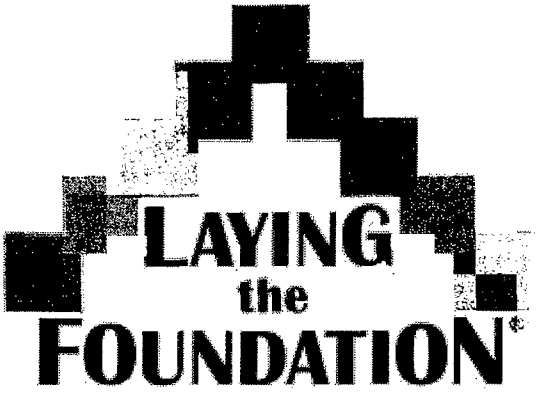


AP* Statistics Review

Descriptive Statistics

Teacher Packet

Describing or comparing distributions:
Always give the shape, center, and spread in the context of the question.

Shape - Tell what the graph looks like.

- Symmetric
- Skewed and direction of skewness
- Uniform
- Peaks (modes) and the location of the peaks
- Gaps in data and the location of the gaps
- Unusual values in the data and the location of those values

Center - Tell (numerically) where the center of the data is.

- Mean - average; μ for population; $\bar{x}$ for sample
- Median - middle value when data is listed from low to high
- Mode - peak in the distribution or most frequently occurring value

Spread or variability - Tell (numerically) how spread out the data is.

- Standard deviation – "average" distance between the data points and the mean
- Inter-Quartile Range (IQR) = range of the middle 50% of the data = $Q_3 - Q_1$
- Range = maximum – minimum

Identifying an outlier

- An unusual value in the data set which is too far from its quartiles
- Numerically, this is any value that exceeds $Q_3 + 1.5IQR$ or is less than $Q_1 - 1.5IQR$

Choosing appropriate measures of center and spread

- If the data is fairly symmetric, use the mean and standard deviation. These are not *resistant* to the presence of skewness or outliers.
- If the data is skewed or outliers are present, use the median and IQR instead.
- Use the mode or range for additional information or if the other measures cannot be calculated from the information given.

Drawing and/or interpreting graphs (by hand or on your calculator)

- Histogram (with either frequency or relative frequency on the vertical axis)
- Stemplot (also called stem-and-leaf plot)
- Dot plot (for small data sets)
- Boxplot and modified boxplot (which shows outliers)
- Cumulative frequency plot or ogive
- Normal quantile plot

Transforming data by multiplying/dividing or adding/subtracting

- Measures of center (mean, median, mode) transform just like any data point.
- Measures of spread (standard deviation, IQR, or range) are transformed only when the data is multiplied or divided by a constant.
- Example: You have measured the barefooted height (in inches) of everyone in your class and determined these values: mean = 67.0 inches, standard deviation = 1.9 inches, median = 66.2 inches, Q_1 = 64.0 inches, Q_3 = 68.9 inches. Now everyone puts on shoes with a 4.0 cm heel. Find the following measures in centimeters: mean, standard deviation, median, IQR. (1 inch = 2.54 centimeters.)

  Solution:
  Here is how the transformation is written:

$$\text{new height in cm} = \left(2.54\frac{\text{cm}}{\text{in}}\right)(\text{old height in inches}) + 4.0\text{cm}$$

  Now find the values asked for in the problem.

$$\text{new mean} = \left(2.54\frac{\text{cm}}{\text{in}}\right)(67.0\text{ in}) + 4.0\text{ cm} = 174\text{ cm}$$

$$\text{new standard deviation} = \left(2.54\frac{\text{cm}}{\text{in}}\right)(1.9\text{ in}) = 4.8\text{ cm}$$

$$\text{new median} = \left(2.54\frac{\text{cm}}{\text{in}}\right)(66.2\text{ in}) + 4.0\text{ cm} = 172\text{ cm}$$

$$\text{original } IQR = Q_3 - Q_1 = 68.9\text{ in} - 64.0\text{ in} = 4.9\text{ in}$$

$$\text{new } IQR = \left(2.54\frac{\text{cm}}{\text{in}}\right)(4.9\text{ in}) = 12.4\text{ cm}$$

Multiple Choice Questions on Descriptive Statistics

1. If the largest value of a data set is doubled, which of the following is **false**?

(A) The mean increases.
(B) The standard deviation increases.
(C) The interquartile range increases.
(D) The range increases.
(E) The median remains unchanged.

