Format: Two test grades in the second marking. Four free response questions and twenty multiple-choice questions. The free response and multiple-choice portions are weighted equally. These are some main topics you should know:

**Unit 1 Probability**
- Binomial
- Geometric
- Discrete Probability Distribution-mean/sd
- Combinations
- P(A or B), P(A and B)
- Normal Distributions(z score)
- Invnorm problems
- Central Limit Theorem
- Mean and proportions problems for samples
- CLT

**Unit 2 Describing Data**
- Dot plot/ bar chart/scatter plot
- Stem Plot-regular and comparative
- Box Plot-skeletal and modified. 5 # summary, markers for outliers
- Histogram-frequency table and percentile
- Calculate mean and standard deviation.
  - know what happens to each when data is increased by 15%
  - know what happens to each when each data point is increased by 15
- 4 measures of center
- 4 measures of variability
- 2 measures of relative location
- Skewed right/skewed left. Examples. Mean vs median

**Unit 3 Correlation and Regression**
- Lin Reg, r, r², SE, residuals, good fit, transformations

**Unit 4 Experimental Design**
- Experiments vs Observations
- 5 sampling techniques
- Definition of SRS
- 3 types of bias
- Design an experiment
- Key concepts of experimental design

**Unit 5- Part 1- Confidence Intervals**
- Confidence Intervals for means (T score)
- Confidence Intervals for proportions
- Sample size

**Some Vocabulary**

**Terms**
- Skewed Right
- Skewed Left
- Bell shaped
- Symmetric
- Uniform
- Normal curve
- Mean/SD/Med
- Variability
- Range/IQR
- Variance
- Quartile
- Gaps/clusters
- Observational Study
- Experiment
- Treatment
- Response Var
- Census
- Voluntary resp.
- Ramdon sample
- Control group
- Placebo
- Stratified
- Bias
- Randomization
- Blinding
- Double blind
- Confounding
Multiple Choice

1. The following are prices for a 25 inch T.V. found in different stores around Roseville:
   100, 98, 121, 111, 97, 135, 136, 104, 135, 138, 189, 114, 92, 69
   a. Skewed to the right
   b. Symmetric
   c. Skewed to the left
   d. Uniform
   e. Bell shaped

2. Which of the following are true statements?
   I. Not every symmetric bell shaped curves are normal
   II. All normal curves are bell shaped and symmetric
   III. All symmetric histograms are unimodal
   a. I only  b. II only  c. III only  d. I and II  e. None of the above

3. Which of the following are true statements?
   I. Stemplots are useful for extremely large data sets.
   II. In histograms, relative areas correspond to relative frequencies.
   III. Both dotplots and stemplots show symmetry, clusters, gaps, and outliers.
   a. I only  b. II only  c. I and II  d. II and III  e. III only

4. Which of the following distributions are more likely to be skewed to the left than skewed to the right?
   I. Scores on an easy test
   II. Scores on a hard test
   III. # of sodas drank in a week
   a. I only  b. I and II  c. I and III  d. II and III  e. I, II, and III

5. Suppose the average score on a national test is 600, with a standard deviation of 50.
   If each score is increased by 10, what are the new mean and standard deviation.
   a. 600, 60  b. 610, 50  c. 610, 55  d. 610, 60  e. 600, 50

6. When a set of data has suspect outliers, which of the following are preferred measures of central tendency and of variability?
   a. mean and variance
   b. median and range
   c. mean and range
   d. median and interquartile range
   e. mean and standard deviation
7. Use the following boxplot to answer the following

I. The mean score is 25
II. The interquartile range is 8
III. The 75th percentile rank is 19

a. I only  b. II only  c. III only  d. II and III  e. I and II

8. A student poll here on campus shows 20% like McDonalds, 30% like Burger King, 25% like Wendy’s and 25% like Smash Burger.
   Which of the following visual displays is most appropriate.
   a. Boxplot
   b. Dotplot
   c. Scatterplot
   d. Bar chart
   e. Stem and leaf plot

9. Consider the following two histograms, then decide which of the statements are true.

a. Both sets have the same standard deviation
b. Both sets have gaps and clusters
c. Only one of the graphs is symmetric
d. Both sets have the same mean and same range
e. Both sets have the same variance
10. Consider the following back to back stem and leaf plot

```
<table>
<thead>
<tr>
<th>4</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 5 2</td>
<td>5 1 1 4 8 8</td>
</tr>
<tr>
<td>4 1</td>
<td>6 1 4 5 6 7 8 9 9</td>
</tr>
<tr>
<td>8 7 5 4</td>
<td>7 3 4 6 6 8</td>
</tr>
<tr>
<td>9 6 5 3 3 0</td>
<td>8 6</td>
</tr>
<tr>
<td>8 5 4 3 3 1</td>
<td>9 2 4</td>
</tr>
</tbody>
</table>
```

Which of the following is a true statement

I. The ranges are the same
II. The variances are the same
III. The means are the same
IV. The medians are the same
V. The sample sizes are the same

