a poll asking students if they had smoked marijuana in the past semester, 38% replied yes. When a similar poll was taken by the college administration, only 13% of students said they had smoked marijuana in the past semester. What type(s) of bias caused the large difference in the results?

I. Sampling bias

II. Wording bias

III. Response bias

(A) II only.

(B) III only.

(C) II and III only.

(D) I and II only.

(E) I and III only.

20. After a school shooting in a small town, a local disc jockey dedicates his radio program to have residents phone in and voice their opinions on a possible nationwide ban on the sale of firearms to civilians. After listening to 30 people, the DJ gleans opinions and, in view of the fact that 20 of the 30 callers were in favor of the ban, concludes that the local population supports the new measure. What can we say about this?

(A) The sample used was simple random because the DJ allowed all local residents to participate, used no biased mechanism to filter them, and took a sizable amount of opinions.

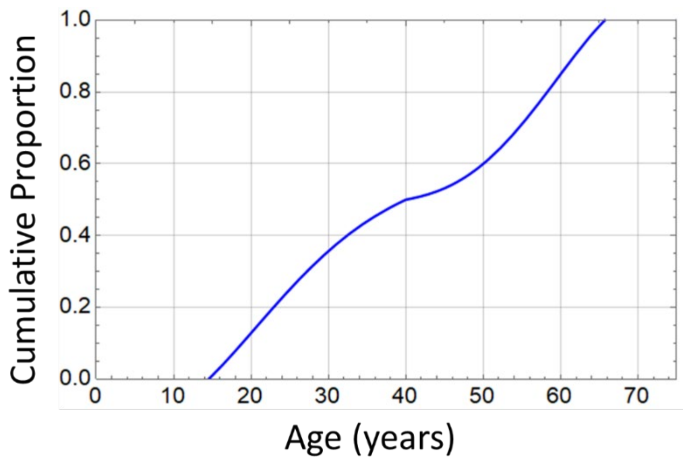
(B) No meaningful conclusion is possible without knowing something more about the characteristics of the callers.

(C) The survey would have been more meaningful if the DJ had picked a random sample of the 30 callers.

(D) The survey would have been more meaningful if the DJ had used a control group.

(E) The survey is meaningless because of voluntary response bias.

21. The following graph shows the cumulative proportions plotted against average age (in years) of subscribers of a local newspaper. What is the median subscriber age?



(A) 32

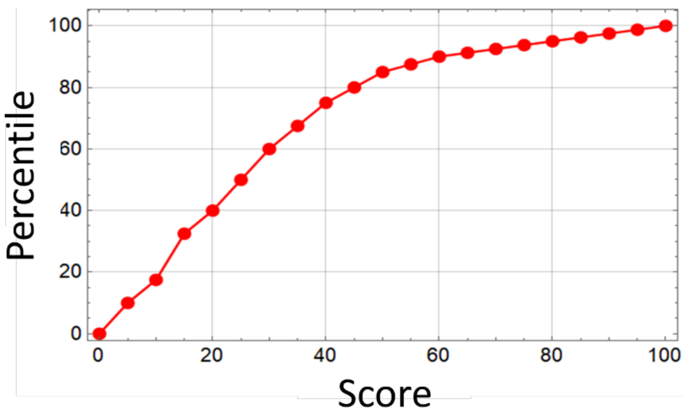
(B) 40

(C) 50

(D) 62

(E) 68

22. Illustrated below is a cumulative frequency plot for the scores of American students of a certain university in Level C1 of the DALF French language aptitude test. The minimum score is zero, and the maximum is 100. Which of the following observations is correct?



(A) The distribution of scores is skewed to the left.
(B) The distribution of scores is skewed to the right.
(C) The distribution of scores is roughly symmetric.
(D) The median score is greater than 60.
(E) If the passing score is 50, most students passed the test.

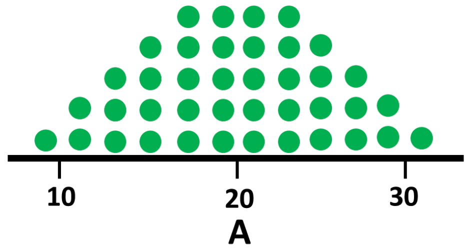
23. Argentina and Brazil are the world's second and fifth largest consumers of meat, respectively. The following back-to-back stem plot shows the annual meat consumption amounts in kilograms for a random sample of Argentinians, to the left, and Brazilians, to the right.

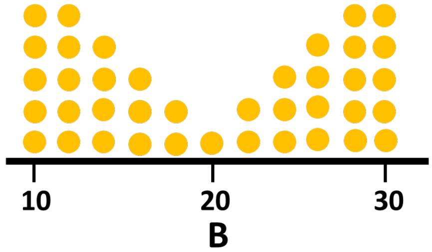

9	5	6 7 7 9 9
4 4 4	6	0 1 1 1 3 4
4 3 1	7	0 1 1 3 4
9 7 7 6	7	7 8
4 4 1 1 1 0 0	8	0 2 3 4
9 8 7 7 6	8	5 6
1	9	

Regarding this plot, which of the following statements is/are correct?
I. The range of the sample of Argentinians is 32.
II. The range of the sample of Brazilians is 27.
III. The mode of the sample of Brazilians is 61.

(A) I only.
(B) II only.
(C) III only.
(D) I and II only.
(E) I and III only.

24. Considering the two following dot plots, which of the following is true?





Considering the two following dot plots, which of the following is true?
(A) Both sets have roughly the same variance.
(B) The mean of plot A is seemingly greater than the mean of plot B.
(C) The mean of plot B is seemingly greater than the mean of plot A.
(D) The empirical rule applies only to set B.
(E) The standard deviation of set A is less than 6.

25. A group of biologists begins monitoring the breeding behavior in a population of catadromous fish. The scientists want to estimate the proportion of fish that will return to the sea in the next mating season. In the first phase of their study, the scientists track a sample of 1,200 individuals representative of the fish population. Due to financial constraints, the scientists are forced to reduce the sample size to 600 in the second phase of the study. How will this affect the distribution of the sample proportion?
(A) The distribution of the sample proportion will more closely resemble the chi-square distribution.
(B) The distribution of the sample proportion will more closely resemble the binomial distribution.
(C) The mean of the distribution of the sample proportion will double when the sample size is halved.
(D) The distribution of the sample proportion will be more spread out.
(E) The distribution of the sample proportion will be less spread out.

26. One paleontological method estimates that there is a 95% confidence level that populations of *Homo erectus*, an early evolutionary descendent of humans, were already present in Africa by 1.6 $\pm$ 0.4 mya (million years ago). What is meant by "95% confidence" in this context?
(A) A confidence interval of the age of *H. erectus* populations has been calculated using z-scores of $\pm$1.96.
(B) A confidence interval of the age of *H. erectus* populations has been calculated using t-scores consistent with a number of degrees of freedom df = $n - 1$, where n is the number of data points, and tail probabilities of $\pm$0.025.
(C) There is a 95% probability that *H. erectus* were present in Africa between 1.2 and 2.0 mya.
(D) If 100 random samples of *H. erectus* paleontological data are gathered by the method in question and a 95% confidence interval is calculated from each sample, the actual age of *H. erectus* populations will be in 95 of these samples.
(E) Of all the random samples of paleontological data obtained by the method in question, 95% will yield intervals that capture the true age of *H. erectus* populations.