2. The five-number summary for scores on a statistics exam is 35, 68, 77, 83, 97. In all, 196 students took the test. About how many had scores between 77 and 83?

(A) 6
(B) 39
(C) 49
(D) 98
(E) It cannot be determined from the information given.

3. The following list is a set of data ordered from smallest to largest. All values are integers. 2 12 y y y 15 18 18 19

I. The median and the first quartile cannot be equal.
II. The mode is 18.
III. 2 is an outlier.

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

4. A substitute teacher was asked to keep track of how long it took her to get to her assigned school each morning. Here is a stem plot of the data. Would you expect the mean to be higher or lower than the median?

2	0002344578
3	0257
4	12789
5	028
6	05

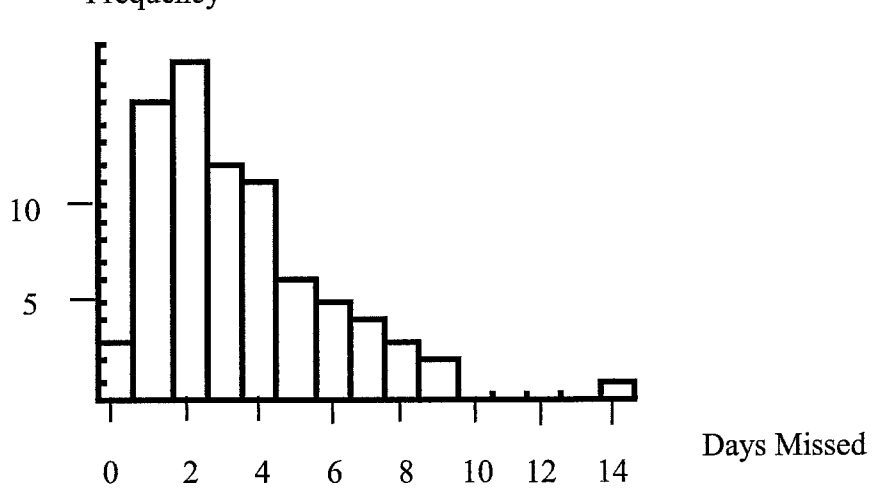
Key: 4|1 = 41 kilometers

(A) Lower, because the data are skewed to the left.
(B) Lower, because the data are skewed to the right.
(C) Higher, because the data are skewed to the left.
(D) Higher, because the data are skewed to the right.
(E) Neither, because the mean would equal the median.

5. A professor scaled (curved) the scores on an exam by multiplying the student's raw score by 1.2, then adding 15 points. If the mean and standard deviation of the scores before the curve were 51 and 5, respectively, then the mean and standard deviation of the scaled scores are respectively:

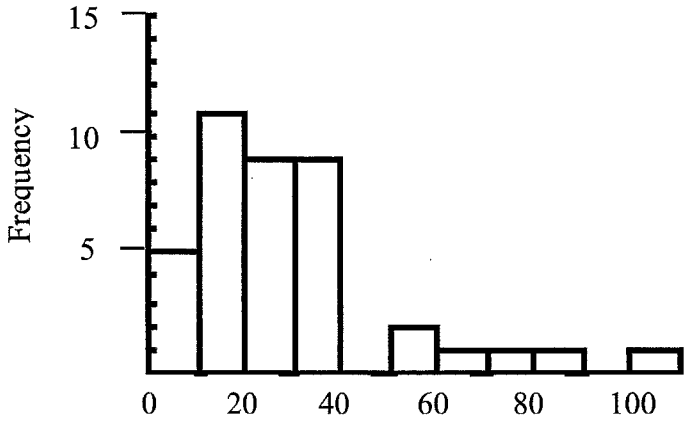
(A) 76.2 and 21
(B) 76.2 and 6
(C) 76.2 and 5
(D) 61 and 6
(E) cannot be determined without knowing if the scores are normally distributed

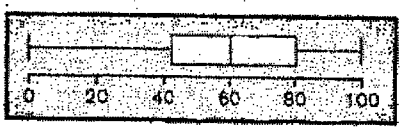
6. In the northern U.S., schools are often closed during severe snowstorms. These missed days must be made up at the end of the school year. The following histogram shows the number of days missed per year for a particular school district using data from the past 75 years. Which of the following should be used to describe the center of the distribution?