a. I and II  b. I and V  c. II and V  d. III and V  e. I, II, and III

11. Consider the following parallel boxplots.

```
90 ------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|              |              | 100             | 110             | 120             | 130             | 140             |
|              | 110           | 120             | 130             | 140             |

I. They both have the same range
II. They both have the same interquartile ranges
III. They both have the same mean
IV. They both have the same medians

a. I and IV  b. II and IV  c. III and IV  d. II and III  e. I and II

12. Which of the following statements about the correlation coefficient r are true?

I. It is not affected by changes in the measurement units of the variables.
II. It is not affected by which variable is called x and which is called y.
III. It is not affected by extreme values.

a. I only  b. I and II  c. I and III  d. II and III  e. I, II, and III
13. Which of the following statements about residuals are true?

   I. Outliers in the y direction have large residuals.
   II. A definite pattern in the residual plot is an indication that a nonlinear model will show a better fit to the data than a straight regression line.
   III. A residual plot that has no definite pattern indicates that a nonlinear relationship will show a better fit to the data than a straight line.

   a. II only
   b. I and III
   c. II and III
   d. I, II, and III
   e. I and II

14. The heart disease death rates per 100,000 people in the United States for certain years were

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Death rate</td>
<td>307.6</td>
<td>286.2</td>
<td>253.6</td>
<td>217.8</td>
<td>202.0</td>
</tr>
</tbody>
</table>

Find the regression line and predict the death rate for the year 1983.

   a. 145.8 per 100,000 people
   b. 192.5 per 100,000 people
   c. 196.8 per 100,000 people
   d. 198.5 per 100,000 people
   e. None of the above

Questions 15 – 17 are based on the following table which shows the number of high school students taking an AP class in a subject by grade level

<table>
<thead>
<tr>
<th></th>
<th>Math</th>
<th>English</th>
<th>Foreign Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Senior</td>
<td>70</td>
<td>80</td>
<td>70</td>
</tr>
<tr>
<td>Junior</td>
<td>150</td>
<td>60</td>
<td>35</td>
</tr>
<tr>
<td>Sophomore</td>
<td>180</td>
<td>30</td>
<td>15</td>
</tr>
</tbody>
</table>

15. What Percentage of students are taking Math?

   a. 10.14%
   b. 30.43%
   c. 57.97%
   d. 88.4%
   e. None of the above

16. What percentage of the students are juniors and taking a Foreign Language?

   a. 14.29%
   b. 17.39%
   c. 5.07%
   d. 52.89%
   e. None of the above

17. What percentage of those taking English are Seniors?

   a. .4706
   b. .3636
   c. .1159
   d. .3043
   e. .3188
18. Which of the following are true statements?
   I. In an experiment some treatment is intentionally forced on one group to note the response.
   II. In an observational study information is gathered on an already existing situation.
   III. Sample surveys are observational studies, not experiments.
   a. I and II
   b. I and III
   c. II and III
   d. I, II, and III
   e. None of these are true statements.

19. In one study bodybuilders were given 500 and 1,000 milligrams of Creatine, and the increase in strength gained in the bench press was noted. In a second study people responded to a questionnaire asking about the average time they spend in the gym in a day and the amount of weight they can bench press.
   a. Both studies were controlled experiments
   b. Both studies were observational.
   c. None of the above is a correct statement.
   d. The first study was an experiment without a control group, while the second was an observational study.
   e. The first study was an observational study, while the second was a controlled experiment.

20. Which of the following are true statements?
   I. Based on careful use of control groups, experiments can often indicate cause and effect relationships.
   II. An observational study may suggest cause and effect if one can control lurking variables.
   III. A complete census is the only way to establish a cause and effect relationship absolutely.
   a. I, II, and III
   b. I and II
   c. I and III
   d. II and III
   e. None of the above.

21. Jerry Springer’s special show on would you date your own sister asked his viewers “would you date your own sister?” Of more than 10,000 viewers who responded, 70% said yes. What does this show?
   a. The survey would have been more meaningful if he had picked a random sample of the 10,000 viewers.
   b. The survey would have been more meaningful if he had used a control group.
   c. This was a legitimate sample, randomly drawn from his viewers and of sufficient size to allow the conclusion that most of his viewers would in fact date their own sister.
   d. No meaningful conclusion is possible without knowing something more about the characteristics of his viewers.
   e. The survey is meaningless because of voluntary response.

22. Each of the 7 basketball teams in the SFL has 12 players. A sample of 14 players is to be chosen as follows. Each team will be asked to place 12 cards with their players’ names into a hat and randomly draw out two names. The two names from each team will be combined to make up the sample. Will this method result in a simple random sample of the 84 basketball players.
   a. Yes, b/C this is an example of stratified sampling, which is a special case of simple random Sampling.
   b. No, because the teams are not chosen randomly.
   c. No, because not each group of 14 players has the same chance of being selected.
   d. Yes, because each player has the same chance of being selected.
   e. Yes, because each team is equally represented.
23. A researcher planning a survey of heads of households in a particular state has census lists for each of the 23 counties in that state. The procedure will be to obtain a random sample of heads of households from each of the counties rather than grouping all the census lists together and obtaining a sample from the entire group. Which of the following is a true statement about the resulting stratified sample?
   I. It is easier and less costly to obtain than a simple random sample.
   II. It gives comparative information that a simple random sample wouldn’t give.
   III. It is not a simple random sample.
   a. I and III
   b. I, II, and III
   c. I only
   d. I and II
   e. None of the above

24. In designing an experiment, blocking is used
   a. As a substitute for a control group
   b. As a first step in randomization
   c. To control the level of the experiment.
   d. To reduce bias
   e. To reduce variation by controlling extraneous factors.

25. Consider the following studies being run by three different nursing homes.
