AP Final Review II – Exploring Data (20%–30%)

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?

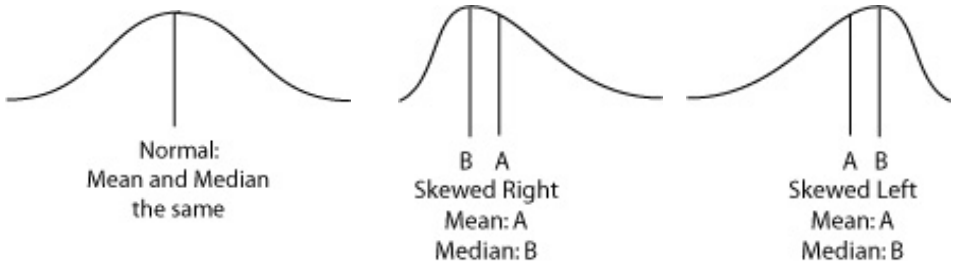
1) zip codes
2) a list of subject heights
3) the times it takes to make a button hole
4) the ages of several subjects
5) subjects listed according to gender
6) students listed by social security number
7) students listed by recent test scores

When looking at a distribution – CUSS and BS

1) locate the center
2) examine the overall shape
3) check for gaps and outliers
4) describe the spread

1) mean – not a resistant measure – outliers really pull it toward them
   median – a very resistant measure of center – outliers have no effect

2)



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

Note – outliers are more than 1.5 times the IQR above Q_3 or below Q_1

Make a histogram and compare the shapes – (looks more skewed on the histogram)

Standard deviation – use AP formula

Use the following numbers: 23 36 21 40 27

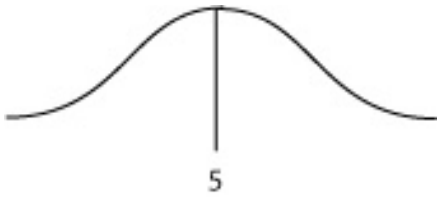
a) First find the mean: 29.4
b) subtract the mean from each number then square it: 40.96 43.56 70.56 112.36 5.76
c) Find the sum of the squared deviation scores: 273.2
d) Divide the sum by n–1: 68.3
e) Take the square root for the standard deviation: 8.234 (NOTE:calculator gives 8.264 due to rounding error on our part)

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

Transform sample A with the linear equation $y = 5x - 8$
{New mean = 5(7) – 8 = 27 new standard deviation = 5(2) = 10}

Find the mean and standard deviation of A–B: $\bar{x}_{A-B} = 7-5 = 2$; $s_{A-B} = \sqrt{2^2 + 3^2} = 3.606$

The Normal Curve
The area under the **standard normal curve** is: (1)
The standard deviation of the standard normal curve is: (1)
The mean of the standard normal curve is: (0)



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.

1) What proportion of the students view more than six movies per semester?

2) What proportion of the students view between 3 and 8 movies per semester?

3) What proportion of the students view less than two movies per semester?

4) What number separates the bottom 15% from the rest?

Example 3: 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.

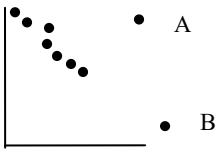
Find the z-score for each test.
Test A: z = 1.333 Test B: z = 1.2 Rudy did better on test A compared to the other students.

REMEMBER: on the standard normal chart, the z-score represents the number of standard deviations from the mean. All the 4-digit decimals in the middle represent the area to the left of z under the curve.

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 what x is

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



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

(case A because it is an outlier – B is an influential point)

Do the points have a positive or negative association, why? (negative, slope is negative)

What does the "least-squares regression" line mean? (the sum of the squared residuals is the smallest possible)

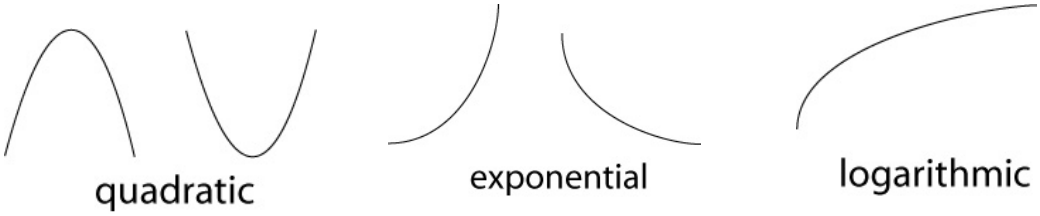
If a set of points has a least squares regression equation of $y = 2.3x + 17$, what is the residual of the actual point (3,19.5)? residual = observed – predicted = 19.5 – 23.9 = –4.4

Would this point be above or below the linear regression line? (below)

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



Example 4: 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 what the coefficients mean.