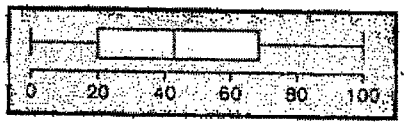


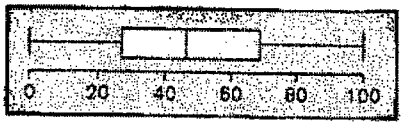
(A) Mean, because it uses information from all 75 years.
(B) Median, because the distribution is skewed.
(C) IQR, because it excludes outliers and includes the middle 50% of the data.
(D) Quartile 1, because the distribution is skewed to the right.
(E) Standard deviation, because it is unaffected by outliers.

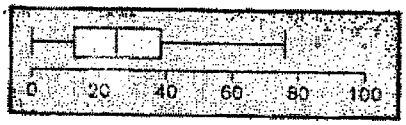
7. Which boxplot was made from the same data as this histogram?



(A) 

(B) 

(C) 

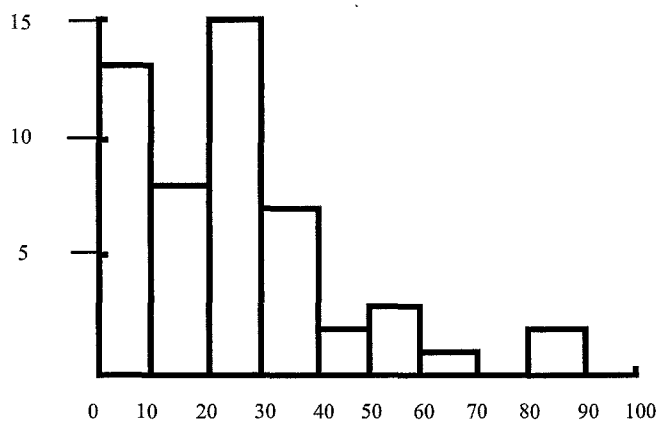
(D) 

(E) None of the above.

8. One advantage of using a stem-and-leaf plot rather than a histogram is that the stem-and-leaf plot

(A) shows the shape of the distribution more easily than the histogram.
(B) changes easily from frequency to relative frequency.
(C) shows all of the data on the graph.
(D) presents the percentage distribution of the data.
(E) shows the mean on the graph.

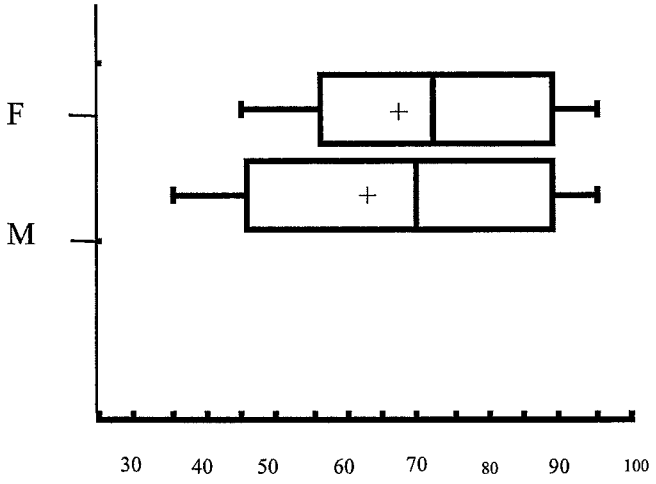
9. This histogram shows the closing price of a stock on 50 days.



In which range does the first quartile lie?

(A) 0 to 10
(B) 10 to 20
(C) 20 to 30
(D) 30 to 40
(E) 80 to 90

10. The scores of male (M) and female (F) students on a statistics exam are displayed in the following boxplots. The pluses indicate the location of the means.