   I. One nursing home brings in pets for an hour every day to see if patient morale is improved.
   II. One nursing home allows hourly visits every day by kindergarten children to see if patient morale is improved.
   III. One nursing home administers antidepressants to all patients to see if patient morale is improved.

Which of the following statements are true?
   a. All of the (b-e) are true
   b. None of these studies uses randomization
   c. None of these studies uses control groups
   d. None of these studies uses blinding
   e. Important information can be obtained from all these studies, but none will be able to establish cause and affect relationships.

26. Which of the following are important in the design of an experiment?
   I. Control of confounding variables
   II. Randomization in assigning subjects to different treatments.
   III. Using a single treatment to gain knowledge about the response variable
   a. I only
   b. II only
   c. III only
   d. I, II, and III
   e. I and II
27. 30% of the students on campus bring their text book to class with them. In a random group of 10 people, what is the probability that exactly six have brought their books to class?
   a. .0367
   b. .1029
   c. .2000
   d. .3503
   e. .9890

28. One of the lottery tickets that you can purchase at a local store has these payoff probabilities.
   Payoff ($)       0       200       500
   Probability      .5       .2       .3

   What are the mean and standard deviation for the payoff variable?
   a. \( \mu = 233.33 \), \( \sigma = 355.9 \)
   b. \( \mu = 233.33 \), \( \sigma = 220.85 \)
   c. \( \mu = 190 \), \( \sigma = 363.13 \)
   d. \( \mu = 190 \), \( \sigma = 216.56 \)
   e. None of the above

29. If \( P(A) = .3 \) and \( P(B) = .2 \), what is \( P(A \cup B) \) if \( A \) and \( B \) are Independent?
   a. .06
   b. .44
   c. .50
   d. .56
   e. There is insufficient information.

30. There are two games involving flipping a coin. In the first game you win a prize if you can throw between 40% and 60% heads. In the second game you win if you can throw more than 75% heads. For each game would you rather flip the coin 50 or 500 times?
   a. It does not matter.
   b. 50 times for each game
   c. 500 for each game
   d. 50 for the first game and 500 for the second
   e. 500 for the first game and 50 for the second

31. Given that 55% of the U.S. Population are female and 20% are older than age 65, can we conclude that \((.55)(.20) = 11\% \) are women older than 65?
   a. Yes, by the multiplication rule.
   b. Yes, by conditional probability
   c. Yes, by law of large numbers.
   d. No, because the events are not independent.
   e. No, because the events are mutually exclusive.
32. Consider the following table of ages of U.S. senators

<table>
<thead>
<tr>
<th>Age (yrs.)</th>
<th>&lt; 40</th>
<th>40 – 49</th>
<th>50 – 59</th>
<th>60 – 69</th>
<th>70 – 79</th>
<th>&gt; 79</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of senators</td>
<td>10</td>
<td>25</td>
<td>30</td>
<td>29</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

What is the probability that a senator is less than 60 years old given that he is over 39 years old?

a. .055
b. .111
c. .611
d. .90
e. .94

33. Given \( P(A) = .3 \) and \( P(A \cup B) = .7 \), what is the probability \( P(B) \) if \( A \) and \( B \) are mutually exclusive? If \( A \) and \( B \) are independent?

a. .4, .3
b. .4, .57
c. None of the answers are correct
d. .7, .57
e. .7, .3

34. Granite Bay High school receives 15% of its packages from company C1, 45% from company C2, and the rest from company C3. The probability that a package is late is .03, .02, and .04, depending on whether it is from company C1, C2, or C3, respectively. If a randomly chosen package is late, what is the probability that it is from company C1?

a. .153
b. .0045
c. .0295
d. .016
e. .009

35. Which of the following are true statements?

f. None of the above.

I. Normal curves with different means can be centered around the same numbers.

II. The area under a normal curve is always equal to one, no matter what the mean and standard deviation are.

III. The smaller the standard deviation of a normal curve, the shorter and wider the graph.

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

36. A trucking firm determines that its fleet of trucks averages a mean of 18.2 miles per gallon with a standard deviation of .8 miles per gallon on cross country hauls. What is the probability that one of the trucks averages more than 19 miles per gallon?

a. .212
b. .788
c. None of the answers is correct
d. .159
e. .841
37. The mean score on a college entrance exam is 600 with a standard deviation of 50. 80% of the test takers score above what value?
   a. 642
   b. 558
   c. 525
   d. 650
   e. None of the answers are correct

38. Suppose that 25% of all business executives are willing to switch companies if offered a higher salary. If a headhunter randomly contacts a random sample of 75 executives, what is the probability that over 30% will be willing to switch companies if offered a higher salary?
   a. .146
   b. .182
   c. .436
   d. .841
   e. .159

39. Which of the following are true statements?
   I. The sampling distribution of \( \hat{p} \) has a mean equal to the population proportion \( p \).
   II. The sampling distribution of \( \hat{p} \) has a standard deviation equal to \( \sqrt{np(1-p)} \).
   III. The sampling distribution of \( \hat{p} \) is considered close to normal provided that \( n \geq 30 \).
   a. None of the below.
   b. I and II
   c. I and III
   d. II and III
   e. I, II, and III

40. Given that 58% of all gold dealers believe next year will be a good one to speculate in South African gold coins, in a simple random sample of 150 dealers, what is the probability that between 55% and 60% believe that it will be a good year to speculate?
   a. .4619
   b. .0500
   c. .1192
   d. .3099
   e. .9215

41. Assume that a baseball team has an average pitcher. That is one whose probability of winning any decision is .5. If this pitcher has 40 decisions in a season, what is the probability that he will win less than 30 games?
   a. .9989
   b. .9997
   c. .002
   d. .001
   e. .0003
42. The mean income per household in a certain state is $9000 with a standard deviation of $1500. The middle 90% of incomes are between what two values?

a. $6060 and $11940  
   b. $6532 and $11467  
   c. $7078 and $8999  
   d. $7737.5 and $10262.4

Mixed Problem Set

This histogram shows the ages of the last 150 people who rode The Beast at Kings Island. Use it to answer questions #1-3.
1. The distribution of the graph can best be described as which of the following: symmetric, skewed to the right, or skewed to the left?
2. Which of the following statements is correct about this histogram:
   a. The mean will most likely be less than the median.  
   b. The mean will most likely be more than the median.  
   c. The mean will be exactly equal to the median.  
3. In which range will the median of this data lie? (i.e. 5-10, 10-15, 15-20, etc.)

The following data is a list of the ages of the last 30 people to ride the carousel at Kings Island. Use it to answer questions #4-8.

35 71 24 54 55 68 4 29 31 6 10 73 45 48 52 27 3 43 52 81 78 36 39 11 8 63 60 29 35 9

4. Organize the data into a stemplot.
5. Find the mean, median, standard deviation, and five-number summary of the data.
6. What is the interquartile range?
7. Does the data have any outliers? How can you tell?
8. Organize the data into a boxplot.
The following shows the ages of the last 1000 people to ride the FireHawk at Kings Island. Use it to answer questions #9-13.

9. If a 32-year-old rides FireHawk, in what percentile would he be?
10. What age corresponds to the 40th-percentile?
11. What is the IQR for this set of data, approximately?
12. Should a FireHawk rider that is 55 be considered an outlier in this set of data? Why or why not?
13. If you were to draw a histogram of this data, would it symmetrical or skewed? How can you tell?

14. Suppose that the mean of a set of data is 55.8 and the standard deviation of a set of data is 12.2.
   a. What would the new mean and standard deviation be if you added 10 to each data point?
   b. What would the new mean and standard deviation be if you multiplied each data point by 5?
   c. What is the variance of this set of data?
15. Describe the two main differences between a bar graph and a histogram.

The lengths of time (in hours) an incandescent light bulb can stay lit solidly are normally distributed, with a mean of 62 hours and a standard deviation of 4.4 hours. Use this information to answer questions #16-20.

16. What percent of light bulbs can last within one standard deviation of the mean, between 57.6 hours and 66.4 hours?
17. A company considers a bulb defective if it can only last 50 hours straight. What portion of light bulbs would this company consider defective?
18. An energy-efficient bulb can stay lit for at least 68 hours. What percent of incandescent bulbs can stay lit for that long?
19. What portion of light bulbs can stay lit for between 55 and 65 hours straight?
20. The top 2% of light bulbs can stay lit for at least how long?

21. Sophia got a 95% on her Statistics mid-term and a 91% on her Calculus mid-term. The grades on both tests were normally distributed. The Statistics grades had a mean of 87%, with a standard deviation of 7%, while the Calculus grades had a mean of 85% with a standard deviation of 4%. On which test did Sophia do better, compared to the rest of her class? How can you tell?
22. A set of data has the following normal probability plot. Is the data normal? How can you tell?
23. A bad statistician heard that the mean age of the riders of Invertigo at Kings Island is 23, with a standard deviation of 5.5 years. He concluded that 95% of the riders of Invertigo must be between the ages of 12 and 34. What is incorrect about his conclusion?
The table below compares the average weight and average life span of several common dog breeds. Use it to answer questions #24-29.

<table>
<thead>
<tr>
<th>Breed</th>
<th>Avg. weight (lbs)</th>
<th>Avg. life span (yrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beagle</td>
<td>26</td>
<td>13.0</td>
</tr>
<tr>
<td>Boxer</td>
<td>70</td>
<td>12.5</td>
</tr>
<tr>
<td>Bulldog</td>
<td>50</td>
<td>11.1</td>
</tr>
<tr>
<td>Chihuahua</td>
<td>4</td>
<td>14.3</td>
</tr>
<tr>
<td>Dachshund</td>
<td>19</td>
<td>13.2</td>
</tr>
<tr>
<td>German shepherd</td>
<td>82</td>
<td>12.3</td>
</tr>
<tr>
<td>Golden retriever</td>
<td>70</td>
<td>11.0</td>
</tr>
<tr>
<td>Labrador retriever</td>
<td>73</td>
<td>12.0</td>
</tr>
<tr>
<td>Pomeranian</td>
<td>5</td>
<td>13.9</td>
</tr>
<tr>
<td>Poodle</td>
<td>42</td>
<td>12.7</td>
</tr>
<tr>
<td>Rottweiler</td>
<td>113</td>
<td>10.8</td>
</tr>
<tr>
<td>Yorkshire terrier</td>
<td>6</td>
<td>13.5</td>
</tr>
</tbody>
</table>

24. Calculate \( r \) for this set of data. What can you conclude about the relationship between average weight and average life span of a dog breed?

25. Calculate the equation of the least-squares regression line for this set of data. How accurate will the regression line be at predicting the average life span of a dog breed from its weight? How do you know?

26. A Shih Tzu weighs an average of 13 pounds. Approximately how long is a Shih Tzu’s average life span?

27. A Shih Tzu’s actual average life span is 12.5 years. What is the residual of that data point?

28. A Mastiff weighs an average of 200 pounds. Approximately how long is a Mastiff’s average life span? Are you confident in your answer as the true estimate of a Mastiff’s average life span? Why or why not?

29. What percent of a dog breed’s average life span can be explained by its average weight?

30. A researcher runs a detailed study and concludes the following: “The correlation between the age when a child first walks and the age when a child says their first word appears to be approximately zero.” Describe what this means to someone who does not know anything about statistics.

31. A study was conducted to see if a baby’s birth weight was related to their birth length. A sample of 200 babies were measured and the following data was gathered:

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth weights:</td>
<td>8.7 lbs</td>
<td>2.8 lbs</td>
</tr>
<tr>
<td>Birth lengths:</td>
<td>15.3 in</td>
<td>4.2 in</td>
</tr>
</tbody>
</table>

Correlation coefficient: \( r = 0.895 \)

A scatterplot revealed that the data was fairly linear. Use all of that information to write the regression line that predicts birth length from birth weight.

32. In the scatterplot to the right, what would happen to the value of \( r \) if point X was removed? What would happen to the slope of the regression line? Based on that, is point X an outlier? Why or why not?
33. Match each of the following functions to the transformation that would make them linear:

i. x vs. y

a. \( y = 3(2)^5 \)

b. \( y = 3x^2 \)

d. Transform that relationship back into a direct relationship between x and y. What is the value of y when x=0.2?

34. A set of data is shown to have a linear relationship if transformed into the function \( \log(y) = 2 + 5x \).

35. Label each relationship below with its correct type: cause-and-effect, common response, or confounding variables.

a. There is a strong, negative correlation between one’s height and the length of one’s hair. However, it should be noted that men are generally taller and have shorter hair.

b. There is a strong, positive correlation between one’s height and one’s age. However, it should be noted that one grows taller as they age.

