AP Statistics NAME:_____

Exam Review: Strand 1: Exploring Data Date:_____ Block:_____

Exploring Data (20 – 30%): Describing patterns and departures from patterns

(Statistics in Action 2nd Edition – Chapters 2 & 3)

Exploratory analysis of data makes use of graphical and numerical techniques to study patterns and departures from patterns. Emphasis should be placed on interpreting information from graphical and numerical displays and summaries.

1. Constructing and interpreting graphical displays of distributions of univariate data (dotplot, stemplot, histogram, cumulative frequency plot)
    1. Center and spread
    2. Clusters and gaps
    3. Outliers and other unusual features
    4. Shape
2. Summarizing distributions of univariate data
    1. Measuring center: median, mean
    2. Measuring spread: range, interquartile range, standard deviation
    3. Measuring position: quartiles, percentiles, standardized scores (z-scores)
    4. Using boxplots
    5. The effect of changing units on summary measures
3. Comparing distributions of univariate data (dotplots, back-to-back stemplots, parallel boxplots)
    1. Comparing center and spread: within group, between group variation
    2. Comparing clusters and gaps
    3. Comparing outliers and other unusual features
    4. Comparing shapes
4. Exploring bivariate data
    1. Analyzing patterns in scatterplots
    2. Correlation and linearity
    3. Least-squares regression line
    4. Residual plots, outliers, and influential points
    5. Transformations to achieve linearity: logarithmic and power transformations
5. Exploring categorical data
    1. Frequency tables and bar charts
    2. Marginal and joint frequencies for two-way tables
    3. Conditional relative frequencies and association
    4. Comparing distributions using bar charts

Review Notes

UNIVARIATE DATA

Types of Variables	Appropriate Graphical Display(s)

Shapes of Distributions			
Graph			
Important Info			
Descriptions (Shape, Center, Spread)			

Effects of Changing Units

Change	Shape	Measure of Center	Measure of Spread

Resistance to Outliers

Examples: During the early part of the 2004 baseball season, many sports fans and baseball players noticed that the number of home runs being hit seemed to be unusually large. Here are the data on the number of home runs hit by American League teams.

American League: 35, 40, 43, 49, 51, 54, 57, 58, 58, 64, 68, 68, 75, 77

Construct an appropriate graph to display of the data and describe the distribution.

The table below displays the ages at the time of their inauguration of the first 43 presidents of the United States. Construct an appropriate display and describe the distribution.
57, 65, 55, 61, 52, 51, 57, 56, 54, 57, 46, 51, 58, 54, 60, 57, 49, 61, 61, 51, 43, 54, 47, 55, 68, 55, 56, 51, 55, 61, 49, 54, 52, 64, 42, 69, 50, 51, 64, 48, 56, 46, 54

AP Statistics NAME:_____

HW: AP Statistics Review: Exploring Data #1 Date:_____ Block:_____

Use this data for Questions 1–5: A survey was conducted to gather ratings of the quality of service at a restaurant at a nearby mall. Customers were to rate overall service using values between 0 (terrible) and 100 (excellent). The ratings of a sample of customers are given in this stemplot. (*Note:* A calculator should not be necessary for Questions 1–4.)

```
Stem  Leaves
3     24
4     03478999
5     0112345
6     12566
7     01
8
9     2
```

1. What percentage of the respondents rated quality as very good or higher (a rating of 80 or more)?

(A) 0% (B) 4% (C) 25%
(D) 96% (E) 100%

2. What is the shape of this distribution?

(A) symmetric
(B) skewed left
(C) skewed right
(D) uniform
(E) approximately normal

3. The mean of these ratings is

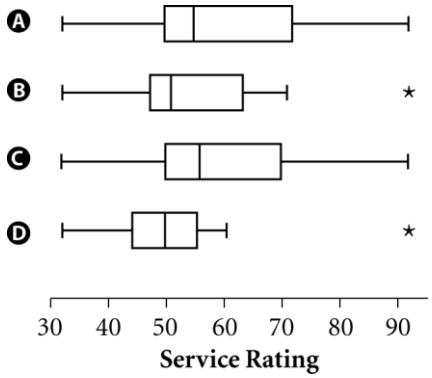
(A) equal to the median
(B) less than the median
(C) greater than the median
(D) an integer
(E) at the 13th rating

4. What is the largest a rating could be without being an outlier?

(A) 63.5 (B) 70 (C) 75
(D) 87.5 (E) 92

5. The boxplot of the ratings is

(A)
(B)
(C)
(D)



(E) None of these is a boxplot of the ratings.