Which of the following is correct?

(A) The mean grade of the females is about 72.
(B) About 75% of the males score above 82.
(C) The median of the male students is about 66.
(D) The scores of the males have a higher variability than the scores of the females.
(E) About 25% of the females scored above 72.

Free Response Questions on Descriptive Statistics

1. As a project in their physical education classes, elementary school students were asked to kick a soccer ball into a goal from a fixed distance away. Each student was given 8 chances to kick the ball, and the number of goals was recorded for each student. The number of goals for 200 first graders is given in the table.

Number of goals scored	Number of first graders
0	14
1	37
2	51
3	33
4	30
5	14
6	11
7	7
8	3

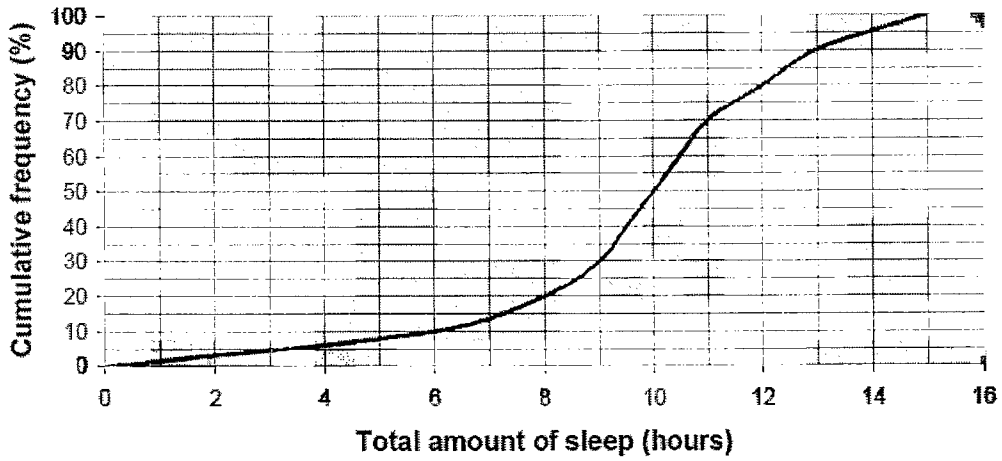
In order to compare whether older children are better at kicking goals, the exercise was repeated with 200 fourth graders.

Number of goals scored	Number of fourth graders
0	5
1	11
2	18
3	24
4	27
5	34
6	39
7	28
8	14

(a) Graph these two distributions so that the number of goals scored by the first graders and the number of goals scored by the fourth graders can be easily compared.

(b) Based on your graphs, how do the results from the fourth graders differ from those of the first graders? Write a few sentences to answer this question.

2. Students at a weekend retreat were asked to record their total amount of sleep on Friday and Saturday nights. The results are shown in the cumulative frequency plot below.



(a) The graph goes through the point (11, 70). Interpret this point in the context of the problem.

(b) Find the interquartile range for the total hours of sleep. Show your work. (Work on the graph counts as work shown.)

(c) Check the appropriate space below **and** explain your reasoning.

In this distribution,

____ the mean amount of sleep will be less than the median amount of sleep.
____ the mean amount of sleep will be equal to the median amount of sleep.
____ the mean amount of sleep will be greater than the median amount of sleep.

3. Employees of a British company are paid monthly salaries in British pounds (£). One division of the company will be relocated to France for a year, where their salaries will be paid in the currency of the European Union, the euro (€). One British pound is equal to 1.27 euros. While the employees are in France, each will also get a monthly bonus of €325 (or 325 euros).

The following are statistics for the employees' original salaries in Great Britain.

Minimum	£ 800
First quartile	£ 1250
Median	£ 1470
Third quartile	£ 2250
Maximum	£ 4500
Mean	£ 2025
Standard deviation	£ 475

(a) One employee earns £ 1600 per month in Great Britain. Calculate this person's monthly salary in euros (including bonus) after the relocation to France.

(b) Find the mean **and** standard deviation of the employees' monthly salaries in euros after the move to France. Show your work.

(c) Based on the salaries in Great Britain, are there any outliers in the salary data? Explain why or why not.