27. The mean length of all screws being produced by a certain factory is assumed to be 10 centimeters. The factory operators do not know what the population standard deviation is. A sample of 25 screws is drawn, and the mean and sample standard deviation of the screw lengths are determined to be 9.6 and 0.8 centimeters, respectively. Which of the following is the 95% confidence interval for the screw lengths' population mean?
(A) (9.15, 10.0) cm
(B) (9.27, 9.93) cm
(C) (9.33, 9.87) cm
(D) (9.42, 9.98) cm
(E) (9.53, 10.1) cm

28. A country's health ministry wants to estimate, using a 95% confidence interval, the proportion of smokers that would actually enroll in a massive smoking cessation program sponsored by the government. *Of the following choices*, which is the smallest sample size of interviewees they need to question in order to ensure a margin of sampling error less than or equal to 3 percentage points?
(A) 700
(B) 1100
(C) 1500
(D) 1900
(E) 2000

29. When leaving for school on an overcast morning, you make a judgement on the null hypothesis: *the weather will remain dry*. What would be the results of Type I and Type II errors?
(A) Type I error: *Carry an umbrella and it rains.* Type II error: *Carry no umbrella, but weather remains dry*.
(B) Type I error: *Get drenched.* Type II error: *Carry no umbrella, but weather remains dry*.
(C) Type I error: *Get drenched.* Type II error: *Carry an umbrella, and it rains.*
(D) Type I error: *Get drenched.* Type II error: *Needlessly carry around an umbrella.*
(E) Type I error: *Needlessly carry around an umbrella.* Type II error: *Get drenched.*

30. Researchers are conducting an experiment using a significance level of 0.05. The null hypothesis is, in fact, false. If they modify their experiment to use twice as many experimental units for each treatment, which of the following would be true?
(A) The probability of a Type I error would stay the same and the power would increase.
(B) The probability of a Type II error would stay the same and the power would increase.
(C) The probability of a Type I error and the power would both decrease.
(D) The probability of a Type II error and the power would both decrease.
(E) The probability of a Type I error and the probability of a Type II error would both decrease.

31. Regarding the p-value, which of the following is true?
(A) The p-value is a conditional probability.
(B) The p-value is usually chosen before an experiment is conducted.
(C) The p-value is based on a specific test and thus should not be used in a two-sided test.
(D) p-Values are more appropriately used with t-distributions than with z-distributions.
(E) If the p-value is less than the level of significance, the null hypothesis is proved false.

32. After a cruise jet full of people crashes in American air space, the jet's manufacturer conducts an independent investigation and estimates that the number of casualties should be around 30. A skeptical statistician at the Federal Aviation Administration draws a random sample of 16 accidents in the same region with the same aircraft and estimates that the real number should be around 45, with a standard deviation of 9.5. What is the p-value?
(A) $P(z > 45 - 30(9.5/\sqrt{16}))$
(B) $2P(z > 45 - 30(9.5/\sqrt{16}))$
(C) $P(t > 45 - 30(9.5/\sqrt{16}))$ with $df = 15$
(D) $2P(t > 45 - 30(9.5/\sqrt{16}))$ with $df = 15$
(E) $P(t > 45 - 30(9.5/\sqrt{16}))$ with $df = 16$

33. A student is researching prices for statistical inference textbooks on Amazon and Barnes & Noble. To compare prices, the student picks 20 titles and checks their prices at the two e-stores. Which test should the student use to determine if the prices are different in the two sites?
(A) Two-sample z-test.
(B) Two-sample t-test.
(C) Matched pairs t-test.
(D) χ^2-test for goodness-of-fit.
(E) χ^2-test for independence.

34. A group of psychologists and political scientists is screening American adults to determine if political inclination is related to color preference. A random sample of people, classified according to their political leaning as *liberal*, *moderate* or *conservative*, was given 7 choices of favorite color; each person could choose only one favorite color. The results are tabulated below.

	Liberal	**Moderate**	**Conservative**
Black	5	4	4
Blue	7	2	1
Green	2	3	4
Purple	6	10	2
Red	1	2	11
White	5	6	5
Yellow	4	3	3
Total	30	30	30

How many degrees of freedom would be used in a test of independence of political standing and color preference?
(A) 12
(B) 14
(C) 18
(D) 21
(E) 119

35. The owner of a local motorcycle dealer believes that each of the five brands it works with are equally successful in producing motorcycle sales. During the course of one month, the following sales were observed:

BMW	Honda	Yamaha	Harley-Davidson	Ducati
24	29	34	16	17

Which of the following would be the value of the χ^2 statistic for the goodness-of-fit test, testing the null hypothesis that the motorcycle dealers are equally preferred?
(A) 9.72
(B) 9.92
(C) 10.1
(D) 10.6
(E) 24

36. Data were collected on two variables x and y and a least squares regression line was fitted to the data. The estimated equation for this data is $y = -1.81 + 0.45x$. One point has $x = 8$, $y = 1.8$. What is the residual of this point?
(A) −0.01
(B) +0.01
(C) −1.79
(D) +1.79
(E) The information is not sufficient to determine the residual of the point in question.

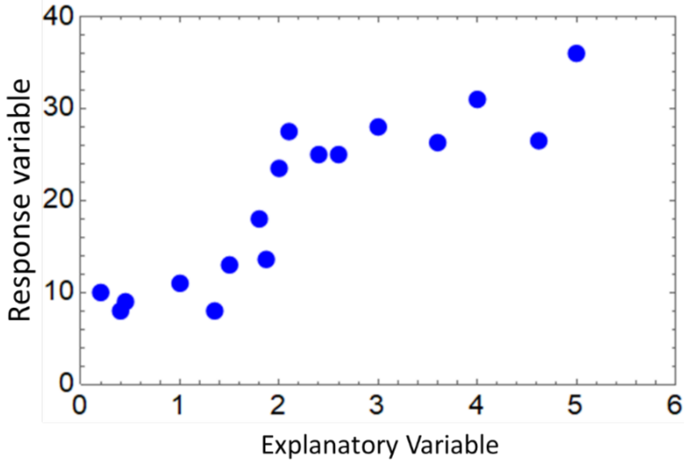
37. Suppose that the correlation between a set of scores X and a set of scores Y is equal to 0.7. If the scores are reversed, so that the X scores become the Y scores and the Y scores become the X scores, what will the new correlation between the scores be?
(A) −0.7
(B) −0.3
(C) 0.3
(D) 0.7
(E) 1/0.7

