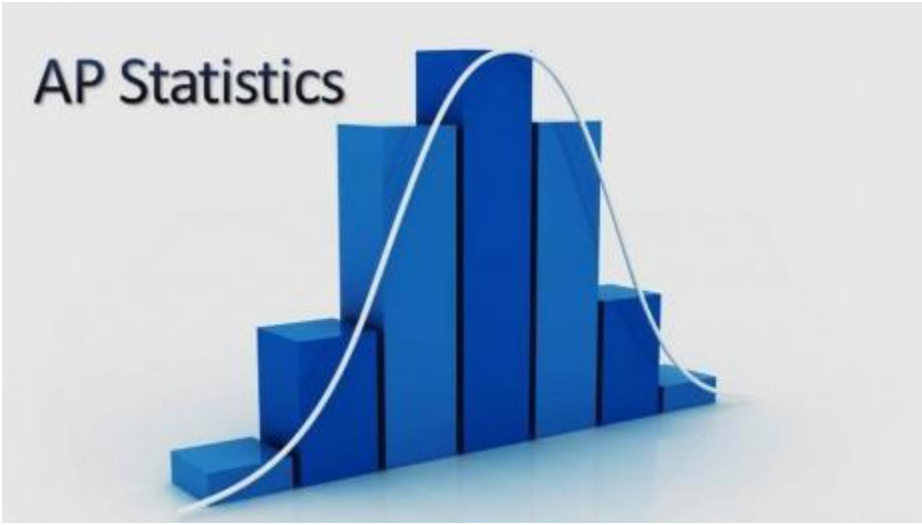
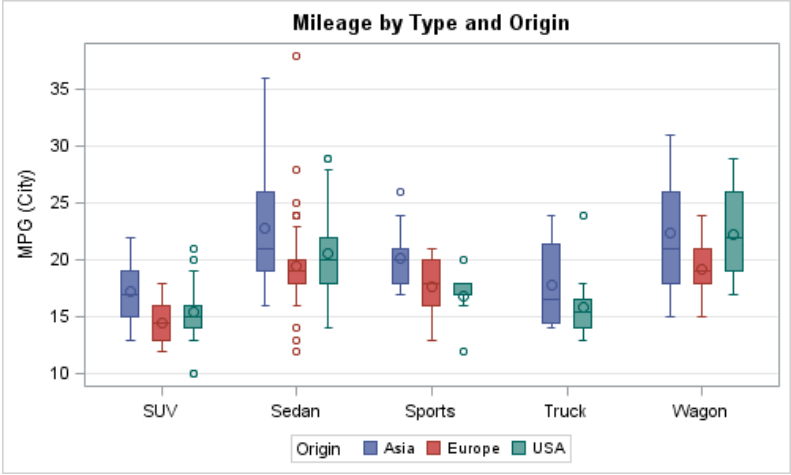
NAME:_____ DATE:_____ PERIOD:_____

REVIEW SET 2:



Exploring Data

Day Students		Evening Students
8	2	05
2	3	6778
	4	00056
98	5	233789
8876632	6	001556
876653211	7	35567
9988644200	8	2449
522110	9	39



THINGS TO REMEMBER

Types of Variables/Data

- **Categorical** – (Qualitative) – basic characteristics
  - Examples: eye color, favorite candy, whether condition (sunny, rainy, cold, etc.), grade level, zip code, area code
- **Numerical** – (Quantitative) –data where you can take the average/mean of it; usually measurements or counts (number of students)
  - **Discrete** – listable sets (counts)
  - **Continuous** – any value over an interval of values (measurements)
- **Univariate** – one variable
- **Bivariate** – two variables
- **Multivariate** – many variables

How to describe numerical graphs – C.U.S.S. & B.S.

Use for boxplots, histograms, dotplots, and stem and leaf plots.

- **Center** – middle of the data (usually the median but can be the mean)
  - **Median** – the middle point of the data (50^{th} percentile) when the data is in numerical order. If two values are present, then average them together.
  - **Mean** – μ is for a population (parameter) and $\bar{x}$ is for a sample (statistic).

- **Unusual features** –outliers, gaps, clusters, etc.
  - Can't see gaps on a boxplot
  - Don't skip numbers (stems) on stemplots so that you can see gaps

- **Shape** –
  - **Symmetrical** – data on which both sides are fairly the same shape and size. "Bell Curve"
  - **Uniform** – every class has an equal frequency (number) "a rectangle"
  - **Skewed** – one side (tail) is longer than the other side. The skewness is in the direction that the tail points (left or right)
    - **Skewed Right**
    - **Skewed Left**
  - **Bimodal** – 2 groups/humps/clusters of data

- **Spread** – refers to variability (range, standard deviation, and IQR)
  - **Range** – a single value – (Max – Min)
  - **IQR** – interquartile range – (Q3 – Q1)
  - **Standard deviation** – measures the typical or average deviation of observations from the mean
    - σ for population (parameter) & s for sample (statistic)
  - **Variance** – standard deviation squared

- **Be Specific -** *Everything must be in **context** to the data/situation of the graph.

- When **comparing two distributions** – MUST CUSS and BS and use ER words (comparative language)!
  - Bigger, larger, smaller, greater than, less than, more than, etc.

How to describe categorical graphs –L.S.S.

Use for bar graphs, segmented bar graphs, and pie charts.

- **Largest** – what's the largest percentage for each group/category
- **Smallest** – what's the smallest percentage for each group/category
- **Something that stands out** – what's something that standout about the data
- **Be Specific -** *Everything must be in **context** to the data/situation of the graph.