6. The average income, in dollars, of people in each of the 50 states was computed for 1980 and for 2000. Summary statistics for these two distributions are given here.

	1980	2000
Mean	9,725	28,336
Standard Deviation	1,503	4,413
Minimum	7,007	21,007
Lower Quartile	8,420	25,109
Median	9,764	28,045
Upper Quartile	10,746	30,871
Maximum	14,866	41,495

a. Explain the meaning of $7,007 for the minimum in 1980.

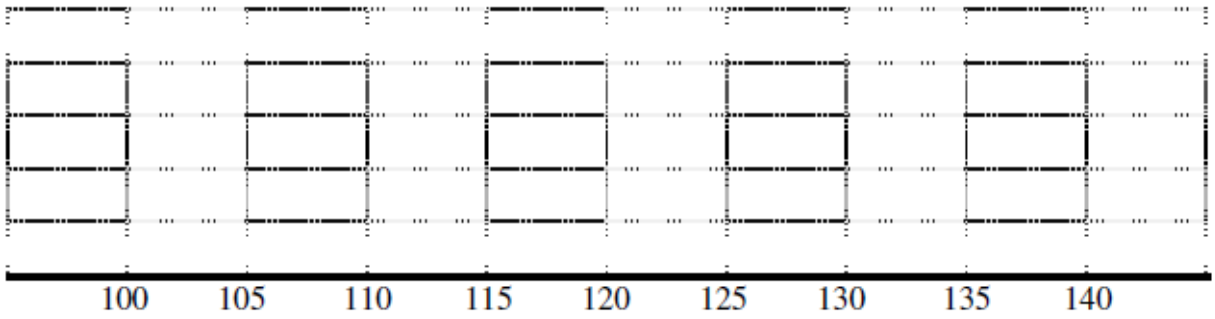
b. Are any states outliers for either year?

c. In 2000 the average personal income in Alabama was $23,768, and in 1980 it was $7,836. Did the income in Alabama change much in relation to the other states? Explain your reasoning.

1. A researcher thinks that modern Thai dogs may be descendants of golden jackals. A random sample of 16 animals was collected from each of the two populations. The length (in millimeters) of the mandible (jawbone) was measured for each animal. The lower quartile, median, and upper quartile for each sample are shown in the table below, along with all values below the lower quartile and all values above the upper quartile

Sample	Values Below Q_1	Q_1	Median	Q_3	Values Below Q_3
Modern Thai dog	114, 116, 116, 120	121	125	128	129, 130, 130, 132
Golden jackal	104, 104, 105, 106	107	108	112	114, 122, 124, 125

(a) Display parallel boxplots of mandible lengths (showing outliers, if any) for the modern Thai dogs and the golden jackals on the grid below.



Based on the boxplots, write a few sentences comparing the distributions of mandible lengths for the two types of dogs.

AP1. These summary statistics are for the distribution of the populations of the major cities in Brazil.

Variable	N	Mean	Median	TrMean	StDev	SEMean
Population	222	381056	191348	261985	820246	55051

Variable	Min	Max	Q1	Q3
Population	100049	10009231	129542	324323

Which of the following best describes the shape of this distribution?

(A) skewed right without outliers
(B) skewed right with at least one outlier
(C) roughly normal, without outliers
(D) skewed left without outliers
(E) skewed left with at least one outlier

AP2. Which of these lists contains only summary statistics that are sensitive to outliers?

(A) mean, median, and mode
(B) standard deviation, IQR, and range
(C) mean and standard deviation
(D) median and IQR
(E) five-number summary

AP3. This stem-and-leaf plot shows the ages of CEOs of 60 corporations whose annual sales were between $5 million and $350 million. Which of the following is not a correct statement about this distribution?

