AP Review - Ch. 1

Chapter Summary: Exploring Data

In this chapter, we learned that statistics is the art and science of data. When working with data, it is important to know whether the variables are categorical or quantitative, as this will determine the most appropriate display for the distribution. For categorical data, the display will help us describe the distribution. For quantitative data, the display will help us describe the shape of the distribution and suggest the most appropriate numeric measures of center and variability. Always begin with a graph of the distribution, then move to a numerical description. When exploring quantitative data, we want to be sure to interpret the shape, outliers, center, and variability. Look for an overall pattern to describe your data and note any striking departures from that pattern.

As you study, be sure to focus on *understanding*, not just mechanics. While it may be easy to "plug the data" into your calculator, simply making graphs and calculating values is not the point of statistics. Rather, focus on being able to explain HOW a graph or value is constructed and WHY you would choose a certain display or numeric summary. Get in this habit early...your calculator is a powerful tool, but it cannot replace your thinking and communication skills!

After You Read: What Have I Learned?
Complete the vocabulary puzzle, multiple-choice questions, and FRAPPY. Check your answers and performance on each of the learning targets. Be sure to get extra help on any targets that you identify as needing more work!

Learning Target	Got It!	Almost There	Needs Work
I can identify individuals and variables for a set of data			
I can classify variables as categorical or quantitative			
I can make and interpret bar graphs for categorical data			
I can identify what makes some graphs of categorical data misleading			
I can calculate marginal and joint relative frequencies from a two-way table			
I can calculate conditional relative frequencies from a two–way table			
I can use bar graphs to compare distributions of categorical data			
I can describe the association between two categorical variables			
I can make and interpret dotplots, stemplots, and histograms of quantitative data			
I can identify the shape of a distribution from a graph			
I can describe the overall pattern (shape, center, and variability) of a distribution and identify any major departures from the pattern (outliers)			
I can compare distributions of quantitative data using dotplots, stemplots, and histograms			
I can calculate and interpret measures of center (mean and median) for a distribution of quantitative data			
I can calculate and interpret measures of variability (range, *IQR*, and standard deviation) for a distribution of quantitative data			
I can explain how outliers and skewness affect measures of center and variability			
I can identify outliers using the 1.5 x *IQR* Rule			
I can make and interpret boxplots of quantitative data			
I can use boxplots and numerical summaries to compare distributions of quantitative data			

Chapter 1 Multiple Choice Practice

Directions. *Identify the choice that best completes the statement or answers the question. Check your answers and note your performance when you are finished.*

1. You measure the age (years), weight (pounds), and breed (beagle, golden retriever, pug, or terrier) of 200 dogs. How many variables did you measure?

(A) 1
(B) 2
(C) 3
(D) 200
(E) 203

2. You open a package of Lucky Charms cereal and count how many there are of each marshmallow shape. The distribution of the variable "marshmallow" is:

(A) The shape: star, heart, moon, clover, diamond, horseshoe, balloon.
(B) The total number of marshmallows in the package.
(C) Seven—the number of different shapes that are in the package.
(D) The seven different shapes and how many there are of each.
(E) Since "shape" is a categorical variable, it doesn't have a distribution.

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

	Male	Female
Democrat	300	600
Republican	500	300
Other	200	100

The proportion of males that are registered as Democrats is

(A) 300
(B) 30
(C) 0.33
(D) 0.30
(E) 0.15

4. For a physics course containing 10 students, the maximum point total for the quarter was 200. The point totals for the 10 students are given in the stemplot below.

```
11 | 6  8
12 | 1  4  8
13 | 3  7
14 | 2  6
15 |
16 |
17 | 9
```

Which of the following statements is NOT true?

(A) In a symmetric distribution, the mean and the median are equal.
(B) About fifty percent of the scores in a distribution are between the first and third quartiles.
(C) In a symmetric distribution, the median is halfway between the first and third quartiles.
(D) The median is always greater than the mean.
(E) The range is the difference between the largest and the smallest observation in the data set.

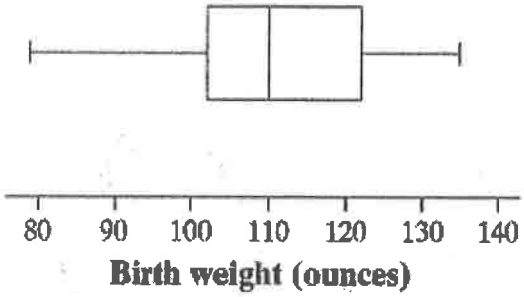
5. When drawing a histogram it is important to

(A) have a separate class interval for each observation to get the most informative plot.
(B) make sure the heights of the bars exceed the widths of the class intervals so that the bars are true rectangles.
(C) label the vertical axis so the reader can determine the counts or percent in each class interval.
(D) leave large gaps between bars. This allows room for comments.
(E) scale the vertical axis according to the variable whose distribution you are displaying.