THINGS TO REMEMBER

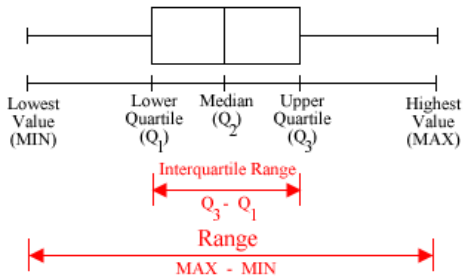
Comparison of mean & median based on graph type

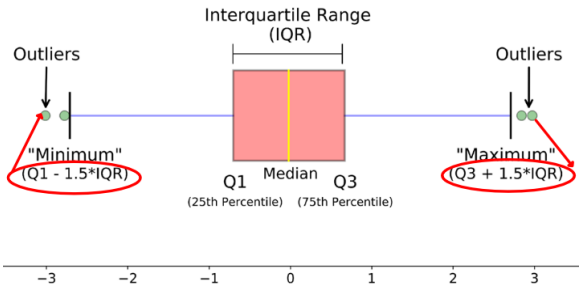
- **Symmetrical** – mean and the median are the same value.
- **Skewed Right** – mean is a larger value than the median.
- **Skewed Left** – the mean is smaller than the median.

*The mean is always pulled in the direction of the skew away from the median.

Boxplots – are for medium or large numerical data. It does not contain original observations.

- Displays the **5-Number Summary** – Minimum, Q_1, Q_2 (Median), Q_3, and Maximum





Resistant – not affected by outliers.

Resistant Values	**Non-Resistant Values**	
• Median	• Mean	• Correlation Coefficient (r)
• IQR	• Range	• Least Squares Regression Line (LSRL)
	• Variance	• Coefficient of Determination (r^2)
	• Standard Deviation	

Linear Transformations of random variables

- $\mu_{a+bx} = a + b\mu_x$ The mean is changed by **both** addition/subtract & multiplication/division.
- $\sigma_{a+bx} = |b|\sigma$ The standard deviation is changed by multiplication/division ONLY.

Combination of two (or more) random variables

- $\mu_{x\pm y} = \mu_x \pm \mu_y$ Just add or subtract the two (or more) means
- $\sigma_{x\pm y} = \sqrt{\sigma_x^2 + \sigma_y^2}$ Always add the variances – X & Y MUST be independent

Z –Score – is a standardized score. This tells you how many standard deviations from the mean an observation is.

- Formula: $z = \frac{x - \mu}{\sigma}$
- It creates a standard normal curve consisting of z-scores with a $\mu = 0$ & $\sigma = 1$.

Normal Curve – is a bell shaped and symmetrical curve.

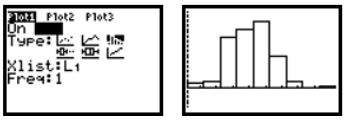
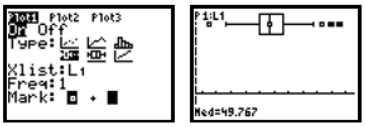
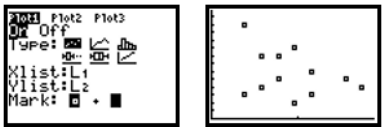
- Notation: N (μ, σ)
- As σ increases the curve flattens; the wider the curve the larger the standard deviation.
- As σ decreases the curve thins; the skinner the curve the smaller the standard deviation

Empirical Rule (68-95-99.7) measures 1 σ, 2 σ, and 3 σ on **normal curves** from a center of μ.

- 68% of the population is within 1σ of the mean
- 95% of the population is within 2σ of the mean
- 99.7% of the population is within 3σ of the mean

CALCULATOR KEY PRESSES

Function	Key Press

To Create Graphs
• **Histogram**
  ○ Put the numbers into list 1 (L1)
  ○ Choose the 3rd graph
 
• **Boxplot**
  ○ Put the numbers into list 1 (L1)
  ○ Choose the 4th graph
 
• **Scatterplots**
• Put the numbers into list 1 (L1) and list 2 (L2)
• Choose the 1st graph
 
	• Stat, Enter
	• 2nd
	• Y=
	• Select appropriate graph
	• Zoom 9 (to see graph)
	• Trace (to see values)

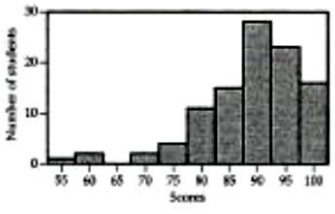
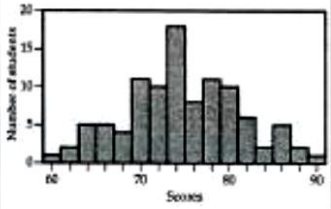
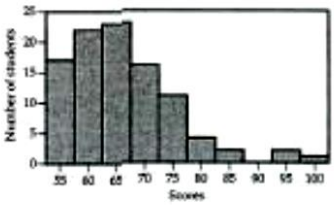
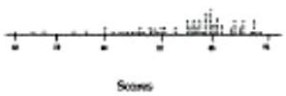
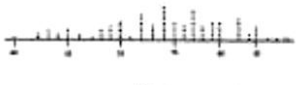
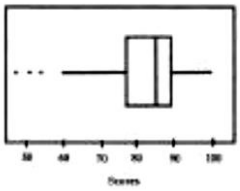
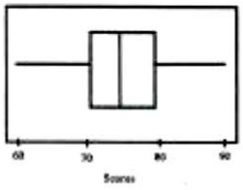
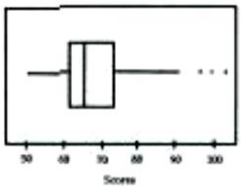
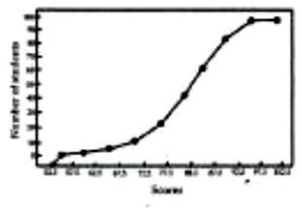
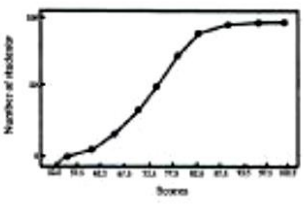
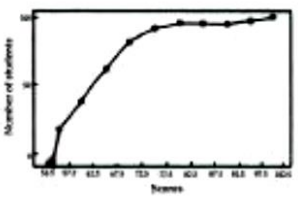
To Get Summary Statistics • Put the numbers into lists • $\bar{x}$ = mean • s_x = standard deviation • n = sample size • minX = minimum • Q_1 = 1st quartile • Med = median/2nd quartile • Q_3 = 3rd quartile • maxX = maximum	• Stat, Enter • Stat • Arrow over to Calc • 1-Var Stats • Calculate