(A) The distribution is skewed left (towards smaller numbers).
(B) The oldest of the 60 CEOs is 74 years old.
(C) The distribution has no outliers.
(D) The range of the distribution is 42.
(E) The median of the distribution is 50.

```
Stem-and-leaf of AGE N = 60
Leaf Unit = 1.0

3    23
3    678
4    013344
4    555566777888889
5    000000112333
5    555666677889
6    0111223
6    99
7    04
```

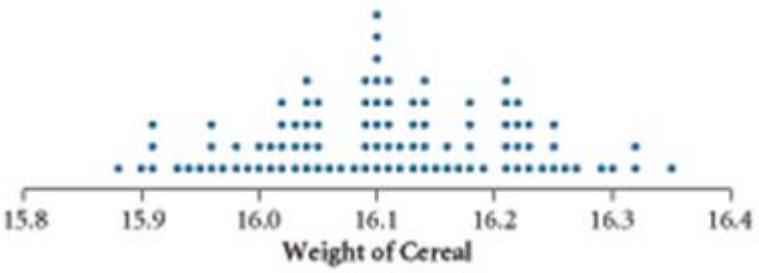
AP4. A traveler visits Europe and stays thirty days in thirty different hotels, paying each day with her credit card. The hotels charged a mean price of 50 euros, with a standard deviation of 10 euros. When the charges appear on her credit card statement in the United States, she finds that her bank charged her $1.20 per euro, plus a $5 fee for each transaction. What is the mean and standard deviation of the thirty daily hotel charges in dollars, including the fee?

(A) mean $50, standard deviation $17
(B) mean $60, standard deviation $12
(C) mean $60, standard deviation $17
(D) mean $65, standard deviation $12
(E) mean $65, standard deviation $17

AP5. The scores on a nationally administered test are approximately normally distributed with mean 47.3 and standard deviation 17.3. Approximately what must a student have scored to be in the 95th percentile nationally?

(A) 55 (B) 61 (C) 73 (D) 76 (E) 81

AP6. A particular brand of cereal boxes is labeled "16 oz." This dot plot shows the actual weights of 100 randomly selected boxes. Which of the following is the best estimate of the standard deviation of these weights?



(A) 0.04 oz. (B) 0.1 oz. (C) 0.2 oz. (D) 0.4 oz.
(E) between 16.0 and 16.2 oz.

AP7. The distribution of the number of points earned by the thousands of contestants in the Game of Pig World Championship has mean 20 and standard deviation 6. What proportion of the contestants earned more than 26 points?

(A) Less than 1% (B) 16%
(C) 32% (D) 84%
(E) This proportion cannot be determined from the information given.

1. The following list is ordered from smallest to largest: 25, 26, 26, 30, y, y, y, 33, 150. Which of the following statements is (are) true?

   I. The mean is greater than the median
   II. The mode is 26
   III. There are no outliers in the data

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

2. Jenny is 5′10″ tall and is worried about her height. The heights of girls in the school are approximately normally distributed with a mean of 5′5″ and a standard deviation of 2.6″. What is the percentile rank of Jenny's height?

   a. 59
   b. 65
   c. 74
   d. 92
   e. 97

3. The mean and standard deviation of a normally distributed dataset are 19 and 4, respectively. 19 is subtracted from every term in the dataset and then the result is divided by 4. Which of the following best describes the resulting distribution?

   a. It has a mean of 0 and a standard deviation of 1.
   b. It has a mean of 0, a standard deviation of 4, 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 normal.
   e. It has a mean of 0, a standard deviation of 4, and its shape is unknown.

4. The five-number summary for a one-variable dataset is {5, 18, 20, 40, 75}. If you wanted to construct a modified boxplot for the dataset (that is, one that would show outliers if there are any), what would be the maximum possible length of the right side "whisker"?

   a. 35
   b. 33
   c. 5
   d. 55
   e. 53

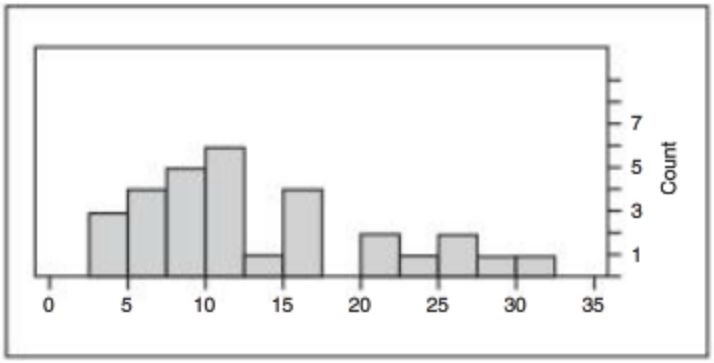
5. A set of 5,000 scores on a college readiness exam are known to be approximately normally distributed with mean 72 and standard deviation 6. To the nearest integer value, how many scores are there between 63 and 75?

   a. 0.6247
   b. 4,115
   c. 3,650
   d. 3,123
   e. 3,227

6. For the data given in #5 above, suppose you were not told that the scores were approximately normally distributed. What can be said about the number of scores that are less than 58 (to the nearest integer)?

   a. There are at least 919 scores less than 58.
   b. There are at most 919 scores less than 58.
   c. There are approximately 919 scores less than 58.
   d. There are at most 459 scores less than 58.
   e. There are at least 459 scores less than 58.