c. There is a strong, positive correlation between one’s height and the distance one can run before getting winded. However, it should be noted that there are other factors besides how long one’s legs are that contribute to how far you can run, including health and level of training.

36. Studies have indicated that there is a strong, negative correlation between the number of times per week that you brush your teeth and the number of cavities you get each year. Which of the following can be concluded from this information?

a. Not brushing your teeth causes you to get cavities.

b. People who don’t brush their teeth often are more likely to have cavities than people who do.

c. There is statistically significant evidence that brushing your teeth prevents cavities.

d. None of these can be concluded.

37. A politician wants to know how the residents of his district will react to a bill that lowers the driving age to 15 years old. He runs an ad during the evening news on a local television station that says:

Let us know what you think! Would you be in favor of allowing 15-year-old children to get their driver’s license, or would you rather keep the driving age at 16-years-old, when they are more mature and ready to take on the responsibility of driving? Give us a call at 555-7834, and give us your opinion!

Which of the following types of bias are present in this ad: voluntary response bias, convenience sampling, undercoverage bias, non-response bias, poor wording effect bias?

38. When asked what their favorite sport is, 44% of American men say football, 26% say baseball, 22% say basketball, 5% say soccer, 2% say hockey, and 1% say something else. Use the following sequence of random numbers to simulate asking 30 men about their favorite sport. Clearly explain the process you used in your simulation.

14459 26056 31429 80371 65103 62253 50490 61181 38967 98532 62183 70632 23417 26185
39. A report on a new brand of headache medicine, Probanol, is published that says, “After extensive research, there is statistically significant evidence that Probanol reduces the likelihood of getting a migraine headache.” Explain what that means to someone who doesn’t know anything about statistics.

An eye doctor believes that he has invented a new drop that improves eyesight. He randomly selects 200 people, and he has them read an eye chart. He then administers the drops and has them read a similar eye chart, noting any improvement. However, despite telling everyone that they will be receiving the new drops, he only gives 100 of the subjects the actual drops; the other half of the sample is simply given water drops.

40. What is the purpose of the water drops? Why couldn’t the eye doctor simply have administered his drops and noted improvement?
41. What is the factor in this experiment? What are the treatments?
42. Is this experiment blind? Is it double-blind?

The principal wants to know if the JJHS student body would like the library to stay open longer during the day. He gathers a random sample of 100 students from each grade level.

43. What is the population of this study?
44. Which of these most accurately describes this sample: simple random sample, stratified random sample, or census?

45. The probability that Event A occurs is 0.48, and the probability that Event B occurs is 0.62. The probability that both occur at the same time is 0.23. Draw a Venn diagram to represent this situation. Then use it to find P(A and B), P(A′ and B), P(A and B′), P(A′ and B′), P(A|B), and P(B|A).
46. Event A is that you will complete this exam review. Event B is that you will get an A on the semester exam. Are Events A and B disjoint? Are they independent?

You have a large bag of marbles, with proportions of each color listed below. Use that chart to answer questions #47-54.

<table>
<thead>
<tr>
<th>Color</th>
<th>Red</th>
<th>Yellow</th>
<th>Blue</th>
<th>Green</th>
<th>Orange</th>
<th>Purple</th>
<th>Black</th>
<th>White</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prob.</td>
<td>0.13</td>
<td>0.06</td>
<td>0.21</td>
<td>0.08</td>
<td>0.18</td>
<td>0.05</td>
<td>0.11</td>
<td>??</td>
</tr>
</tbody>
</table>

47. What is the probability of drawing a white marble from the bag?
48. What is the probability of drawing either a red or blue marble from the bag?
49. What is the probability of drawing a marble from the bag that is not yellow?
50. I draw a marble from the bag and tell you that it isn’t black or white. What is the probability that it is green?
51. Imagine that you draw two marbles from the bag, replacing the first before drawing the second. What is the probability that they are both orange?
52. What is the probability that neither of them is orange?
53. Are questions #51 and #52 above complements? Why or why not?
54. Imagine that you draw five marbles from the bag, replacing them each time before drawing the next. What is the probability that you get at least one purple?
Suppose that, in Mr. Willets’ AP Calculus class, 71% of the students are seniors, 26% are juniors, and 4% are sophomores. 36% of the seniors and 15% of the juniors are also in AP Statistics, but none of the sophomores are. Use that information to answer questions #55-60.

55. Draw a tree diagram to organize the above information. Make sure to include the final probability of each branch.

56. What is the probability that a randomly-selected student from Mr. Willets’ class is a junior in AP Statistics?

57. What is the probability that a randomly-selected student from Mr. Willets’ class is a senior who is not in AP Statistics?

58. What is the probability that a randomly-selected student from Mr. Willets’ class is a sophomore who is in AP Statistics?

59. What is the probability that a randomly-selected student from Mr. Willets’ class is also in AP Statistics?

60. A student is randomly selected from Mr. Willets’ class, and that student is also in AP Statistics. What is the probability that they are a senior?

61. Which of the following sequences of heads/tails is most likely to occur, if you flipped a fair coin 6 times?
   b. H, T, H, T, H, T
   c. H, H, H, T, T, T
   d. T, H, H, T, H, T
   e. None of these

62. A fair die is rolled 600 times. Label each of the following statements as true or false:
   a. Exactly 100 of the rolls will be 1s.
   b. As the number of rolls approaches 600, the proportion of 5s rolled will get closer to 1/6.
   c. A run of 8 odd numbers in a row is impossible, since the proportion of evens and odds has to stay close to 50%.
   d. The number of 3s rolled should be approximately equal to the number of 6s rolled, by the end.
   e. The first twelve rolls will include two of each number, to keep the proportions equal.

In the following probability distribution, X = the number of 10s rolled on three ten-sided dice. Use it to answer questions #63-65.