To Get Linear Regression Equation, correlation coefficient, and coefficient of determination
• Make sure diagnostics are on
  ○ Mode
  ○ Arrow Up
  ○ Select Stat Diagnostics – On
• Equation: $\hat{y} = ax + b$
  ○ a = slope
  ○ b = y-intercept
• r^2 = coefficient of determination
• r = correlation coefficient
	• Stat, Enter
	• Stat
	• Arrow over to Calc
	• LinReg (ax + b) #4
	• Calculate

How to Describe Data

In any graph of data, look for the *overall pattern* and for any striking *deviations* from the pattern. In statistics when we describe data we **CUSS** and **BS.** You can describe the overall pattern of a distribution by its *center*, *unusual features*, *shape*, and *spread* and *being very specific.*

Comparison of Shapes of Different Graphs

Graph	LEFT-SKEWED	SYMMETRIC	RIGHT-SKEWED
Histogram			
Dotplot			
Boxplot			
Cumulative Frequency Chart			
Stemplot	4 8 5 2 5 5 6 14 6 889 7 0112233444 7 56667778899 8 00111111222222333444444 8 55566666677777777788888999999 9 222233444 9 5555567	6 0 6 22333 6 44555 6 667 6 88899999 7 0000001 7 222222222333333 7 444444444455555 7 66677777777 7 88899999 8 00000 8 222222 8 445555 8 67 8 89	5 2333344 5 55556667799 6 000000000111111111222222233333444 6 555555666666667777888899 7 0011122224444 8 003 8 7 9 2 9 59

Exploring Data Review Problems

Quantitative vs Categorical Variables

Quantitative variables are numerical values for which arithmetic operations such as means make sense. It is usually a *measure* of some sort.

Categorical variables simply count which of several categories a person or thing falls.

Examples: Are the following categorical or quantitative?

1) zip codes
2) a list of subject heights
3) the times it takes to make a button hole
4) the ages of several subjects
5) subjects listed according to gender
6) students listed by social security number
7) students listed by recent test scores

When looking at a distribution how do we describe or compare them? – CUSS and BS

C – locate the center

Mean or Median

mean – not a resistant measure – outliers really pull it toward them

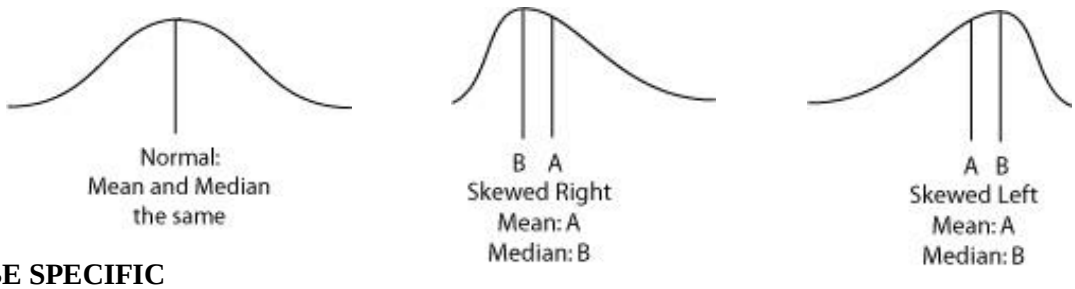
median – a very resistant measure of center – outliers have no effect

U – check for unusual features, gaps, and outliers

Outliers are more than 1.5 times the IQR above Q_3 or below Q_1

$IQR = Q_3 - Q_1$

S – examine the overall shape



BS – BE SPECIFIC

DO NOT USE IT, SHE, HE, THEY

note: when looking at a graph be sure to check the vertical axis to learn if it is a

a) frequency graph – the numbers count the data in each bar
b) relative frequency graph – the numbers represent the percent of the data in each bar
c) cumulative frequency graph – the numbers represent the TOTAL up to that point

S – Describe the spread

Standard deviation or Range or IQR

Use the following numbers: 23 36 21 40 27

1. Calculate the mean
2. Calculate the standard deviation
3. Calculate the five number summary

4. Make a boxplot of the data.

5. Use the following numbers to create a boxplot weight of a newborns in ounces: 17, 27, 36, 45, 80, 85, 50

Stem & Leaf Plots

The following data represents the weight of a toddler in pounds: 23, 36, 21, 40, 27. Create a Stem and Leaf Plot

Linear Transforming Data

Example 1: sample A: $\bar{x} = 7$ $s = 2$ and sample B: $\bar{x} = 5$ $s = 3$

6. Transform sample A with the linear equation $y = 5x - 8$

7. Find the mean and standard deviation of A–B:

Mean	Standard Deviation

8. Given $\mu_X = 15$ and $\sigma_X = 5$. Y can be described as the 2X + 4. Find the mean and standard deviation of Y.

The Normal Curve

The area under the **standard normal curve** is:

The standard deviation of the standard normal curve is:

The mean of the standard normal curve is:

What is the rule that describes 1, 2, and 3 standard deviations from the mean?