6. A set of data has a mean that is much larger than the median. Which of the following statements is most consistent with this information?

(A) The distribution is symmetric.
(B) The distribution is skewed left.
(C) The distribution is skewed right.
(D) The distribution is bimodal.
(E) The data set probably has a few low outliers.

7. The following is a boxplot of the birth weights (in ounces) of a sample of 160 infants born in a local hospital.



About 40 of the birth weights were below

(A) 92 ounces.
(B) 102 ounces.
(C) 112 ounces.
(D) 122 ounces.
(E) 132 ounces.

8. A sample of production records for an automobile manufacturer shows the following figures for production per shift:

705 700 690 705

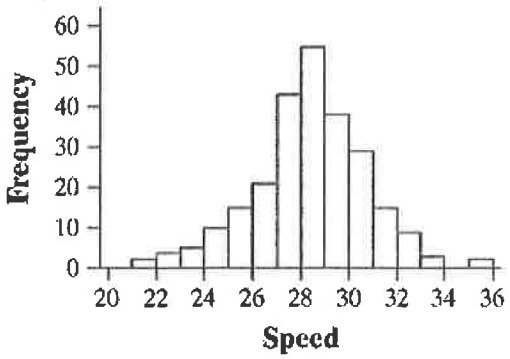
The variance of the sample is approximately

(A) 8.66.
(B) 7.07.
(C) 75.00.
(D) 50.00.
(E) 20.00.

9. You catch 10 cockroaches in your bedroom and measure their lengths in centimeters. Which of these sets of numerical descriptions are *all* measured in centimeters?

(A) median length, variance of lengths, largest length
(B) median length, first and third quartiles of lengths
(C) mean length, standard deviation of lengths, median length
(D) mean length, median length, variance of lengths.
(E) both (B) and (C)

10. A policeman records the speeds of cars on a certain section of roadway with a radar gun. The histogram below shows the distribution of speeds for 251 cars.



Which of the following measures of center and spread would be the best ones to use when summarizing these data?

(A) Mean and interquartile range.
(B) Mean and standard deviation.
(C) Median and range.
(D) Median and standard deviation.
(E) Median and interquartile range.

Check your answers below. If you got a question wrong, check to see if you made a simple mistake or if you need to study that concept more. After you check your work, identify the concepts you feel very confident about and note what you will do to learn the concepts in need of more study.

#	Answer	Concept	Right	Wrong	Simple Mistake?	Need to Study More
1	C	Variables				
2	D	Categorical variables				
3	D	Two-way table				
4	D	Distribution basics				
5	C	Constructing histograms				
6	C	Skewed distributions				
7	B	Interpreting boxplots				
8	D	Variance				
9	E	Summary statistics units				
10	B	Choosing statistics				

Chapter 1 Reflection

Summarize the "Big Ideas" in Chapter 1:

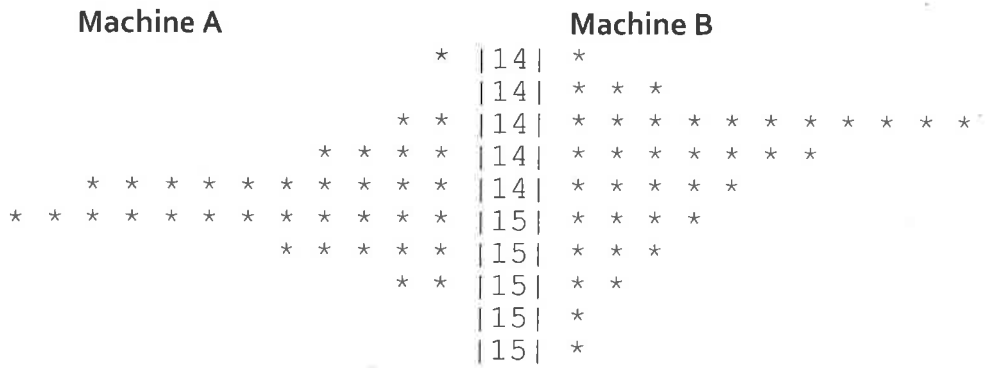
My strengths in this chapter:

Concepts I need to study more and what I will do to learn them:

FRAPPY! Free Response AP® Problem, Yay!

The following problem is modeled after actual Advanced Placement Statistics free response questions. Your task is to generate a complete, concise response in 15 minutes. After you generate your response, view two example solutions and determine whether you feel they are "complete", "substantial", "developing" or "minimal". If they are not "complete", what would you suggest to the student who wrote them to increase their score? Finally, you will be provided with a rubric. Score your response and note what, if anything, you would do differently to increase your own score.