38. A mechanical engineer working for an industry research group has investigated 20 popular sedan models and found that the maximum velocity in mph they can achieve is linearly proportional to powertrain horsepower (hp). Given the power x, the top speed y can be estimated by the linear fit

$$y = 0.631x + 11$$

The correlation coefficient obtained was 0.881. A friend of the engineer, who owns a luxury Audi A8 sedan, couldn't help but notice that the linear fit above greatly underestimates the top speed of his vehicle. Which of the following best explains this underprediction?
(A) The closer the correlation coefficient is to 1, the worse a linear fit will be. Accordingly, the correlation coefficient obtained by the engineer is high and hence one can only expect that the linear fit will yield inaccurate results for a number of sedan models.
(B) Luxury sedans were not included in the vehicle sample; only popular models were assessed. Accordingly, the power rating of a vehicle such as the A8 extrapolates the range of x values encompassed by the engineer's data.
(C) The vehicle sample was too small; the engineer should include another 20 popular sedans, preferably within the same range of power ratings that he used in the original analysis, and establish a new least-squares line.
(D) The engineer was jealous of his friend and did not include his luxury vehicle in the dataset; thus, this is a typical instance of a study marred by social comparison bias.
(E) There was an error made in the computation of the original least-squares line; the engineer should check his results and repeat his calculations.

39. What is the correct regression output for the scatterplot below?



(A)

Predictor	Coef	Stdev	t	p
Constant	2.89451	1.93445	3.79857	0.00175
Explanatory	−5.66063	0.735974	7.69135	0.00001
s = 0.8033	R-sq = 100.0		R-sq(adj) = 99.8	

(B)

Predictor	Coef	Stdev	t	p
Constant	2.89451	1.93445	3.79857	0.00175
Explanatory	−5.66063	0.735974	7.69135	0.00001
s = 0.8033	R-sq = 79.8		R-sq(adj) = 73.5	

(C)

Predictor	Coef	Stdev	t	p
Constant	−2.89451	1.93445	3.79857	0.00175
Explanatory	5.66063	0.735974	7.69135	0.00001
s = 0.8033	R-sq = 79.8		R-sq(adj) = 73.5	

(D)

Predictor	Coef	Stdev	t	p
Constant	2.89451	1.93445	3.79857	0.00175
Explanatory	5.66063	0.735974	7.69135	0.00001
s = 0.8033	R-sq = 79.8		R-sq(adj) = 73.5	

(E)

Predictor	Coef	Stdev	t	p
Constant	2.89451	1.93445	3.79857	0.00175
Explanatory	5.66063	0.735974	7.69135	0.00001
s = 0.8033	R-sq = 100.0		R-sq(adj) = 99.8	

40. A marine biologist is monitoring a population of a species of starfish in a nearby coast. He has established that, in a population of 12 starfish, there is a strong linear relationship between age, expressed in years, and radius, expressed in cm.

Predictor	Coef	Stdev	t	p
Constant	10.8485	1.87115	−3.1314	0.004
Age	1.15152	0.08923	14.5698	0.000
s = 0.88			R-sq = 90.1	

Which of the following would represent a 95% confidence interval for the slope of the regression line?

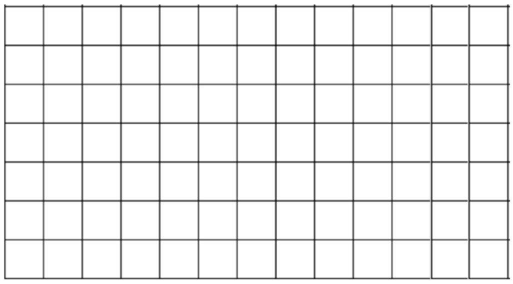
(A) 1.15 ± 0.111
(B) 1.15 ± 0.162
(C) 1.15 ± 0.199
(D) 1.24 ± 0.162
(E) 1.24 ± 0.199

SECTION II – Part A

1. The following data shows the number of wildfires detected by a satellite monitoring technique within a small county in the Pacific Northwest. The data concern the three summer months and summarize information gleaned in the past seven years.

	2014	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
June	50	150	50	400	350	500	250
July	400	450	450	100	550	150	200
August	100	50	100	350	300	350	600

A. Draw three parallel boxplots describing the number of fires for June, July and August over the years monitored. Use the grid below to save time.



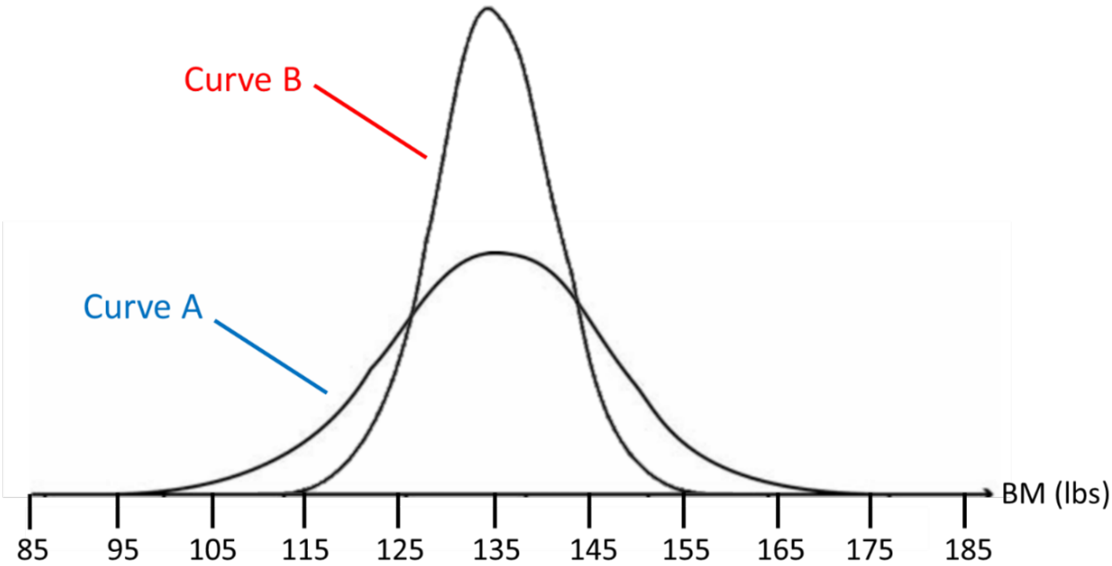
B. Which of the summer months has a median number of fires, over the years monitored, greater than 350?
C. Which of the summer months has more than one mode for number of fires over the years monitored? What is the term attributed to such a dataset?
D. Which of the summer months has the greatest interquartile range (IQR) for number of fires over the years monitored?
E. Suppose 500 fires were detected in the county during the month of July 2021, so the dataset for this month now contains 6 measurements. Will this affect the median for the month in question? Explain.

2. A group of labor statisticians has been commissioned by the American Society of Civil Engineers to produce a career outlook for civil engineering in the USA for the near future and hopefully attract more college students to the profession. As part of their work, the researchers looked at the association between mean annual wage and number of years of work experience. They hoped to relate average yearly winnings to the number of years an engineer has been in the market. The statistical summary of the data from a random sample of 2000 engineers is shown below.