Example 2: There are 4 basic types of normal curve problems.

Students at the fine arts academy view, on average, five movies per semester with a standard deviation of two movies.

9. What proportion of the students view more than six movies per semester?

10. What proportion of the students view between 3 and 8 movies per semester?

11. What proportion of the students view less than two movies per semester?

12. What number separates the bottom 15% from the rest?

Example 3:

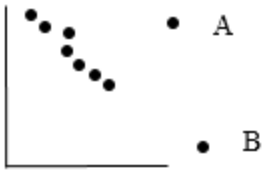
13. Test A has a mean of 79 with a standard deviation of 3. Test B has a mean of 84 with a standard deviation of 5. If Rudy made an 83 on test A and a 90 on test B, which test he did score higher on compared to the rest of the class.

Bivariate data

Explanatory variables – attempt to explain the observed outcome – the independent variable (x)

Response variable – measures the outcome of the study – the dependent variable (y) because it depends on (x)

*****study the scatter plot to the right



14. Would it be most appropriate to remove case A or case B?

15. Do the points have a positive or negative association, why?

16. What does the "least-squares regression" line mean?

17. How do you calculate the residual?

18. If a set of points has a least squares regression equation of $y = 2.3x + 17$, what is the residual of the actual point (3, 19.5)?

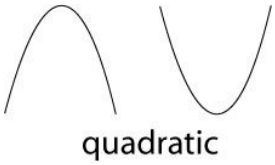
19. Would this point be above or below the linear regression line?

20. How do we describe a scatterplot!

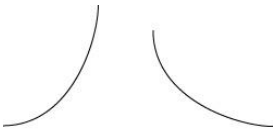
What to do with a set of quantitative bivariate data

1) make a scatter plot and look at it
2) find linear regression equation
3) make a residual plot to check linear regression equation
4) if regression equation has a pattern, try to straighten the curve using a transformation

Other patterns I'd expect on the AP test



quadratic



exponential

logarithmic

21. If the linear regression equation to find the temperature on top of Flattop Mountain based on the temperature of Denver is $y = 1.2x - 31$ where x is Denver's temperature and y is the temperature on the top of Flatttop Mountain, describe/interpret what the coefficients/constants mean.

22. For the scenario in question 21, if $r = 0.87$, what does it mean?

23. For the scenario in question 21, if $r^2 = .7569$, what does it mean?

Notes:

- Making predictions using the regression equation may not be useful outside the range of the data. Often it doesn't make sense. This is called ***extrapolation.***

- Just because two things are associated doesn't mean one causes the other. There may be a lurking variable that has an effect on both. This is called **confounding.** In order to determine causation, one must do an **experiment.**

- When it comes to **Categorical bivariate or multivariate data,** make a table and figure the totals. Percentages of either rows or columns may be useful. A chi-squared test may be useful.

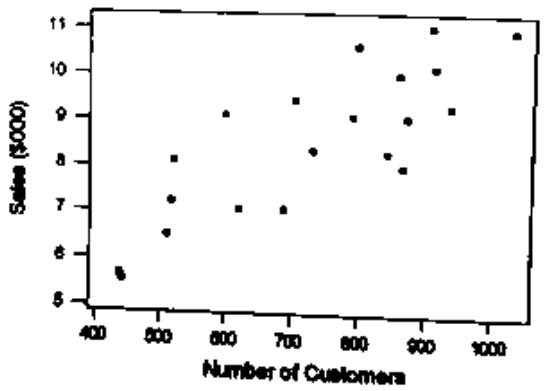
Computer Output

Regression Analysis

Number of Customers vs. Amount of Sales (in $100s)

Predictor	Coef	StDev	T	P
Constant	3.0821	0.9176	3.36	0.003
Number	0.007500	0.001224	6.13	0.000

S = 0.9610 R-Sq = 67.6% R-Sq(adj) = 65.8%



(03 Q1) Since Hill Valley High School eliminated the use of bells between classes, teachers have noticed that more students seem to be arriving to class a few minutes late. One teacher decided to collect data to determine whether the students' and teachers' watches are displaying the correct time. At exactly 12:00 noon, the teacher asked 9 randomly selected students and 9 randomly selected teachers to record the times on their watches to the nearest half minute. The ordered data showing minutes after 12:00 as positive values and minutes before 12:00 as negative values are shown in the table below.

Students	–4.5	–3.0	–0.5	0	0	0.5	0.5	1.5	5.0
Teachers	–2.0	–1.5	–1.5	–1.0	–1.0	–0.5	0	0	0.5

a) Construct parallel boxplots using these data.

b) Based on the boxplots in part (a), which of the two groups, students or teachers, tend to have watch times that are closer to the true time. Explain your choice.

c) The teacher wants to know whether individual students' watches tend to be set correctly. She proposes to test H_o: $\mu = 0$ versus H_a: $\mu \neq 0$, where μ represents the mean amount by which all student watches differ from the correct time. Is this an appropriate pair of hypothesis to test to answer the teacher's question? Explain why or why not. <u>Do not carry out the test.</u>

(07FBQ1)

The Better Business Council of a large city has concluded that students in the city's schools are not learning enough about economics to function in the modern world. These findings were based on test results from a random sample of 20 twelfth-grade students who completed a 46-question multiple-choice test on basic economic concepts. The data set below shows the number of questions that each of the 20 students in the sample answered correctly.

12	16	18	17	18	33	41	44	38	35
19	36	19	13	43	8	16	14	10	9

(a) Display these data in a stemplot.

(b) Use your stemplot from part (a) to describe the main features of this score distribution.

(c) Why would it be misleading to report only a measure of center for this score distribution?