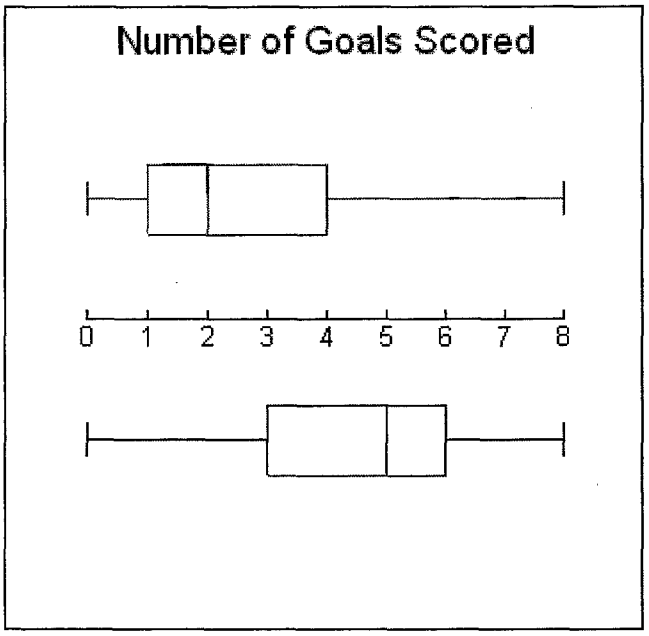
Key to Descriptive Statistics Multiple Choice

1. C Median and IQR are resistant, so they do not change.
2. C 25% of scores lie between median and Q_3; 25% of 196 is 49.
3. D The median and Q_1 are equal if $y = 12$. The mode is y. 2 is an outlier if the IQR is 6, which is the largest it could be.
4. D Skewed right so mean is pulled up toward outliers; mean > median
5. B Don't add 15 when transforming the standard deviation.
6. B Skewed, so you need to use resistant measures
7. D
8. C Can't do B, D, or E; can see shape equally well on both
9. A When the 46 data points are put in order from lowest to highest, Q_1 would be the 12^{th} value. Since there are 13 values in the first histogram bar, Q_1 is in the range 0 to 10.
10. D IQR for males > IQR for females (width of box)

Rubric for Descriptive Statistics Free Response

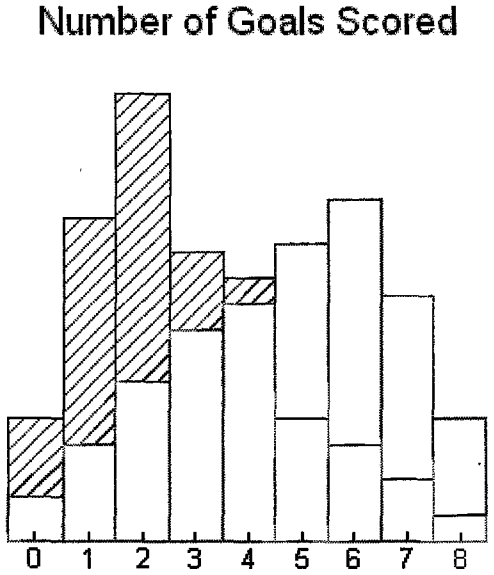
1. **Solution**

Part (a): Show either side-by-side boxplots OR histograms on the same scale.



First graders

Fourth graders



Shaded bars are first graders;
Unshaded bars are fourth graders

Part (b):

The fourth graders tended to score more goals The 4th graders' mean (4.695), median (5) and mode (6) were all larger than those of the first graders (mean = 2.835, median = 2, mode = 2). Both grades have approximately the same variability (for fourth graders, standard deviation = 2.1, IQR = 3 and for first graders, standard deviation = 1.9 and IQR = 3). The distribution for the goals scored by the fourth graders is skewed left while the distribution for the first graders is skewed right.

Scoring

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

Part (a) is correct if either side-by-side boxplots, overlapping histograms, or individual histograms are made on the same scale. Graphs must be labeled and have values on the axes for scaling. Since sample sizes are equal, either frequency or relative frequency could be used on the vertical axis of the histogram(s).

Part (a) is partially correct if graphs are started correctly but are incomplete OR scaling is absent.