<table>
<thead>
<tr>
<th>X</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(X)</td>
<td>.729</td>
<td>.243</td>
<td>.027</td>
<td>.001</td>
</tr>
</tbody>
</table>

63. Is the distribution above discrete or continuous?

64. Find the mean and standard deviation of X.

65. If you multiplied each X-value above by 3 and added 5, what would be the new mean and standard deviation of X?

66. A man on the street offers you a wager. He’ll fan out a deck of cards and let you pick one at random. If it’s a face card, he’ll give you $2. If it’s an ace, he’ll give you $4. If it’s anything else, you give him $1. What is the amount you are expected to win each time you play? Should you take his offer?
67. For each of the following situations, say whether or not a binomial distribution could be used:
   a. You draw three cards out of deck, without replacing them. Let X be the number of red cards drawn.
   b. You survey a SRS of 100 college students. Let X be the number of them that are over 21.
   c. You roll a die twenty times. Let X be the number of times you roll a prime number.
   d. You spin a spinner ten times. Let X be the total sum of the numbers you roll.

You will roll a twenty-sided die ten times in a row, and you will record the number of times you roll a perfect square. Use this situation to answer questions #68-72.

68. What is the probability that you roll a perfect square exactly 4 times?
69. What is the probability that you roll a perfect square no more than 2 times?
70. What is the probability that you roll a perfect square at least 6 times?
71. How many times should you expect to roll a perfect square?
72. Can you use a normal curve to approximate this distribution? Why or why not?

You want to know how many women in Ohio are democrats. You will interview 200 women and ask them their political party. Imagine that, in reality, 58% of Ohio women are democrats. Use this information to answer questions #73-75.

73. Strictly speaking, this situation isn’t binomial, because the 200 trials aren’t technically independent. Why can we still use a binomial distribution to calculate probabilities here?
74. What is the mean and standard deviation of the number of democrats in your sample?
75. Use a normal curve to approximate the probability that at least 130 women in your sample are democrats.
Answers to Mixed Problem Set:

1. Skewed to the right
2. b
3. 20-25
4. Stemplot
5. $\mu = 39.3; M = 37.5; \sigma = 23.52; \text{five-number summary} = (3\ 24\ 37.5\ 55\ 81)$
6. $\text{IQR} = 31$
7. No, because $1.5\times\text{IQR}$ yields a range from -22.5 to 101.5, and none of the data is outside of that range.
8. Boxplot
9. 90th-percentile
10. 18 years old
11. 10
12. Yes, because $1.5\times\text{IQR}$ is 15, and 55 is more than $15+Q_3=40$.
13. Skewed to the right, because most of the data is on the left side of the age axis.
14. a. $\mu = 65.8; \sigma = 12.2$
    b. $\mu = 279; \sigma = 61$
    c. 148.84
15. Bar graphs describe categorical data, while histograms describe quantitative data. In a bar graph, the bars do not touch, while in a histogram, the bars are connected.
16. 68%
17. 0.0032
18. 0.0869
19. 0.6958
20. 71.02 hours
21. Calculus exam; her z-score for Statistics (1.14) was lower than her z-score for Calculus (1.5)
22. No, because the normal probability plot isn’t straight.
23. The data probably is not normal, as it is most likely skewed toward younger riders.
24. $r = -0.849; \text{there is a strong negative relationship between a dog breed’s average weight and average life span.}$
25. $y = 13.8-0.027x; \text{it will be pretty accurate, since the correlation is so strong.}$
26. 13.5 years
27. -1
28. 8.4 years; not very confident, since the Mastiff is far outside of our range of data.
29. 72.2%
30. There is no relationship between the age when a child first walks and the age when a child first says their first word. You cannot predict one based off of the other.
31. $y = 3.616+1.343x$
32. $r$ would most likely decrease, as the scatter would increase. The slope of the least-squares regression line would be unaffected, as the line would follow a similar pattern. Point X is an outlier, as it affects the value of $r$.
33. a. ii
    b. iv
34. $y = 100(10^5); y = 1,000$
35. a. common response (girls tend to be shorter, and they also tend to have longer hair)
    b. cause-and-effect (getting older causes you to grow taller)
    c. confounding variables (there are other factors that contribute to how far you can run)
36. b
37. voluntary response bias, undercoverage bias, poor wording bias
38. I broke the numbers 1-100 into the following categories: 01-44 = football, 45-70 = baseball, 71-92 = basketball, 93-97 = soccer, 98-99 = hockey, 00 = other. Then I created two-digit numbers from the random number sequence and came up with the following results: 15 said football, 8 said baseball, 5 said basketball, 1 said soccer, and 1 said hockey
39. There is enough evidence, gathered through many experiments, to conclude that the relationship between taking Probanol and the reduction of migraine is not happening by random coincidence. Thus, it can be concluded that Probanol causes migraine reduction.

40. He needed a placebo, to counteract the placebo effect.

41. The factor is the eye drops, while the treatments are the actual drops versus the placebo group.

42. It is blind, as the subjects do not know which group they are in, but it is not double-blind, as the eye doctor himself does know.

43. The population is the JJHS student body.

44. Stratified random sample

45. Venn diagram; \( P(A \text{ and } B) = 0.23, P(A^c \text{ and } B) = 0.39, P(A \text{ and } B^c) = 0.25, P(A^c \text{ and } B^c) = 0.13, P(A | B) = 0.371, P(B | A) = 0.479 \)

46. They are neither disjoint nor independent.

47. 0.18

48. 0.34

49. 0.94

50. 0.113

51. 0.0324

52. 0.6724

53. No, because they don’t add up to 1. The complement of “both orange“ is “not both orange.”

54. 0.226

55. Tree diagram

56. 0.039

57. 0.454

58. 0

59. 0.295

60. 0.8676

61. e (They are all equally likely!)

62. a. false
   b. true
   c. false
   d. true
   e. false

63. discrete

64. \( \mu = 0.3; \sigma = 0.520 \)

65. \( \mu = 5.9; \sigma = 1.559 \)

66. $0.08; yes, you should take his bet, because you should expect to win 8 cents every bet, in the long run

67. a. no (not independent)
   b. yes (technically not, but the population is significantly large enough)
   c. yes
   d. no (more than two possible outcomes)

68. 0.088

69. 0.678

70. 0.0063

71. 2

72. No, because \( np \) is less than 10.

73. The population of Ohio women is so large that taking a sample of 200 out of it will not affect individual probabilities significantly.

74. \( \mu = 116; \sigma = 6.98 \)

75. \( P(X > 130) = P(z > 2.01) = 1 - .9778 = .0222 \)
Free Response B (from old AP Exams)