(**08Q6)** Administrators in a large school district wanted to determine whether students who attended a new magnet school for one year achieved greater improvement in science test performance than students who did not attend the magnet school. Knowing that more parents would want to enroll their children in the magnet school than there was space available, the district decided to conduct a lottery of all families who expressed interest in participating. In their data analysis, the administrators would then compare the change in test scores of those children who were selected and those who were not selected.

Administrators were also interested in using pretest scores on this test as a predicator of posttest scores in the test. The following computer output contains the results from separate regression analyses on the magnet school scores and on the original school scores. The accompanying graph displays the data and separate regression lines for the magnet and original schools.

Regression Analysis: Post_Magnet versus Pre_Magnet

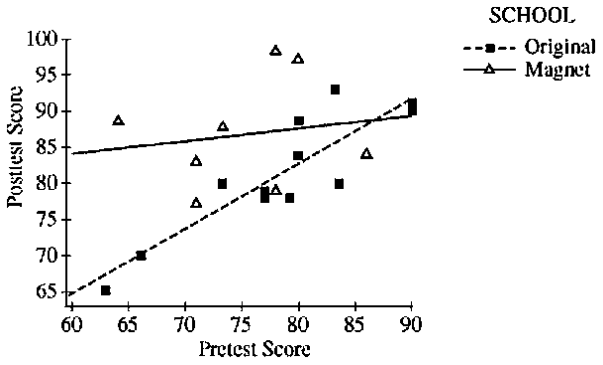
Predictor	Coef	SE Coef	T	P
Constant	73.27	34.55	2.12	0.078
Pre_Magnet	0.1811	0.4583	0.40	0.706

S = 8.20920 R-Sq = 2.5% R-Sq(adj) = 0.0%

Regression Analysis: Post_Original versus Pre_Original

Predictor	Coef	SE Coef	T	P
Constant	9.24	11.91	0.78	0.456
Pre_Original	0.9204	0.1512	6.09	0.000

S = 4.11463 R-Sq = 78.8% R-Sq(adj) = 76.6%



b)

(i) State the equation of the regression line for the magnet school and interpret its slope in context of the questions.

(ii) State the equation of the regression line for the original school and interpret its slope in context of the questions.

c) To determine whether there is a significant correlation between pretest score and posttest score, a test of the following hypotheses will be performed.

H_0: There is no correlation between pretest score and posttest score
(true slope = 0)

H_a: There is a correlation between pretest score and posttest score
(true slope $\neq$ 0)

(i) Using the regression output, state the p-value and conclusion for this test at the magnet school. Assume the conditions for inference have been met.

(ii) Using the regression output, state the p-value and conclusion for this test at the original school. Assume the conditions for inference have been met.

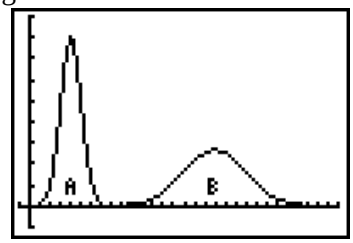
Exploring Data Review Problems

Multiple Choice Problems

Circle the letter for the statement that is the **best answer** for each multiple choice question.

1. In the display of distributions A and B, which has the larger mean and which has the larger standard deviation?

(a) Larger mean, A; larger standard deviation, A
(b) Larger mean, A; larger standard deviation, B
(c) Larger mean, B; larger standard deviation, A
(d) Larger mean, B; larger standard deviation, B
(e) Larger mean, B; same standard deviation



2. The average cost per ounce for glass cleaner is 7.7 cents with a standard deviation of 2.5 cents. What is the Z score of the glass cleaner, Windex, that costs 10.1 cents per ounce?

(a) 0.96 (b) 1.31 (c) 1.94 (d) 2.25 (e) 3.00

3. What characteristic of a distribution does standard deviation measure?

(a) shape (b) center (c) spread (d) skewness (e) frequency

4. Scores on the American College Test (ACT) are normally distributed with a mean of 18 and a standard deviation of 6. The interquartile range of the scores is approximately:

(a) 8.1 (b) 12 (c) 6 (d) 10.3 (e) 7

5. Ms. Jackson's Algebra II class had a standard deviation of 2.4 on their last test, while her statistics class had a standard deviation of 1.2 on their last test. What can be said about these two classes? (The word homogeneous means alike, consistent, and similar)

(a) The algebra class's scores are more homogeneous than the statistics class's scores.
(b) The statistics class's scores are more homogeneous than the algebra class's scores.
(c) The statistics class did less well on the test than the algebra class.
(d) The algebra class performed twice as well on their test as did the statistics class.
(e) The algebra class performed 1.2 points better on their test than did the statistics class.

6 The test grades at a large school have an approximately normal distribution with a mean of 50. What is the standard deviation of the data so that 80% of the students are within 12 points (above or below) the mean?

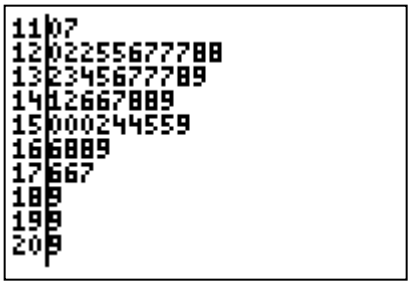
(a) 5.875 (b) 9.375 (c) 10.375
(d) 14.5 (e) cannot be determined from the given information

7. In a frequency distribution of 3000 scores, the mean is 78 and the median as 95. One would expect this distribution to be:

(a) skewed to the right (b) skewed to the left (c) bimodal
(d) symmetrical and mound-shaped (e) symmetrical and uniform

8. The stemplot displays the 1988 per capita income (in hundreds of dollars) of the 50 states. Which of the following best describes the data?