Variable	**Mean**	**SD**
Mean annual wage ($)	89,060	26,500
Work experience (years)	8	7.08
r **= 0.58**		

A. What is the slope of the regression line? Interpret it in context.
B. What is the equation of the least squares regression line?
C. Alfred says he will only crack the books to get into a civil engineering school if he manages to make over $80,000 a year with six years of experience or less. According to the research ordered by the ASCE, is this a realistic prospect?

3. The population of adolescents of a certain age in country 1 has normally distributed body mass with a mean of 135 lbs and a standard deviation of 31 lbs.
A. If an adolescent is randomly chosen from this population, what is the probability that he/she will have a weight between 100 and 140 lbs?
B. What body mass marks the upper 10% of the distribution of adolescents in this population?
C. The body mass of adolescents in country 2 is also normally distributed with a mean of 135 lbs, but the standard deviation in this case is 45 lbs. Aware of this information, a AP Statistics student says that if the distribution of weights for adolescents of country 1 can be represented by Gaussian curve A below, it follows that the distribution of weights for adolescents in country 2 can be represented by Gaussian curve B. Is the student correct? If not, why?



4. Four fifths of the members of a sports club are adults, and one fifth are children. Three quarters of the adults, and three fifths of the children, are male. Half the adult males, and a quarter of the adult females, use the swimming pool at the club; the corresponding proportion for children of either sex is three quarters.
A. Find the probability that a member of the club is female.
B. Find the probability that a member of the club uses the swimming pool.
C. Find the probability that a member of the club who does not use the swimming pool is either female or an adult.

5. In an opinion poll, randomly chosen Californians and Texans were asked about their thoughts (*pro* or *contra*) about the federal legalization of euthanasia for terminally ill patients.

	Pro	***Contra***
California	94	31
Texas	56	38

Assuming independence between rows and columns, investigate whether there is a significant difference between California and Texas in opinions about euthanasia.

SECTION II – Part B

1. Research conducted in the United States indicates that about 41.8% of marriages in the country end in divorce. XY is reportedly the state with the highest divorce rate. Aiming to study this phenomenon further, a group of researchers working for a psychology research periodical will sample 81 couples from this state, follow their daily lives for a time, and eventually ascertain whether divorces in the region are more common than the national average.

Let p denote the proportion of marriages in XY that end in divorce. The hypotheses of interest are the following:

$H_0: p = 0.418$	$H_a: p > 0.418$

A. Describe a Type II error in the context of the present investigation and a possible consequence.

B. What values of the sample proportion $\hat{p}$ would represent sufficient evidence to reject the null hypothesis at a significance level $\alpha = 0.05$?

Suppose the actual proportion of marriages in XY that end in divorce is 46%.

C. Using the actual proportion of 0.46 and the result from (B), find the probability that the null hypothesis will be rejected. Show your work.

D. What statistical term describes the probability calculated in (C)?

E. Suppose the size of the sample was greater than 81. How would that affect the probability of rejecting the null hypothesis calculated in (C)? Explain.

Solutions

→ Section I

1.A. Since the non-profit has all the data for November, there is no inference to do regarding the donations for the month of interest; this excludes alternatives (C) to (E). Alternative (B) is also false, because 50% of donations are above the *median*, not the mean. Option (A) is the most we can infer with the data at hand.

2.B. The revenue values are discrete random variables, and the best design choice will be the one that has the greatest expected revenue; standard deviation has little bearing on a first analysis of which design to produce, which excludes alternatives (C) and (D). The expected revenue from Design A is

$$E(X_A) = 0.6 \times 500 + 0.4 \times 400 = \$460\,\text{million}$$

The expected revenue from Design B, in turn, is

$$E(X_B) = 0.2 \times 600 + 0.3 \times 500 + 0.5 \times 400 = \$470\,\text{million}$$

Since $E(X_B) > E(X_A)$, it would be a better choice for the car maker to work with Design B.

3.E. A discrete random variable must be countable; length/height (alt. C), velocity (alt. D), mass (alt. B) and temperature (alt. A) are usually taken as continuous, non-countable measurements in most settings of statistical interest. The number of times a student guesses the answers to questions on a test, on the other hand, necessarily takes integer values 0, 1, 2, ..., n, where n is the number of questions in the test, and can thus be taken as a discrete RV.

4.C. An elementary inequality tells us that the arithmetic mean is always greater than or equal to the geometric mean, which in turn is always greater than or equal to the harmonic mean. During negotiations, teachers will likely quote the lowest possible measurement, i.e., the measurement that suggests that they are in fact underpaid; thus, they will probably use the harmonic mean. Administrators, in turn, will prefer using the measurement that suggests that they are well paid as it is; thus, they will probably quote the arithmetic mean.

5.D. Observational studies are generally cheaper and quicker to conduct than experiments. However, such studies are subject to bias, and it is very difficult to conclude cause and effect from observational studies; accordingly, statements (B) and (C) are false. Obviously, experiments also rely on randomization, which invalidates statement (A). Blocking in experimental design roughly corresponds to stratification in sampling design, thus alternative (E) is also false.

6.E. We can state that $E(X+Y) = E(X) + E(Y) = 40 + 56 = 96$, but without *independence* we cannot establish the value of $var(X+Y)$.

7.C. With independence, variances add; accordingly, the total student costs should have a standard deviation such that

$$\sigma = \sqrt{120^2 + 500^2 + 600^2} = \boxed{\$590}$$

8.C. Blocking for gender (male, female) and blood type in the MN system (M, N, MN) results in six blocks: male-M, male-N, male-MN, female-M, female-N, female-MN. Since each block is to be represented by two groups, a treatment group and a control group, we have 6 × 2 = 12 groups. Each group has 25 subjects, so the total number of subjects required becomes 12 × 25 = 300.

9.D. Note that $0.6 \times 0.3 = 0.18 = P(M \cap N)$; since $P(M \cap N) = P(M)P(N)$, we surmise that M and N are independent. In view of $P(M \cap N) \neq 0$, M and N are not mutually exclusive.