SugarBitz candies are packaged in 15 oz. snack-size bags. The back-to-back plot below displays the weights (in ounces) of two samples of SugarBitz bags filled by different filling machines. The weights ranged from 14.1 oz. to 15.9 oz.



(a) Compare the distributions of weights of bags packaged by the two machines.

(b) The company wishes to be as consistent as possible when packing its snack bags. Which machine would be *least* likely to produce snack bags of SugarBitz that have a consistent weight? Explain.

(c) Suppose the company filled its bags using the machine you chose in part (b). Which measure of center, mean or median, would be closer to the advertised 15oz.? Explain why you chose this measure.

FRAPPY! Student Responses

Student Response 1:

a) Machine A has a slightly higher center than Machine B. Machine B has a much larger range. Machine A is approximately symmetric and Machine B is slightly skewed right. Neither machine has any extreme values.
b) Machine B would be least likely to produce bags containing 15 oz of SugarBitz because it has a much wider range than Machine A.
c) The company should report the mean weight of Machine B. Since the distribution is skewed to the right, the mean will be pulled higher towards the tail. Therefore, the mean will be higher than the median and will be closer to 15.

How would you score this response? Is it substantial? Complete? Developing? Minimal? Is there anything this student could do to earn a better score?

Student Response 2:

a) Machine A is normal and Machine B is skewed. Both have a single peak and wide ranges.
b) Machine B usually fills bags with about 14.6 oz of candy. Machine A usually fills bags with 15 oz of candy. Machine B is least likely to fill the bags with 15 oz. of candy.
c) The mean because it is about 15.

How would you score this response? Is it substantial? Complete? Developing? Minimal? Is there anything this student could do to earn a better score?

FRAPPY! Scoring Rubric

Use the following rubric to score your response. Each part receives a score of "Essentially Correct," "Partially Correct," or "Incorrect." When you have scored your response, reflect on your understanding of the concepts addressed in this problem. If necessary, note what you would do differently on future questions like this to increase your score.

Intent of the Question

The goals of this question are (1) to determine your ability to use graphical displays to compare and contrast two distributions and (2) to evaluate your ability to use statistical information to make a decision.

Solution

(a) Both distributions are single-peaked. However, Machine A's distribution is roughly symmetric while Machine B's is skewed to the right. The center of the weights for Machine A (median A = about 15) is slightly higher than that of Machine B (median B = about 14.5). There is more variability in the weights produced by Machine B. Machine A has one low value (14.1) that does not fall with the majority of weights. However, it does not appear to be extreme enough to be considered an outlier.

(b) Both machines produce bags of varying weight. However, Machine B has a higher variability as evidenced by a wider overall range. Machine B would be least likely to produce a consistent weight for the snack bags.

(c) The mean would be closer to the advertised 15 oz. weight. This is because in a skewed distribution, the mean is pulled away from the median in the direction of the tail. In Machine B's distribution, the peak is at about 14.5 oz so we would expect the mean to be higher and closer to 15 oz.

Scoring:

Parts (a), (b), and (c) are scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is essentially correct if you correctly identify similarities and differences in the shape, center, and spread for the two distributions.
Part (a) is partially correct if you correctly identify similarities and differences in two of the three characteristics for the two distributions.
Part (a) is incorrect if you only identify one similarity or difference of the three characteristics for the two distributions.

Part (b) is essentially correct if Machine B is chosen using rationale based on its measure of spread of the packaged weights.
Part (b) is partially correct if B is chosen, but the explanation does not refer to the variability in the weights.

Part (c) is incorrect if B is chosen and no explanation is provided OR if A is chosen.

Part (c) is essentially correct if the mean is chosen and the explanation addresses the fact that the mean will be greater than the median in a skewed right distribution.
Part (c) is partially correct if the mean is chosen, but the explanation is incomplete or incorrect.
Part (c) is incorrect if the mean is chosen, but no explanation is given OR if the median is chosen.
NOTE: If Machine A was chosen in part (b) and the solution to part (c) indicates either the mean or median would be appropriate due to the fact that they will be approximately equal in a symmetric, mound-shaped distribution, part (c) should be scored as essentially correct.