(a) Skewed distribution, mean greater than median
(b) Skewed distribution, median greater than mean
(c) Symmetric distribution, mean greater than median
(d) Symmetric distribution, median greater than mean
(e) Symmetric distribution with outliers on high end

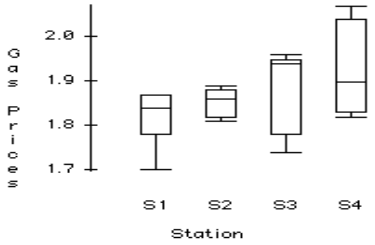


9. Which of the following are true statements?

I. The standard deviation is the square root of the variance.
II. The standard deviation is zero only when all values are the same.
III. The standard deviation is strongly affected by outliers.

(a) I and II (b) I and III (c) II and III (d) I, II, and III (e) I only (f) III only

10. A resident of Auto Town was interested in finding the cheapest gas prices at nearby gas stations. On randomly selected days over a period of one month, he recorded the gas prices (in dollars) at four gas stations near his house. The box plots of gas prices are as follows:



Which station has more consistent gas prices?

(a) Station 1 (b) Station 2 (c) Station 3 (d) Station 4 (e) Cannot be determined

11. A small kiosk at the Atlanta airport carries souvenirs in the price range of $3.99 to $29.99, with a mean price of $14.75. The airport authorities decide to increase the rent charged for a kiosk by 5 percent. To make up for the increased rent, the kiosk owner decides to increase the prices of all items by 50 cents. As a result, which of the following will happen?

(a) The mean price and the range of prices will increase by 50 cents.
(b) The mean price will remain the same, but the range of prices will increase by 50 cents.
(c) The mean price and the standard deviation of prices will increase by 50 cents.
(d) The mean price will increase by 50 cents, but the standard deviation of prices will remain the same.
(e) The mean price and the standard deviation of prices will stay the same.

12. The weights of cockroaches living in a typical college dormitory are approximately normally distributed with a mean of 80 grams and a standard deviation of 4 grams. The percentage of cockroaches weighing between 77 grams and 83 grams is about:

(a) 99.7% (b) 95% (c) 68% (d) 55% (e)34%

13. Which of the following are true statements?

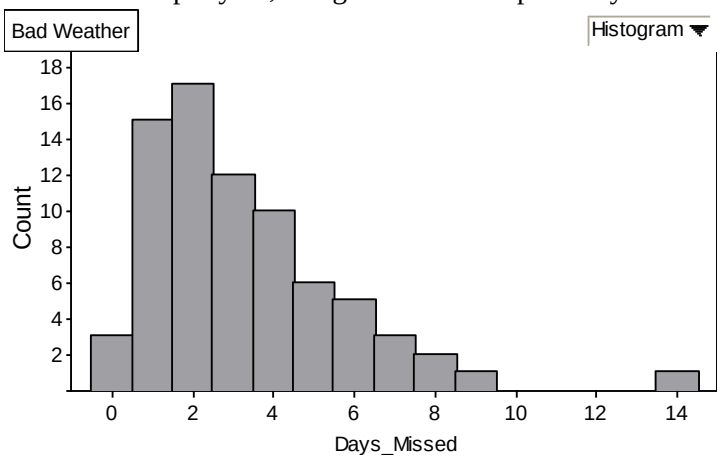
I. In all normal distributions, the mean and median are equal.

II. All bell-shaped curves are normal distributions for some value of μ and σ.

III. Virtually all the area under a normal curve is within three standard deviations of the mean, no matter what the particular mean and standard deviation are.

(a) I and II (b) I and III (c) II and III (d) I, II, and III (e) I only

14. In the northern U.S., schools are sometimes closed during winter due to severe snowstorms. At the end of the school year, schools have to make up for the days missed. The following graph shows the frequency distribution of the number of days missed due to snowstorms per year, 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 is an unbiased estimator.
(b) Median, because the distribution is skewed.
(c) IQR, because it excludes outliers and includes only the middle 50 percent of the data.
(d) First quartile, because the distribution is left skewed.
(e) Standard deviation, because it is unaffected by outliers.

15. A large company has offices in two locations, one in New Jersey and one in Utah. The mean salary of the office assistants in the New Jersey office is $28,500. The mean salary of office assistants in the Utah office is $22,500. The New Jersey office has 128 office assistants and the Utah office has 32 office assistants. What is the mean salary paid to the office assistants in this company?

(a) $22,500 (b) $23,700 (c) $25,500
(d) $27,300 (e) $28,500

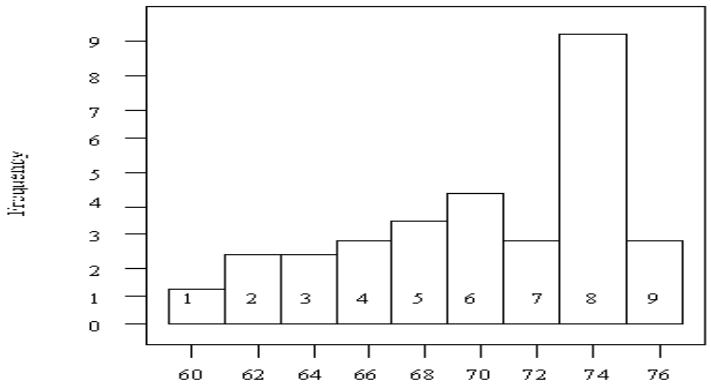
16. A distribution of 6 scores has a median of 21. If the highest score increase 3 points, what will be the value of the median?

(a) 21 (b) 21.5 (c) 24 (d) 27 (e) cannot be determined with the information given

17. A single stem-and-leaf plot is a useful tool because:

(a) It displays the mean and quartiles.
(b) It displays the percentage distribution of data values.
(c) It can display large sets of data easily.
(d) It enables one to see the overall shape of a distribution.
(e) It allows one to use any percentage to display the data.