10.B. The probability that any one person with healthy teeth will experience caries within the next year is 1/20 = 0.05. In "at least n" probability problems, the fastest way to go is to calculate the probability of the event "less than n" and then compute the probability of the complementary event, all the while noting that $P(A) + P(A^c) =$

Method 1: Hand calculation

The probability that zero people will experience caries within the next year is

$$P(0 \text{ people develop caries}) = \binom{50}{0} \times 0.95^{50} \times 0.05^{0} = 0.0769$$

The probability that exactly one person will experience caries within the next year, in turn, is

$$P(1 \text{ person develops caries}) = \binom{50}{1} \times 0.95^{49} \times 0.05^{1} = 0.202$$

The probability that exactly two people will experience caries within the next year, in turn, is

$$P(2 \text{ people develops caries}) = \binom{50}{2} \times 0.95^{48} \times 0.05^{2} = 0.261$$

The sum of the three preceding probabilities defines event A = (Less than 3 people develop caries); mathematically,

$$P(A) = 0.0769 + 0.202 + 0.261 = 0.540$$

The probability that we are looking for is that of event A^c = (At least 3 people develop caries), which happens to be the complementary event of A. Since the sum of probabilities of an event and its complementary must amount to 1, we can state that

$$P(A)+P(A^c)=1 \to P(A^c)=1-P(A)$$

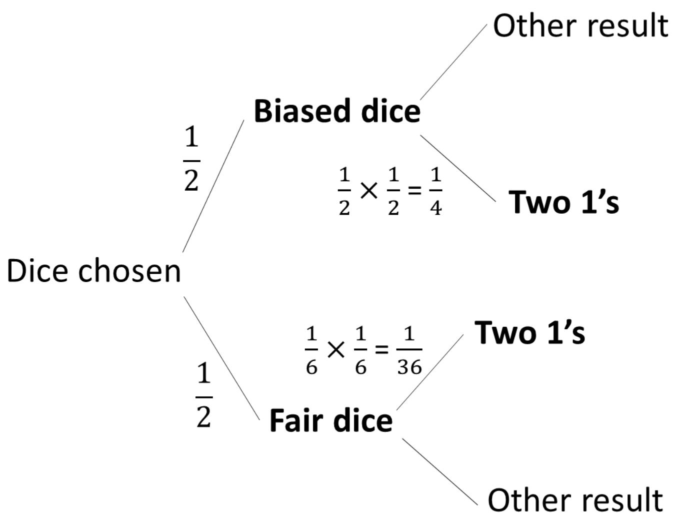
$$\therefore P(A^c)=1-0.54=\boxed{0.46}$$

Method 2: Using the TI-83/4

The probability that we are looking for can be easily determined via the TI-83/4's *binomcdf* function, as in

1 – `binomcdf`(50, 0.05, 2)

11.D. This typical conditional probability problem can be easily solved if we draw a probability tree.



The general probability that a die thrown twice will yield 1 in both throws is

$$P(\text{Two 1's observed})=\frac{1}{2}\times\frac{1}{4}+\frac{1}{2}\times\frac{1}{36}=\frac{5}{36}$$

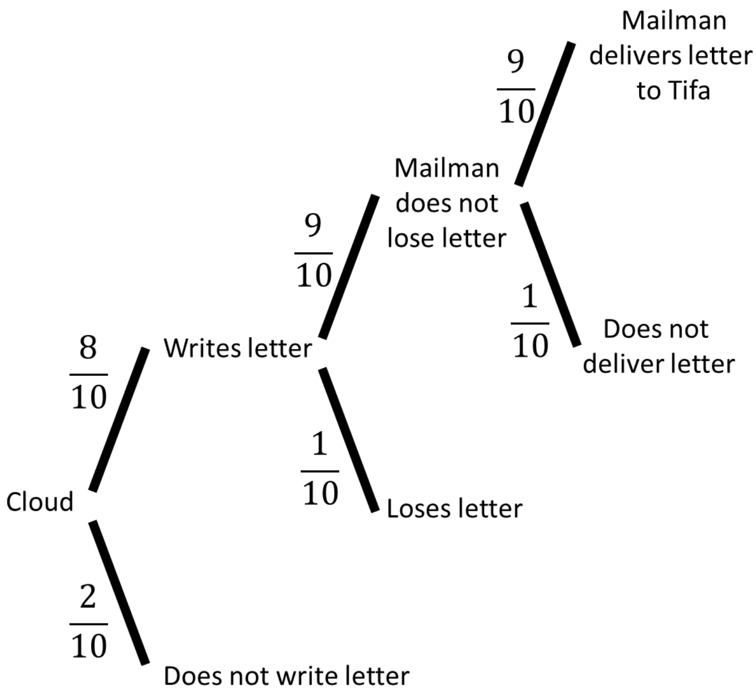
The probability that a biased die will yield 1 in two consecutive throws is

$$P(\text{Biased dice, two 1's observed})=\frac{1}{2}\times\frac{1}{4}=\frac{1}{8}$$

The probability we are looking for is

$$P=\frac{P(\text{Biased dice, two 1's observed})}{P(\text{Two 1's observed})}=\frac{1/8}{5/36}=\boxed{90\%}$$

12.C. Clearly, this is a typical conditional probability problem; the way to go is to use a tree diagram, as illustrated in continuation.



The probability we aim for is determined as

$$P(\text{Does not write}|\text{Does not receive}) = \frac{P(\text{Does not write})}{P(\text{Does not receive})}$$

$$\therefore P = \frac{0.2}{0.2 + 0.8 \times 0.1 + 0.8 \times 0.9 \times 0.1} = \boxed{56.8\%}$$

13.D. The CLT states that any random variable X will have a normally distributed mean, assuming the sample is sufficiently large. This applies regardless of the PDF that fits X.

14.D. $\bar{x}$ is an unbiased estimator of the mean and hence the sampling distribution of those sample means should have the same value as the population; thus, mean = 330. The standard deviation of the sampling distribution is given by $\sigma/\sqrt{n}$ = $30/\sqrt{9}$ = 10. However, the population is substantially skewed right and the sample size is very small, so we cannot say that the shape of the sampling distribution is approximately normal.

15.E. The effect on the mean of a dataset when we subtract the same value is to reduce the old mean by that amount (i.e., μ_{x-k} = $\mu_x - k$). Because the original mean was 50, and every data point was reduced by 50, the mean of the modified dataset will be 50 – 50 = 0. The effect on the standard deviation of a dataset as each term is divided by the same value is to divide the standard deviation by that value, so that $\sigma_{x/k}$ = σ_x/k = 8/8 = 1. The student should be aware that the process of subtracting the mean from each term and dividing by the standard deviation creates a set of z-scores

$$z_x = \frac{x - \overline{x}}{s}$$

so that any complete set of z-scores has a mean of 0 and a standard deviation of 1. The shape is normal since any linear transformation of a normal distribution will still be normal.

16.B. It goes without saying that the range of a dataset that consists of a number of repetitions of the same value equals zero; the same applies to the IQR. The standard deviation and variance depend on the sum $\Sigma(x - \bar{x})$, which will yield zero because $x - \bar{x} = 0$ for all points. However, the mean $\bar{x}$ will not be nil; rather, it will be equal to the repeating value that constitutes the dataset.

17.D. The mean of a dataset can in fact be negative if a considerable amount of data in a sample are lower than zero. The standard deviation is the square root of variance, which is always non-negative, and hence cannot be lower than zero. The skewness is a cubic moment and can assume negative values.

18.D. Every homeowner has to pay property tax; accordingly, a sample of homeowners drawn from property tax records will necessarily encompass the entire population of people who own homes. All other alternatives are marred by undercoverage bias: in (A), we know that some wealthy people do not have their names on phonebooks, and some very poor have no landlines; in (B), we know that not all motorists are in traffic school; in (C), we know that not all students are enrolled in biology courses; in (E), we know that not all licensed geologists work for the USGS.