Part (a) is incorrect if frequencies are interpreted as data.

Part (b) is essentially correct if graphs are interpreted in context, explicitly comparing at least two of shape, center, and spread.

Part (b) is partially correct if interpretation is correct but not in context

OR

comparison of groups is made on only one of shape, center, or spread

OR

values are given but not compared on at least 2 of shape, center, and spread

Part (b) is incorrect if student fails to compare the groups on any of shape, center, and spread.

4 Complete Response

Both parts essentially correct.

3 Substantial Response

One part essentially correct and one part partially correct

2 Developing Response

One part essentially correct and one part incorrect

OR

Partially correct on both parts

1 Minimal Response

Partially correct on one part.

2. **Solution**

Part (a):

70% of the students got 11 hours of sleep or less on the weekend retreat.

Part (b):

Since 25% of the students got 8.5 hours of sleep or less, Q1 = 8.5 hours. Since 75% of the students got 11.5 hours of sleep or less, Q3 = 11.5 hours. The IQR is $Q_3 - Q_1 = 11.5 - 8.5 = 3$ hours.

Part (c):

The distribution is skewed to the left (five number summary is 0, 8.5, 10, 11.5, 15), so the mean is less than the median.

Scoring

Part (a) is either essentially correct (E) or incorrect (I). Parts (b) and (c) are either essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is incorrect if the values are interpreted without context or if the cumulative nature of the graph is not recognized (for example, if the student says that 70% of the students got 11 hours of sleep).

Part (b) is essentially correct if both Q1 and Q3 are found, the IQR is correctly calculated, and work is shown (on the graph) or an appropriate explanation is given.

Part (b) is partially correct if incorrect values are read from the plot but the explanation is correct.

Part (b) is incorrect if a numerical answer (including 3 hours) is given with no supporting work.

Part (c) is essentially correct if the student checks that the mean will be less than the median and states that this is because the distribution is skewed to the left.

Part (c) is partially correct if the student checks the correct space but has an incomplete explanation OR if the student checks the wrong space but states that the distribution is skewed to the left.

Part (c) is incorrect if the student checks the wrong space and gives no explanation or a wrong explanation.

4 Complete Response

All 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
OR
One part essentially correct and two parts partially correct

1 Minimal Response

One part essentially correct and either zero or one part partially correct
OR
No parts essentially correct and two parts partially correct

3. **Solution**

Part (a):

$$(£1600)\left(\frac{1.27\ euros}{£1}\right)+325\ euros = 2357\ euros$$

Part (b):

$$\text{New mean} = (£2025)\left(\frac{1.27\ euros}{£1}\right)+325\ euros = 2896.75\ euros$$

$$\text{New standard deviation} = (£475)\left(\frac{1.27\ euros}{£1}\right) = 603.25\ euros$$

Part (c):

Yes, there is at least one outlier in the salary distribution.

$IQR = Q_3 - Q_1 = 2250 - 1250 = 1000$

$1.5IQR = (1.5)(1000) = 1500$

Outliers are below $Q_1 - 1.5IQR = 1250 - 1500 = -250$ or above $Q_3 + 1.5IQR = 2250 + 1500 = 3750$. Since the maximum salary is 4500, which is greater than 3750, the maximum is an outlier.

Scoring

Part (a) is either essentially correct (E) or incorrect (I). Parts (b) and (c) are either essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is essentially correct if the correct answer is given and work is shown.
Part (a) is incorrect if no work is shown.

Part (b) is essentially correct if both values are computed correctly and work is shown.
Part (b) is partially correct if only one of the two values is correctly calculated (with supporting work).
Part (b) is incorrect if neither value is computed correctly OR if no work is shown.

Part (c) is essentially correct if the student states that the max is an outlier AND supporting work is shown.
Part (c) is partially correct if the supporting work is correct but the student did not answer the question OR if the student answers the question but the explanation is incomplete.
Part (c) is incorrect if no explanation is given for the answer.

4 Complete Response

All 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
OR
One part essentially correct and two parts partially correct

1 Minimal Response

One part essentially correct and either zero or one part partially correct
OR
No parts essentially correct and two parts partially correct