18. Of the following, what best describes the distribution in the histogram below?



(a) Skewed to the right (b) Skewed to the left (c) Approx. Symmetric
(d) Bimodal (e) Uniform

19. In drawing a histogram, which of the following suggestions should be followed?

(a) Leave large gaps between the bins (bars). This allows room for comments.
(b) The height of bars should equal the class frequency.
(c) Generally, the bars should be square so that both the height and width equal the class column.
(d) Histograms should always have at least 15 bins.
(e) The center bar should always be the tallest.

20. The stem-plot below represents the total number of points earned by the 10 students in a statistics class.

Statistics Class Total Points

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

14 | 2 = 142 points

What is this stemplot most similar to?

(a) A histogram with class interval 110 ≤ score < 120, 120 ≤ score < 130, etc
(b) A symmetric like distribution.
(c) A boxplot distribution.
(d) A 5-point summary of the data.
(e) A bimodal distribution.

Use the following information to answer the next three questions.

Rainwater was collected in water collectors at thirty-one different sites near an industrial basin and the amount of acidity (pH level) was measured. The following stem plot shows the pH values that ranged from 2.6 to 6.3.

Rainwater pH

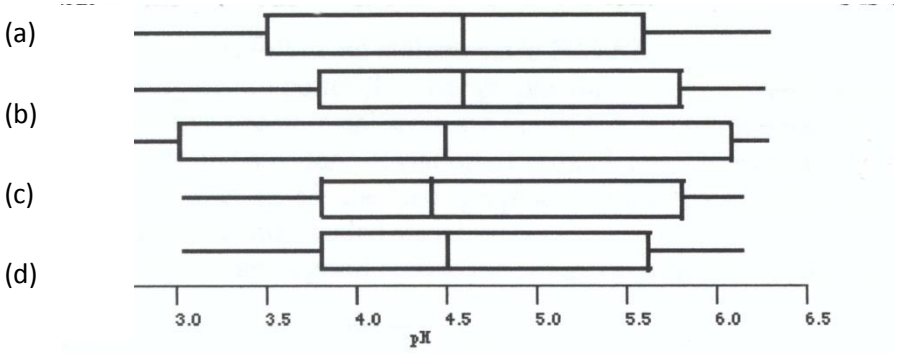
2	679
3	237789
4	1222446899
5	05567888
6	0233

4|2 = 4.2 pH

21. What is the median pH reading?

(a) 4.2 (b) 4.4 (c) 4.5 (d) 4.6 (e) Average of 55 and 56

22. Which boxplot represents the data in the stemplot?



23. What is the interquartile range?

(a) 2.0 (b) 3.7 (c) 3.8 (d) 4.5 (e) 5.6

24. The following is a stem plot of the birth weights of 26 male babies born to a smoking group of mothers.

Birth Weight of Male Babies

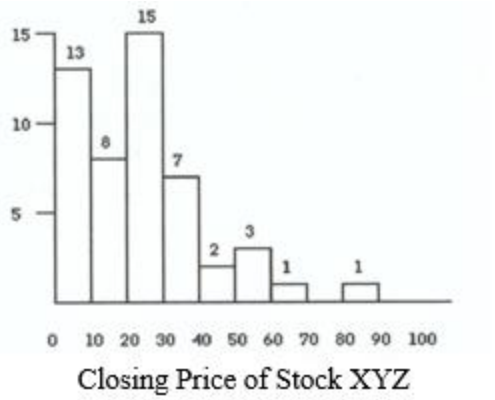
2	346778889
3	22346789
4	12234
5	3556

2 | 3 = 2.3 kg

What is the median weight, in kilograms, of the male babies in this sample?

(a) 13.5 (b) 3.2 (c) 3.5 (d) 3.7 (e) 5.524.

25. The following is a histogram showing the actual frequency of the closing prices for 50 days of trading on the New York exchange for stock XYZ.



Based on the above frequency histogram for New York Stock exchange, which class contains the 80th percentile?

(a) 10 – 20 (b) 20 – 30 (c) 30 – 40

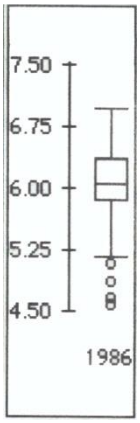
(d) 40 – 50 (e) 50 – 60

26. In the histogram above, what best describes the shape of the distribution?

(a) approximately symmetric (b) skewed to the left (c) skewed to the right

(d) skewed in both directions (e) uniform

27. Use the following output from the statistical software Data Desk when analyzing the pH values of data collected on precipitation events in 1986.



Which of the following is **not correct**?

(a) The interquartile range is about 0.34

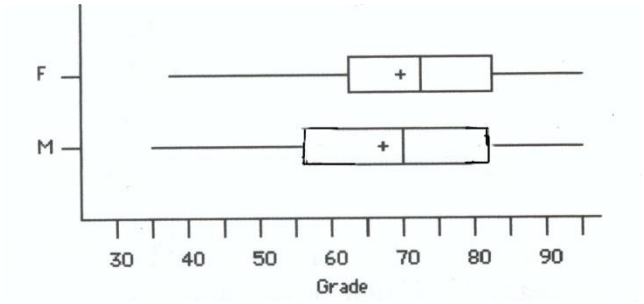
(b) The 25th percentile is about 5.9.

(c) The median is about 5.24.

(d) About 75% of the data is less than 6.4.

e) Some outliers appear to be present below a pH of 5.25.

28. Consider the following box plots of males (M) and females (F) for grades in a course in statistics. These boxplots are drawn according to the convention that the whiskers only reach to the 10th and 90th percentiles, not the minimum and maximum values. The "+" indicates the location of the mean.



Which of the following is correct?