2013 Exam

1. An environmental group conducted a study to determine whether crows in a certain region were ingesting food containing unhealthy levels of lead. A biologist classified lead levels greater than 6.0 parts per million (ppm) as unhealthy. The lead levels of a random sample of 23 crows in the region were measured and recorded. The data are shown in the stemplot below.

   Lead Levels
   
   2 \| 8
   3 \| 0
   3 \| 5 8 8
   4 \| 1 1 2
   4 \| 6 8 8
   5 \| 0 1 2 2 3 4
   5 \| 9 9
   6 \| 3 4
   6 \| 6 8

   Key: \ 2|8 = 2.8 \ ppm

   (a) What proportion of crows in the sample had lead levels that are classified by the biologist as unhealthy?

   (b) The mean lead level of the 23 crows in the sample was 4.90 ppm and the standard deviation was 1.12 ppm. Construct and interpret a 95 percent confidence interval for the mean lead level of crows in the region.

2. An administrator at a large university wants to conduct a survey to estimate the proportion of students who are satisfied with the appearance of the university buildings and grounds. The administrator is considering three methods of obtaining a sample of 500 students from the 70,000 students at the university.

   (a) Because of financial constraints, the first method the administrator is considering consists of taking a convenience sample to keep the expenses low. A very large number of students will attend the first football game of the season, and the first 500 students who enter the football stadium could be used as a sample. Why might such a sampling method be biased in producing an estimate of the proportion of students who are satisfied with the appearance of the buildings and grounds?

   (b) Because of the large number of students at the university, the second method the administrator is considering consists of using a computer with a random number generator to select a simple random sample of 500 students from a list of 70,000 student names. Describe how to implement such a method.

   (c) Because stratification can often provide a more precise estimate than a simple random sample, the third method the administrator is considering consists of selecting a stratified random sample of 500 students. The university has two campuses with male and female students at each campus. Under what circumstance(s) would stratification by campus provide a more precise estimate of the proportion of students who are satisfied with the appearance of the university buildings and grounds than stratification by gender?
3. Each full carton of Grade A eggs consists of 1 randomly selected empty cardboard container and 12 randomly selected eggs. The weights of such full cartons are approximately normally distributed with a mean of 840 grams and a standard deviation of 7.9 grams.

(a) What is the probability that a randomly selected full carton of Grade A eggs will weigh more than 850 grams?

(b) The weights of the empty cardboard containers have a mean of 20 grams and a standard deviation of 1.7 grams. It is reasonable to assume independence between the weights of the empty cardboard containers and the weights of the eggs. It is also reasonable to assume independence among the weights of the 12 eggs that are randomly selected for a full carton.

Let the random variable $X$ be the weight of a single randomly selected Grade A egg.

i) What is the mean of $X$?

ii) What is the standard deviation of $X$?
1. The scatterplot below displays the price in dollars and quality rating for 14 different sewing machines.

(a) Describe the nature of the association between price and quality rating for the sewing machines.
(b) One of the 14 sewing machines substantially affects the appropriateness of using a linear regression model to predict quality rating based on price. Report the approximate price and quality rating of that machine and explain your choice.
(c) Chris is interested in buying one of the 14 sewing machines. He will consider buying only those machines for which there is no other machine that has both higher quality and lower price. On the scatterplot reproduced below, circle all data points corresponding to machines that Chris will consider buying.
2. A charity fundraiser has a Spin the Pointer game that uses a spinner like the one illustrated in the figure below.

A donation of $2 is required to play the game. For each $2 donation, a player spins the pointer once and receives the amount of money indicated in the sector where the pointer lands on the wheel. The spinner has an equal probability of landing in each of the 10 sectors.

(a) Let $X$ represent the net contribution to the charity when one person plays the game once. Complete the table for the probability distribution of $X$.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$$2$</th>
<th>$$1$</th>
<th>$-$8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(x)$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) What is the expected value of the net contribution to the charity for one play of the game?

(c) The charity would like to receive a net contribution of $500 from this game. What is the fewest number of times the game must be played for the expected value of the net contribution to be at least $500$?

(d) Based on last year’s event, the charity anticipates that the Spin the Pointer game will be played 1,000 times. The charity would like to know the probability of obtaining a net contribution of at least $500 in 1,000 plays of the game. The mean and standard deviation of the net contribution to the charity in 1,000 plays of the game are $700$ and $92.79$, respectively. Use the normal distribution to approximate the probability that the charity would obtain a net contribution of at least $500 in 1,000 plays of the game.
Solutions

2013 Exam Question #1

**Part (a):**

Four of the 23 crows in the sample had a lead level greater than 6.0 ppm. Therefore, the proportion of crows in the sample that were classified as unhealthy is \( \frac{4}{23} \approx 0.174 \).

**Part (b):**

Step 1: Identifies the appropriate confidence interval (by name or by formula) and checks appropriate conditions.

The appropriate procedure is a one-sample \( t \)-interval for a population mean.

**Conditions:**
1. The sample is randomly selected from the population.
2. The population has a normal distribution, or the sample size is large.

The first condition is met because we were told that the crows were randomly selected. The sample size of 23 is not considered large, so we need to examine the sample data to assess whether it is reasonable to assume that the population distribution of lead levels for all crows in this region is normal. The stem-and-leaf plot shows no strong skewness or outliers, so we will consider the second condition to be met.

Step 2: Correct mechanics