7. The following histogram pictures the number of students who visited the Career Center each week during the school year.



The shape of this graph could best be described as

a. Mound-shaped and symmetric
b. Bi-modal
c. Skewed to the left
d. Uniform
e. Skewed to the right

8. Which of the following statements is (are) true?
   I. The median is resistant to extreme values.
   II. The mean is resistant to extreme values.
   III. The standard deviation is resistant to extreme values.
   a. I only
   b. II only
   c. III only
   d. II and III only
   e. I and III only

9. One of the values in a normal distribution is 43 and its *z*-score is 1.65. If the mean of the distribution is 40, what is the standard deviation of the distribution?
   a. 3
   b. –1.82
   c. –0.55
   d. 1.82
   e. –0.55

10. Free-response questions on the AP Statistics Exam are graded on 4, 3, 2, 1, or 0 basis. Question #2 on the exam was of moderate difficulty. The average score on question #2 was 2.05 with a standard deviation of 1. To the nearest tenth, what score was achieved by a student who was at the 90th percentile of all students on the test? You may assume that the scores on the question were approximately normally distributed.
    a. 3.5
    b. 3.3
    c. 2.9
    d. 3.7
    e. 3.1

1. B. One out of the 25 ratings, or 4%.
2. C
3. C. The mean of 54.16 is greater than the median of 51.
4. D. $Q_3 + 1.5IQR = 63.5 + 1.5(63.5 - 47.5) = 87.5$
5. B. This boxplot shows a minimum at 32, Q_1 at 47.5, median at 51, Q_3 at 63.5, and a maximum at 71, with an outlier at 92.
6. *Note:* This is Chapter 2, E91.

   a. The state with the lowest average income in dollars in 1980 had an average income of $7,007.

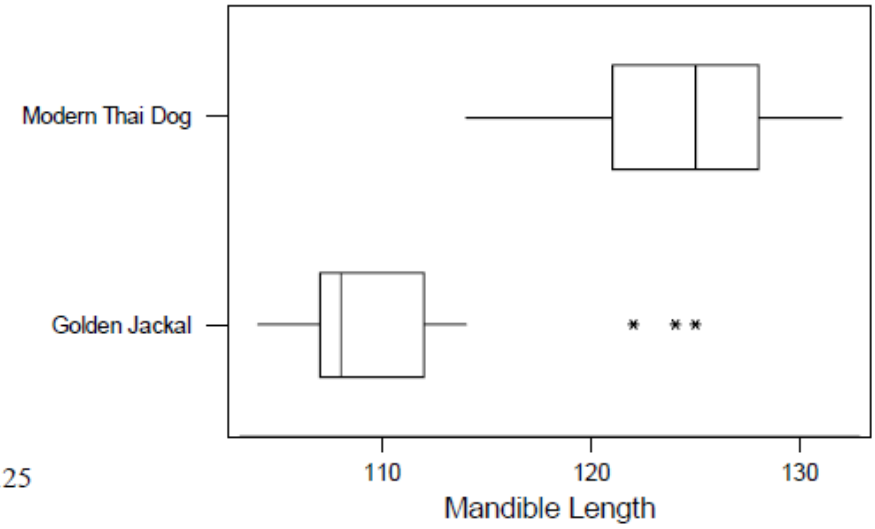
   b. There is at least one outlier for 1980 because $Q_3 + 1.5 \cdot IQR$ = 10,746 + 3,489 = $14,235, which is less than the maximum of $14,866. Thus, the state with the maximum income is an outlier. There is also at least one outlier for 2000 because $Q_3 + 1.5 \cdot IQR$ = 30,871 + 8,643 = $39,514, which is less than the maximum of $41,495. Other states on the high end may also be outliers. There are no outliers on the low end for either year.

   c. No. Alabama remains below the lower quartile of the distribution, but you cannot say exactly where in the lowest quarter it lies. It is tempting to find Alabama's relative position using *z*-scores, but there is no indication that these incomes are normally distributed, and it is only appropriate to use *z*-scores for comparison in that case.