(a) The mean grade of the female students is about 72.

(b) The median of the male students is about 60.

(c) The male IQR has more variability than the female IQR.

(d) About 25% of the female students get grades above 72.

(e) About 10% of the male students get grades below 56.

Exploring Data Review Problems
INVESTIGATIVE TASK

(10 Q6) Hurricane damage amounts, in millions of dollars per acre, were estimated from insurance records for major hurricanes for the past three decades. A stratified random sample of five locations (based on categories of distance from the coast) was selected from each of three coastal regions in the southeastern United States. The three regions were Gulf Coast (Alabama, Louisiana, Mississippi), Florida, and Lower Atlantic (Georgia, South Carolina, North Carolina). Damage amounts in millions of dollars per acre, adjusted for inflation, are shown in the table below.

HURRICANE DAMAGE AMOUNTS IN MILLIONS OF DOLLARS PER ACRE

	Distance from Coast				
	< 1 mile	1 to 2 miles	2 to 5 miles	5 to 10 miles	10 to 20 miles
Gulf Coast	24.7	21.0	12.0	7.3	1.7
Florida	35.1	31.7	20.7	6.4	3.0
Lower Atlantic	21.8	15.7	12.6	1.2	0.3

a) Sketch a graphical display that compares the hurricane damage amounts per acre for the three different coastal regions (Gulf Coast, Florida, and Lower Atlantic) and that also shows how the damage amounts vary with distance from the coast.

b) Describe differences and similarities in the hurricane damage amounts among the three regions.

Because the distributions of hurricane damage amounts are often skewed, statisticians frequently use rank values to analyze such data.

c) In the table below, the hurricane damage amounts have been replaced by the ranks 1, 2, or 3. For each of the distance categories, the highest damage amount is assigned a rank of 1 and the lowest damage amount is assigned a rank of 3. Determine the missing ranks for the 10-to-20-miles distance category and calculate the average rank for each of the three regions. Place the values in the table below.

ASSIGNED RANKS WITHIN DISTANCE CATEGORIES

	Distance from Coast					
	< 1 mile	1 to 2 miles	2 to 5 miles	5 to 10 miles	10 to 20 miles	Average Rank
Gulf Coast	2	2	3	1		
Florida	1	1	1	2		
Lower Atlantic	3	3	2	3		

d) Consider testing the following hypotheses.

H_o : There is no difference in the distributions of hurricane damage amounts among the three regions.

H_a : There is a difference in the distributions of hurricane damage amounts among the three regions.

If there is no difference in the distribution of hurricane damage amounts among the three regions (Gulf Coast, Florida, and Lower Atlantic), the expected value of the average rank for each of the three regions is 2.Therefore, the following test statistic can be used to evaluate the hypotheses above:

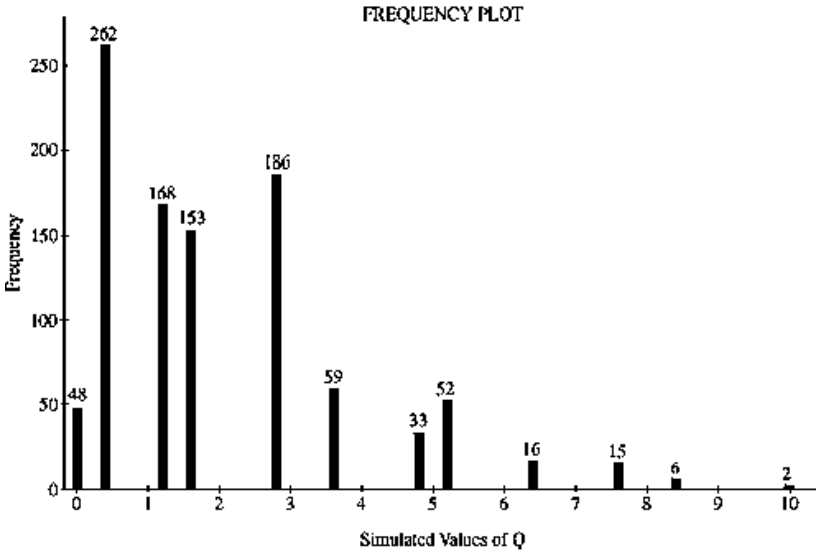
$$Q = 5\left[\left(\overline{R}_G - 2\right)^2 + \left(\overline{R}_F - 2\right)^2 + \left(\overline{R}_A - 2\right)^2\right]$$

where $\overline{R_G}$ is the average rank over the five distance categories for the Gulf Coast (and $\overline{R_F}$ and $\overline{R_A}$ are similarly defined for the Florida and Lower Atlantic coastal regions). Calculate the value of the test statistic Q using the average ranks you obtained in part (c).

e) One thousand simulated values of this test statistic, Q, were calculated, assuming no difference in the distributions of hurricane damage amounts among the three coastal regions. The results are shown in the table below. These data are also shown in the frequency plot where the heights of the lines represent the frequency of occurrence of simulated values of Q.

Frequency Table for Simulated Values of Q

Q	Frequency	Cumulative Frequency	Percent	Cumulative Percent
0.0	48	48	4.80	4.80
0.4	262	310	26.20	31.00
1.2	168	478	16.80	47.80
1.6	153	631	15.30	63.10
2.8	186	817	18.60	81.70
3.6	59	876	5.90	87.60
4.8	33	909	3.30	90.90
5.2	52	961	5.20	96.10
6.4	16	977	1.60	97.70
7.6	15	992	1.50	99.20
8.4	6	998	0.60	99.80
10.0	2	1000	0.20	100.00



Use these simulated values and the test statistic you calculated in part (d) to determine if the observed data provide evidence of a significant difference in the distributions of hurricane damage amounts among the three coastal regions. Explain.