19.B. The method by which the students were selected was probably not an issue, so we can exclude sampling bias. The question is straightforward and the way in which it was posed to students was likely similar in both polls; thus, we can reject wording bias as well. Students may have replied yes to the newspaper poll more often because they felt more comfortable admitting that they smoked marijuana to their peers; they may not have issued similar answers to the college administration's research because of fear of being judged by older adults, many of which likely hold a negative view on recreational drug use. In short, the change in results was probably due to response bias.

20.E. The sample of opinions gathered by the DJ suffers from voluntary response bias, as the people who took the effort to phone in on his show certainly had some implicit incentive to do so, such as a strong preexisting support for gun control – which was likely boosted in that community after the shooting – or perhaps a close friend or relative who was among the victims of the incident. Sampling the sample or knowing more about the callers would do little to reduce the bias of the exercise.

21.B. The median in a cumulative probability plot corresponds to a cumulative proportion of 0.5; inspecting the graph, we see that the median subscriber age is 40.

22.B. Draw a horizontal line at percentile = 50 and you`ll find that half the candidates scored no more than 25 or so, which excludes alternatives (D) and (E). The steeper part of the graph corresponds to the higher bars in a histogram; accordingly, the higher bars would be on the left with a tail stretching toward the right.

23.E. The range of the sample of Argentinians is 91 − 59 = 32. The range of the sample of Brazilians is 86 − 56 = 30. The number 61 occurs three times in the sample of Brazilians and constitutes the mode of that dataset.

24.E. Since both sets are symmetric around 20, it is easy to surmise that the mean of both plots is 20; this excludes alternatives (B) and (C). Since B is more spread out than A, we surmise that B has the greater variance; this excludes alternative (A). For bell-shaped data, about 68% of the values fall within one standard deviation of the mean, 95% within two standard deviations, and 99.7% within three standard deviations. Thus, the standard deviation of set A seems to be closer to 3 than 6.

25.D. Recall that the standard deviation of the sampling distribution of a sample proportion is given by

$$\sigma = \sqrt{\frac{p(1-p)}{n}}$$

Clearly, decreasing n, i.e., reducing sample size will increase the standard deviation, which in graphical terms translates to a more spread out dataset.

26.E. The 95% refers to the method: 95% of all intervals obtained by the method in question will capture the true population parameter. Nothing is certain about any particular set of 100 intervals. For any particular interval, the probability that it captures the true parameter is 1 or 0 depending upon whether the parameter is or isn’t in it.

27.B. The formula for the confidence interval for one population mean, using the t-distribution, is

$$\overline{x} \pm t_{n-1}\frac{s}{\sqrt{n}}$$

In the case at hand, the mean $\bar{x}$ = 9.6 cm, the sample standard deviation s = 0.8 cm, the sample size is n = 25, and the t-distribution value for df = 25 − 1 = 24 degrees of freedom is t_{24} = 2.06390, so that

$$9.6 \pm 2.06390 \times \frac{0.8}{\sqrt{25}} = 9.93,\ 9.27\,\text{cm}$$

Alternative (C) is the interval one would obtain for a 90% confidence interval, in which case t_{24} = 1.710882 and

$$9.6 \pm 1.710882 \times \frac{0.8}{\sqrt{25}} = 9.87,\ 9.33\,\text{cm}$$

Alternative (A) is the interval one would obtain for a 99% confidence interval, in which case t_{24} = 2.79694 and

$$9.6 \pm 2.79604 \times \frac{0.8}{\sqrt{25}} = 10.0,\ 9.15\,\text{cm}$$

28.B. This problem involves the usual formula

$$ME = Z\sqrt{\frac{p(1-p)}{n}}$$

The margin of error is largest for p = 0.5; having $p \neq$ 0.5 would yield even smaller ME's. For a 95% confidence interval, Z = 1.96. Substituting above and solving for n, we get

$$ME = Z\sqrt{\frac{p(1-p)}{n}} \rightarrow 0.03 = 1.96 \times \sqrt{\frac{0.5 \times (1-0.5)}{n}}$$

$$\therefore n = 1067 \approx \boxed{1100}$$

29.E. A Type I error means that the null hypothesis is correct (that is, the weather remains dry) but you end up rejecting it (i.e., you needlessly carry an umbrella). A Type II error means that the null hypothesis is wrong (it rains) but you do not reject it (you get drenched).

30.A. We did not change the significance level, which is the probability of making a type I error. Increasing the sample size is one way to increase the power of the test.

31.A. The p-value is the probability of obtaining a result as extreme as or more extreme than the one seen given that the null hypothesis is true; thus, it is a conditional probability. The p-value depends on the sample chosen, which excludes alternative (B). The p-value can very well be used in a two-sided test, assuming the indicated tail probability is doubled; thus, alternative (C) is also false. p-Values are not restricted to use with any particular distribution, which eliminates alternative (D). With a small p-value, there is evidence to reject the null hypothesis, but we're not *proving* anything.

32.D. This is a comparison of means, so the hypotheses to work with are H_0: μ = 30, H_a: $\mu \neq$ 30. The standard deviation of the sample means is 9.5/$\sqrt{16}$, with df = 16 − 1 = 15, and a two-sided test so that the p-value is twice the tail probability. Accordingly, the p-value equals 2P(45 – 30(9.5/$\sqrt{16}$)).

33.C. Two-sample tests require that the two samples being compared be independent of each other. However, in the case at hand the data occur in related pairs, namely, the prices of the same 20 titles on the two e-stores. The proper procedure is to run a one-sample test on the single variable consisting of the difference between online prices for each book.

34.A. Degrees of freedom are calculated by (*rows* – 1)(*columns* – 1); in the present case, df = (7 – 1)(3 – 1) = 12.

35.B. The value of the χ^2 statistic for a goodness-of-fit test is given by the formula

$$\chi^3 = \frac{\Sigma(O-E)^2}{E}$$

The expected values for each category, in this case, would be the average of the sample size E = (24 + 25 + 38 + 16 + 17)/5 = 24, since it is claimed that the categories should be equal. Therefore, the χ^2 statistic would be computed as

$$\chi^2 = \frac{(24-24)^2}{24} + \frac{(29-24)^2}{24} + \frac{(34-24)^2}{24} + \frac{(16-24)^2}{24} + \frac{(17-24)^2}{24} = \boxed{9.92}$$

36.B. The predicted value of y for x = 8 is

$$\hat{y} = -1.81 + 0.45 \times 8 = 1.79$$

The residual at the point in question is the difference between the y-coordinate of the point, y = 1.8, and the value given by the regression curve; that is,

$$\Delta = y - \hat{y} = 1.8 - 1.79 = \boxed{+0.01}$$

37.D. The correlation will remain unchanged.