4 Complete Response
All three parts essentially correct

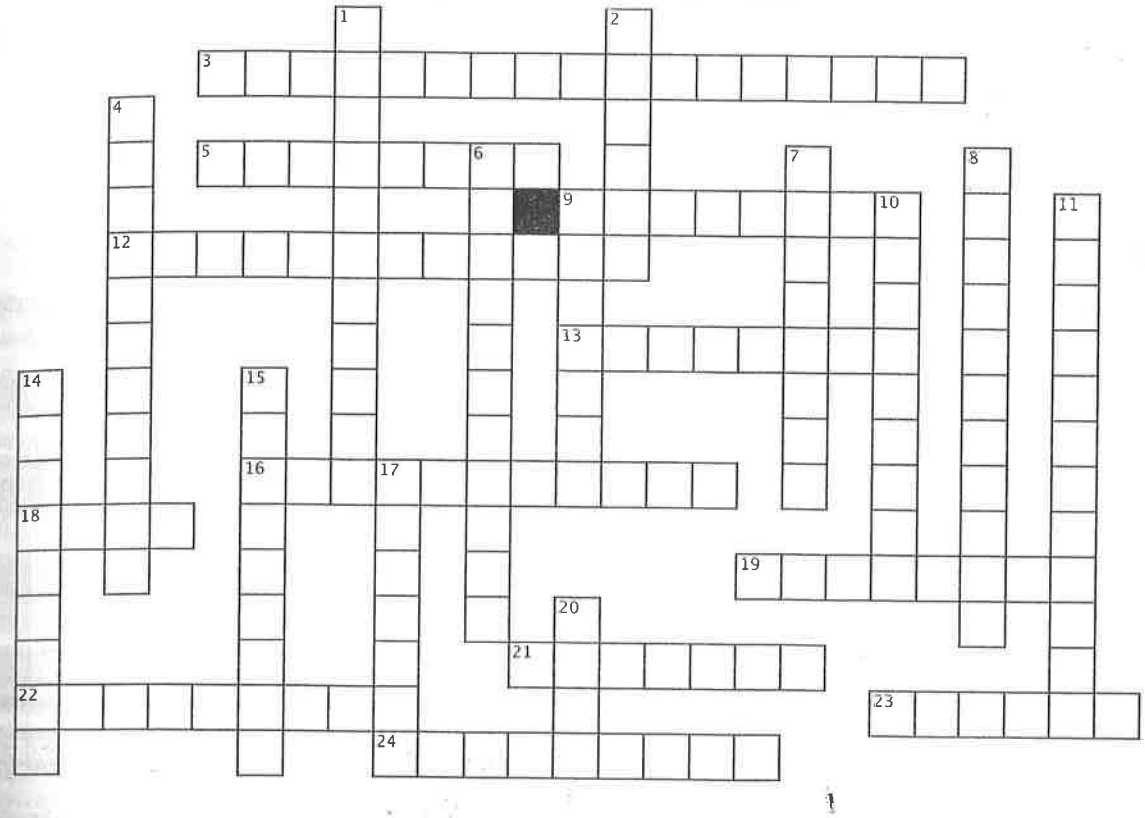
3 Substantial Response
Two parts essentially correct and one part partially correct

2 Developing Response
Two parts essentially correct and no parts partially correct
One part essentially correct and two parts partially correct
Three parts partially correct

1 Minimal Response
One part essentially correct and one part partially correct
One part essentially correct and no parts partially correct
No parts essentially correct and two parts partially correct

My Score:
What I did well:
What I could improve:
What I should remember if I see a problem like this on the AP Exam:

Chapter 1: Exploring Data



Across

3. The average distance of observations from their mean (two words)
5. The average squared distance of the observations from their mean
9. Displays the counts or percents of categories in a categorical variable through differing heights of bars
12. Tells you what values a variable takes and how often it takes these values
13. Displays a categorical variable using slices sized by the counts or percents for the categories
16. When specific values of one variable tend to occur in common with specific values of another
18. A measure of center, also called the average
19. A graphical display of quantitative data that involves splitting the individual values into two components
21. One of the simplest graphs to construct when dealing with a small set of quantitative data
22. Drawing conclusions beyond the data at hand
23. The shape of a distribution if one side of the graph is much longer than the other
24. What we call a measure that is relatively unaffected by extreme observations

Down

1. The objects described by a set of data
2. The midpoint of a distribution of quantitative data
4. A _____ distribution describes the distribution of values of a categorical variable among individuals who have a specific value of another variable.
6. A variable that places an individual into one of several groups or categories
7. A characteristic of an individual that can take different values for different individuals
8. When comparing two categorical variables, we can orgainze the data in a ___-___ ____.
9. A graphical display of the five-number summary
10. A graphical display of quantitative data that shows the frequency of values in intervals by using bars
11. A variable that takes numerical values for which it makes sense to find an average
14. The shape of a distribution whose right and left sides are approximate mirror images of each other
15. These values lie one-quarter, one-half, and three-quarters of the way up the list of quantitative data
17. A value that is at least 1.5 IQRs above the third quartile or below the first quartile
20. When exploring data, don't forget your ___