A 95% confidence interval for the population mean \( \mu \) is given by: \( \bar{x} \pm t^* \frac{s}{\sqrt{n}} \). The critical value for 95% confidence, based on 23 - 1 = 22 degrees of freedom, is \( t^* = 2.074 \). The 95% confidence interval for \( \mu \) is therefore

\[
4.90 \pm 2.074 \times \frac{1.12}{\sqrt{23}} \approx 4.90 \pm 0.484,
\]

which is the interval (4.416, 5.384) ppm.

Using the raw data rather than the given summary statistics, the 95% confidence interval for \( \mu \) is (4.411, 5.3803).

Step 3: Interpretation

We can be 95% confident that the population mean lead level among all crows in this region is between 4.416 and 5.384 parts per million.
2013 Exam Question #2

**Part (a):**

The first 500 students who enter the football stadium were not likely to be representative of the population of all students at the university. In other words, these 500 students were likely to differ systematically from the population with regard to many variables. For example, these 500 students might have more school pride than the population of students as a whole, which might be related to their opinions about the appearance of university buildings and grounds. Perhaps their school pride is related to having more positive opinions about the appearance of university buildings and grounds, in which case the sample proportion of students who were satisfied would be biased toward overestimating the population proportion of students who were satisfied.

**Part (b):**

Obtain a list of all 70,000 students at the university. Assign an identification number from 1 to 70,000 to each student.

Then use a computer to generate 500 random integers between 1 and 70,000 without replacement. The students whose ID numbers correspond to those numbers were then selected for the sample.

**Part (c):**

Stratifying by campus would be more advantageous than stratifying by gender provided that opinions about appearance of university buildings and grounds between the two campuses differ more than the opinions about appearance of university buildings and grounds between the two genders.
Solution

Part (a):

Let $W$ denote the weight of a randomly selected full carton of eggs. $W$ has a normal distribution with mean 840 grams and standard deviation 7.9 grams.

The $z$-score for a weight of 850 grams is $z = \frac{850 - 840}{7.9} \approx 1.27$.

The standard normal probability table reveals that $P(W > 850) = P(Z > 1.27) \approx 1 - 0.8980 = 0.1020$.

Part (b):

(i) Let $W$ represent the weight of a randomly selected full carton of eggs, $P$ the weight of the packaging, and $X_i$ the weight of the $i$th egg, for $i = 1, 2, \ldots, 12$.

Note that $W = P + X_1 + X_2 + \ldots + X_{12}$.

Properties of expected values establish that $E(W) = E(P) + E(X_1) + \ldots + E(X_{12})$.

Because all 12 eggs have the same mean weight, this becomes $E(W) = E(P) + 12 \times E(X_i)$.

We were told that $E(W) = 840$ and $E(P) = 20$, so we can solve $840 = 20 + 12 \times E(X_i)$ to find $E(X_i) = \frac{840 - 20}{12} \approx 68.33$ grams.

(ii) Because of independence, properties of variance establish that $\text{Var}(W) = \text{Var}(P) + \text{Var}(X_1) + \text{Var}(X_2) + \ldots + \text{Var}(X_{12})$.

Because all 12 eggs have the same variance of their weights, this becomes $\text{Var}(W) = \text{Var}(P) + 12 \times \text{Var}(X_i)$.

We were told that $\text{SD}(W) = 7.9$ and $\text{SD}(P) = 1.7$. Therefore, $\text{Var}(W) = (7.9)^2 = 62.41$ and $\text{Var}(P) = (1.7)^2 = 2.89$.

We can solve $62.41 = 2.89 + 12 \times \text{Var}(X_i)$ to find $\text{Var}(X_i) = \frac{62.41 - 2.89}{12} = 4.96$. Thus, $\text{SD}(X_i) = \sqrt{4.96} \approx 2.23$ grams.
2012 Exam Question #1

**Part (a):**

The data show a weak but positive association between price and quality rating for these sewing machines. The form of the association does not appear to be linear. Among machines that cost less than $500, there appears to be very little association between price and quality rating. But the machines that cost more than $500 do generally have better quality ratings than those that cost less than $500, which causes the overall association to be positive.

**Part (b):**

The sewing machine that most affects the appropriateness of using a linear regression model is the one that costs about $2,200 and has a quality rating of about 65. Although the other four sewing machines costing more than $500 generally have higher quality ratings than those costing under $500, their prices and quality ratings follow a trend that suggests that quality ratings may not continue to increase with higher prices, but instead may approach a maximum possible quality rating. The $2,200 sewing machine is the most expensive of all but has a relatively low quality rating, which is consistent with a nonlinear model that approaches a maximum possible quality rating and then perhaps decreases. If a linear model were fit to all of the data, this one machine would substantially pull the regression line toward it, resulting in a poor overall fit of the line to the data.

**Part (c):**

According to Chris’s criterion, there are two sewing machine models that he will consider buying:

1. The model that costs a bit more than $100 and has a quality rating of 65.
2. The model that costs a bit below $500 and has a quality rating of 81 or 82.

The data points corresponding to these two machines have been circled on the scatterplot below.
2012 Exam Question #2

Part (a):

By counting the number of sectors for each value and dividing by 10, the probability distribution is calculated to be:

<table>
<thead>
<tr>
<th>x</th>
<th>$2</th>
<th>$1</th>
<th>$-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(x)</td>
<td>0.6</td>
<td>0.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Part (b):

The expected value of the net contribution for one play of the game is:

\[ E(x) = \$2(0.6) + \$1(0.3) + (-\$8)(0.1) = \$0.70 \] (or 70 cents).

Part (c):

The expected contribution after \( n \) plays is \$0.70n. Setting this to be at least \$500 and solving for \( n \) gives:

\[ 0.70n \geq 500, \text{ so } n \geq \frac{500}{0.70} \approx 714.286, \]

so 715 plays are needed for the expected contribution to be at least \$500.

Part (d):

The normal approximation is appropriate because the very large sample size (\( n = 1,000 \)) ensures that the central limit theorem holds. Therefore, the sample mean of the contributions from 1,000 plays has an approximately normal distribution, and so the sum of the contributions from 1,000 plays also has an approximately normal distribution.

The \( z \)-score is \[ \frac{500 - 700}{92.79} \approx -2.155. \]

The probability that a standard normal random variable exceeds this \( z \)-score of \(-2.155 \) is 0.9844. Therefore, the charity can be very confident about gaining a net contribution of at least \$500 from 1,000 plays of the game.