38.B. The correlation coefficient is good, the calculations are probably fine, and one can assume that the study has no social comparison bias. The reason for the underprediction of top speed for the A8 was that the engineer worked with popular models, and the power rating range with which he worked probably does not include a value close to that of the vehicle in question. As a result, substituting the A8's power rating in the least-squares fit yields an extrapolated, inaccurate output.

39.D. For starters, the slope of the least-squares fit would be positive, which excludes alternatives (A) and (B). Further, the intercept of the line will be likewise positive, which excludes (C). Lastly, the fact that the scatterplot shows significant spread rules out the possibility of having a correlation coefficient very close to unity, as in the case of alternative (E); we are left with option (D) as the one viable answer.

40.C. The radius y is related to the starfish age x in accordance with the relationship

$$y = 1.15152x + 10.8485$$

Further, the standard deviation of the slope is Stdev = 0.08923. With *degrees of freedom* = (*sample size* – 2) = (12 – 2) = 10 and a confidence interval equal to 95%, we read 2.22814 from a t-table. The 95% confidence interval is then

$$1.15152 \pm 2.22814 \times 0.08923 = \boxed{1.15 \pm 0.199}$$

Answer Summary

1	A	21	B
2	B	22	B
3	E	23	E
4	C	24	E
5	D	25	D
6	E	26	E
7	C	27	B
8	C	28	B
9	D	29	E
10	B	30	A
11	D	31	A
12	C	32	D
13	D	33	C
14	D	34	A
15	E	35	B
16	B	36	B
17	D	37	D
18	D	38	B
19	B	39	D
20	E	40	C

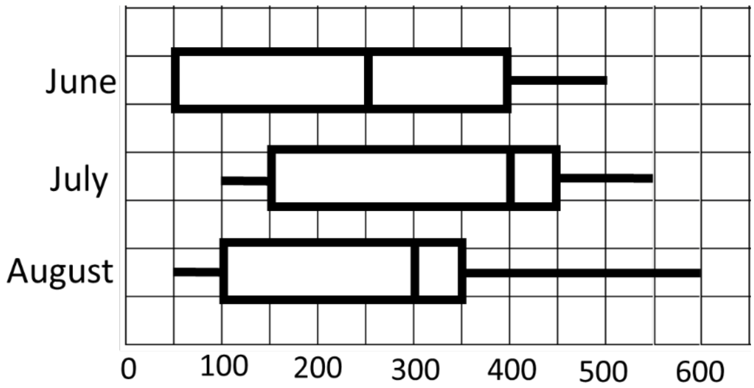
→ Section II – Part A

Problem 1

(A) The first step is to rank the data. The median is term number $(n + 1)/2 = (7 + 1)/2 = 4$. The lower quartile is the term with number $(n + 1)/4 = (7 + 1)/4 = 2$, and the upper quartile is the term with number $3(n + 1)/4 = 6$.

June	50	50	150	250	350	400	500
July	100	150	200	400	450	450	550
August	50	100	100	300	350	350	600

The boxplots are shown below.



(B) The medians for the months of June, July and August are 250, 400, and 300. The month of July has a median number of fires greater than 350.

(C) The month of August has 100 and 350 as modes. Such a dataset is called *bimodal*.

(D) The interquartile range for June is 400 – 50 = 350; the IQR for July is 450 – 150 = 300; the IQR for August is 350 – 100 = 250. Thus, the month of June is associated with the greatest IQR.

(E) The new dataset is ordered below; a 400 is shown in blue to denote the old median, and another 400 is shown in red to denote the new data point.

July	100	150	200	400	400	450	450	550

For an even number of data points, the median should be given by the average of the middlemost data points; in the case at hand, $x_m = (400 + 400)/2 = 400$. Thus, the median number of fires detected in July will remain unchanged after the data for 2021 is included in the list.

Problem 2

(A) The slope b of the regression line is

$$b = r\frac{s_y}{s_x} = 0.58 \times \frac{26,500}{7.08} = \boxed{2170.9}$$

For each additional year in the market, the engineer will have his/her MAW raised by about 2170 dollars.

(B) The least squares regression line passes through point $(\bar{x}, \bar{y})$ = (8;89,060). Thus, intercept a is calculated to be

$$b\bar{x} + a = \bar{y} \to 2170.9 \times 8 + a = 89,060$$
$$\therefore 17,370 + a = 89,060$$

$$\therefore a = 71{,}690$$

Thus, the LSRL is expressed as

$$y = 2170.9x + 71{,}690$$

(C) Substituting x = 6 into the LSRL obtained above, we obtain

$$y = 2170.9 \times 6 + 71{,}690 \approx 84{,}700$$

Per the research commissioned by the ASCE, a civil engineer with six years of experience is likely to make over $80,000/year; accordingly, Alfred should indeed consider civil engineering as a career choice.

Problem 3

(A) This problem can be easily solved in a TI-83/84; just use the syntax `normalcdf(100,140,135,31)`, which should return 0.434. However, a well-prepared student should be also able to perform such calculations with old-fashioned tables, so this is the approach we employ in this solution. We aim for the probability $P(100 \le X \le 140)$; the first step is to convert this X interval to a Z interval. For x = 100, we get

$$z = \frac{x-\mu}{\sigma} = \frac{100-135}{31} = -1.13$$

For x = 140, in turn, we get

$$z = \frac{x-\mu}{\sigma} = \frac{140-135}{31} = 0.16$$

Thus,

$$P(100 \le X \le 140) = P(-1.13 \le Z \le 0.16)$$

or

$$P(-1.13 \le Z \le 0.16) = P(Z \le 0.16) - P(Z \le -1.13)$$

Using a Z-table, we see that $P(Z \le 0.16) = 1 - P(Z \ge 0.16) = 1 - P(Z \le -0.16) = 1 - 0.4364 = 0.5636$. Moreover, $P(Z \le -1.13) = 0.1292$. Lastly,

$$P(-1.13 \le Z \le 0.16) = 0.5636 - 0.1292 = 0.4344 = \boxed{43.4\%}$$

(B) We want to find the value of X where 10% of the population lies below it. In other words, we want to find the 10th percentile of X. The problem can be easily solved in a TI-83/84; just use the syntax `invNorm(0.9,135,31)`; this should return 174.7 lbs. This computation can also be done in a Z-table if we take the usual formula

$$z = \frac{x-\mu}{\sigma}$$

and establish the value of z pertaining to a probability of 0.1. The entry closest to 0.1 in a typícal Z-table is 0.1003, for which $z = -1.28$. Substituting in the general expression above and solving for x brings to

$$z=\frac{x-\mu}{\sigma}\rightarrow -1.28=\frac{x-135}{31}$$
$$\therefore 39.7=x-135$$
$$\boxed{x=174.7\text{ lbs}}$$

The upper 10% of adolescents in the population have a body mass of 174.7 lbs or more.

(C) If the two distributions of body mass are normally distributed with the same mean μ = 135 lbs, both populations of adolescents can be represented by Gaussian curves symmetric with respect to the abscissa 135 lbs. However, because the population of adolescents of country 2 has a greater standard deviation, the Gaussian curve that represents their weights should be shorter and wider (i.e., more spread out) than the curve that represents the weights of country 1's adolescents; it follows that the AP student was incorrect, as curve A likely represents country 2's population of adolescents and curve B best illustrates country 1's population of young people, not the contrary.