Part (a):

Modern Thai Dogs
n = 16
IQR = 128 - 121 = 7
Outlier boundaries:
121 – 1.5(7) = 110.5
128 + 1.5(7) = 138.5
no outliers for Modern Thai Dogs

Golden Jackals
n = 16
IQR = 112 -107 = 5
Outlier boundaries:
107 – 1.5(5) = 99.5
112 + 1.5(5) = 119.5
outliers on the high end: 122, 124, 125

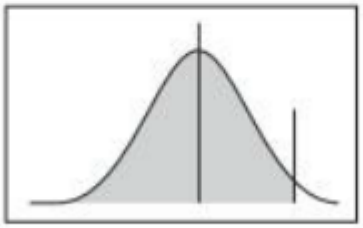


The distributions of mandible lengths for Modern Thai Dogs and Golden Jackals are not similar. The distribution of mandible lengths for Modern Thai Dogs is approximately symmetric and a typical value is about 125, whereas the distribution of mandible lengths for Golden Jackals has a typical value that is much smaller, around 108, and the distribution appears to be skewed to the right with outliers (relative to likely samples from a normal distribution) on the high end. The variability in the lengths is roughly the same for both types of dogs.

AP1. B
AP2. C
AP3. A
AP4. D
AP5. D
AP6. B
AP7. E. Without knowing whether the distribution of scores is close to normal, you can't make an accurate assessment of this probability.

1. The correct answer is (a). I is correct since the mean is pulled in the direction of the large maximum value, 150 (well, large compared to the rest of the numbers in the set). II is not correct because the mode is *y*—there are three *y*s and only two 26s. III is not correct because 150 is an outlier (you can't actually compute the upper boundary for an outlier since the third quartile is *y*, but even if you use a larger value, 33, in place of *y*, 150 is still an outlier).

2. The correct answer is (e).
$z = \frac{70-65}{2.6} = 1.92 \rightarrow$ percentile = 0.9726 (see drawing below):



(On the TI-83/84, `normalcdf (-100, 1.92) = normalcdf (-1000, 70, 65, 206) = 0.9726` up to rounding error.)

3. The correct answer is (d). The effect on the mean of a dataset of subtracting the same value is to reduce the old mean by that amount (that is, $\mu_{x-k} = \mu_x - k$). Because the original mean was 19, and 19 has been subtracted from every term, the new mean is 0. The effect on the standard deviation of a dataset of dividing each term by the same value is to divide the standard deviation by that value, that is,

$$\sigma_{\frac{x}{k}} = \frac{\sigma_x}{k}.$$

Because the old standard deviation was 4, dividing every term by 4 yields a new standard deviation of 1. Note that the process of subtracting the mean from each term and dividing by the standard deviation creates a set of *z*-scores

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

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.

4. The correct answer is (b). The maximum length of a "whisker" in a modified boxplot is 1.5(IQR) = 1.5(40 – 18) = 33.

5. The correct (best) answer is (d). Using Table A, the area under a normal curve between 63 and 75 is 0.6247 ($z_{63} = -1.5 \Rightarrow A_1 = 0.0668$, $z_{75} = 0.5 \Rightarrow A_2 = 0.6915 \Rightarrow A_2 - A_1 = 0.6247$). Then (0.6247)(5,000) = 3123.5. Using the TI-83/84, `normalcdf(63,75,72,6) × 5000 = 3123.3.`

6. The correct answer is (b). Since we do not know that the empirical rule applies, we must use Chebyshev's rule.
   Since $72 - k(6) = 58$, we find $k = 2.333$. Hence, there are at most $\frac{1}{2.333^2}\% = 18.37\%$ of the scores less than 58. Since there are 5000 scores, there are at most (0.1837)(5,000) = 919 scores less than 58. Note that it is unlikely that there are this many scores below 58 (since some of the 919 scores could be more than 2.333 standard deviation *above* the mean)—it's just the strongest statement we can make.

7. The correct answer is (c). The graph is clearly not symmetric, bi-modal, or uniform. It is skewed to the right since that's the direction of the "tail" of the graph.

8. The correct answer is (a). The median is resistant to extreme values, and the mean is not (that is, extreme values will exert a strong influence on the numerical value of the mean but not on the median). II and III involve statistics equal to or dependent upon the mean, so neither of them is resistant.

9. The correct answer is (d). $z = 1.65 = \frac{43-40}{\sigma} \Rightarrow \sigma = \frac{3}{1.65} = 1.82$.

10. The correct answer is (b). A score at the 90th percentile has a z-score of 1.28. Thus, $z_x = \frac{x-2.05}{1} = 1.28 \Rightarrow x = 3.33$.