Problem 4

(A) In our solution we shall let A denote the event that a member of the club is an adult, and C the event that he or she is a child. We shall let M and F stand, respectively, for the events that a member is male or female, and S for the event that he or she uses the swimming pool. The probabilities we were given can be summarized as follows.

$$P(A)=\frac{4}{5}\ ;\ P(C)=\frac{1}{5}$$
$$P(M\,|\,A)=\frac{3}{4}\ ;\ P(F\,|\,A)=\frac{1}{4}$$
$$P(M\,|\,C)=\frac{3}{5}\ ;\ P(F\,|\,C)=\frac{2}{5}$$
$$P(S\,|\,A\cap M)=\frac{1}{2}:P(S\,|\,A\cap F)=\frac{1}{4}$$
$$P(S\,|\,C\cap M)=P(S\,|\,C\cap F)=P(S\,|\,C)=\frac{3}{4}$$

The probability that a member of the club is female is given by

$$P(F)=P(A)P(F\,|\,A)+P(C)P(F\,|\,C)$$
$$\therefore P(F)=\frac{4}{5}\times\frac{1}{4}+\frac{1}{5}\times\frac{2}{5}=\frac{7}{25}=\boxed{28\%}$$

(B) The probability that a given member of the club uses the swimming pool is given by

$$P(S)=P(A\cap M\cap S)+P(A\cap F\cap S)+P(C\cap S)$$

$$\therefore P(S) = P(A)P(M \mid A)P(S \mid A \cap M) + P(A)P(F \mid A)P(S \mid A \cap F) + P(C)P(S \mid C)$$

$$\therefore P(S) = \frac{4}{5} \times \frac{3}{4} \times \frac{1}{2} + \frac{4}{5} \times \frac{1}{4} \times \frac{1}{4} + \frac{1}{5} \times \frac{3}{4} = \frac{1}{2} = \boxed{50\%}$$

(C) The probability we are looking for is

$$P(A \cup F \mid \bar{S}) = \frac{P\{(A \cup F) \cap \bar{S}\}}{P(\bar{S})}$$

The denominator in this expression is the probability that a member of the club does not use the swimming pool. From Part B, we know the probability of the complementary event - that is, the probability that a member *does* use the swimming pool, $P(\bar{S})$ = 50/100; using this result, we clearly have

$$P(\bar{S}) = 1 - P(S) = 1 - \frac{50}{100} = \frac{50}{100} = \frac{1}{2}$$

To obtain the value of the numerator, the way to go is to split the event of interest into three mutually exclusive components, and to sum their probabilities; mathematically,

$$P\{(A \cup F) \cap \bar{S}\} = P(A \cap M \cap \bar{S}) + P(A \cap F \cap \bar{S}) + P(C \cap F \cap \bar{S})$$

$$\therefore P\{(A \cup F) \cap \bar{S}\} = P(A)P(M \mid A)P(\bar{S} \mid A \cap M) + P(A)P(F \mid A)P(\bar{S} \mid A \cap F) + P(C)P(F \mid C)P(\bar{S} \mid C \cap F)$$

$$\therefore P\{(A \cup F) \cap \bar{S}\} = \frac{4}{5} \times \frac{3}{4} \times \left(1 - \frac{1}{2}\right) + \frac{4}{5} \times \frac{1}{4} \times \left(1 - \frac{1}{4}\right) + \frac{1}{5} \times \frac{2}{5} \times \left(1 - \frac{3}{4}\right) = \frac{47}{100}$$

Finally,

$$P(A \cup F \mid \bar{S}) = \frac{47/100}{1/2} = \frac{94}{100} = \boxed{94\%}$$

Problem 5

What we have is a contingency table with two rows and two columns (i.e., a 2×2 table). In total, there are n = 219 observations. If there is independence between rows and columns, the expected numbers will be:

	Pro	*Contra*
California	85.6	39.4
Texas	64.4	29.6

For the upper-left measurement, for instance, we write

$$f_{11} = \frac{R_1 S_1}{n} = \frac{125 \times 150}{219} = 85.6$$

For the other three spaces, in turn, we have

$$f_{12} = \frac{R_1 S_2}{n} = \frac{125 \times 69}{219} = 39.4$$

$$f_{21} = \frac{94 \times 150}{219} = 64.4$$

$$f_{22} = \frac{94 \times 69}{219} = 29.6$$

The statistic we aim for is then

$$\chi^2 = \left(\frac{94 \times 38 - 56 \times 31}{219} \right)^2 \times \left(\frac{1}{85.6} + \frac{1}{39.4} + \frac{1}{64.4} + \frac{1}{29.6} \right) = 6.07$$

The significance probability of the test thus becomes

$$P\left(\chi^2 \geq 6.07\right) = 0.01375$$

In the TI-83/84, use χ^2`cdf(6.07, E99, 1)`; (number `E99` basically means infinity). The probability is lower than 0.05, so we conclude that there is appreciable regional dependence in the opinion poll.

→ Section II – Part B

Problem 1

(A) A Type II error is to mistakenly failing to reject a false null hypothesis. In this situation, it would happen if the proportion of marriages in XY that end in divorce is greater than 41.8%, but the sample proportion does not provide sufficient evidence that it is. The consequence of such an erroneous finding would be to surmise that marriages in XY are just as stable and lasting as elsewhere in the country, which may not be true.

(B) Before anything else, note that the standard deviation is $\sqrt{p(1-p)/n}$ = $\sqrt{0.418(1-0.418)/81}$ = 0.0548. This is a one-sided test, so the critical z-score is 1.645. Accordingly, the sample proportion we are looking for is

$$\frac{\hat{p} - 0.418}{0.0548} > 1.645 \rightarrow \hat{p} - 0.418 = 0.0901$$

$$\therefore \boxed{\hat{p} > 0.508}$$

(C) If the true proportion is 0.46, the sampling distribution of $\hat{p}$ is approximately normal with mean 0.46 and standard deviation $\sqrt{p(1-p)/n}$ = $\sqrt{0.46(1-0.54)/81}$ = 0.0554. Using a Z-table, we find that

$$P(\hat{p} > 0.508) = P\left(z > \frac{0.508 - 0.46}{0.0554} \right) = P(z > 0.866) = \boxed{19.3\%}$$

Typing `normalcdf(0,1,0.866,E99)` should yield the same result.

(D) The concept in question is the *power* of the test.

(E) Increasing the sample size with $z > 1.645$ still as the rejection region would cause the sampling distribution of $\hat{p}$ to have a smaller standard deviation. The minimum value of $\hat{p}$ for which we would reject H_0 would be lower, and the probability of rejecting H_0 would be greater.

Problems researched and solved by Lucas Monteiro Nogueira.
Edited by Lucas Monteiro Nogueira.