1.2 means that for every degree the temperature goes up in Denver, the temperature rises 1.2 degrees on top of the mountain.

–31 means that when the temperature in Denver is 0, the temperature on the top of Flattop Mountain is 31 degrees below 0 or –31, on average.

If $r = 0.87$, what does it mean. (There is a positive linear correlation between the two temperatures.)

If $r^2 = .7569$, what does it mean? (75.69% of the total variation in temperature on top of Flattop Mountain is associated with the variation in temperature in Denver.)

Note: 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.*

Note: 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.**

Categorical bivariate or multivariate data
Make a table and figure the totals. Percentages of either rows or columns my be useful.
A chi-squared test may be useful.

'03 #1

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

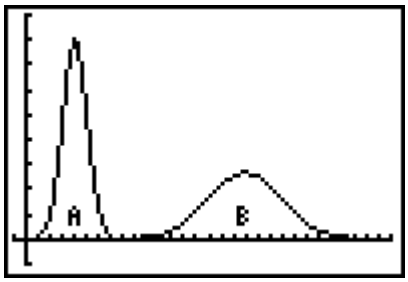
a) Construct parallel boxplots using these data.

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. Do not carry out the test.

Circle the letter for the statement that is the **best answer** for each multiple choice question and then write the letter in the margin to the left of your paper. Remember to do both.

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, 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.

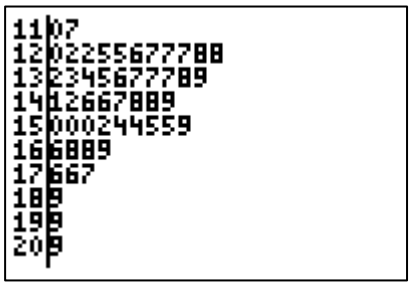
6 The test grades at a large school have an approximately normal distribution with a mean of 50. What is the standard deviation of the data so that 80% of the students are within 12 points (above or below) the mean?

(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?



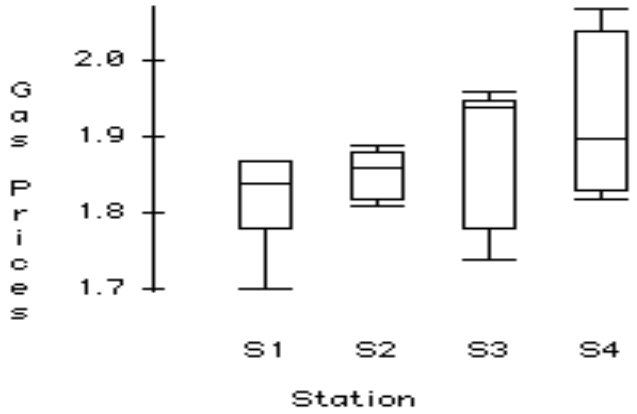
(a) Skewed distribution, mean greater than median
(b) Skewed distribution, median greater than mean
(c) Symmetric distribution, mean greater than median
(d) Symmetric distribution, median greater than mean
(e) Symmetric distribution with outliers on high end

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:

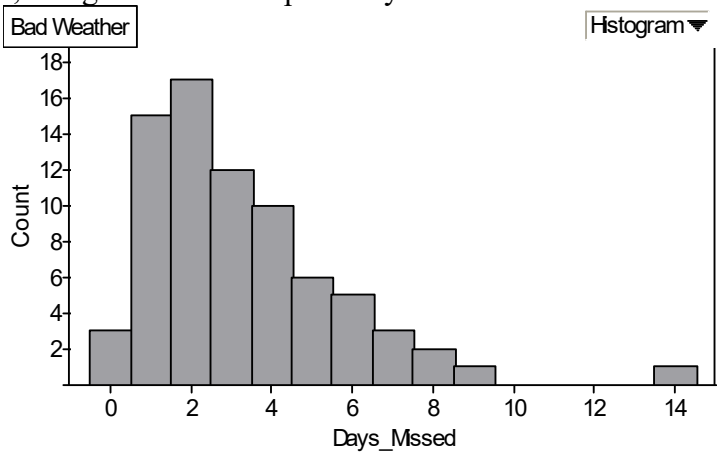
(a) 99.7% (b) 95% (c) 68% (d) 55% (e) 34%

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

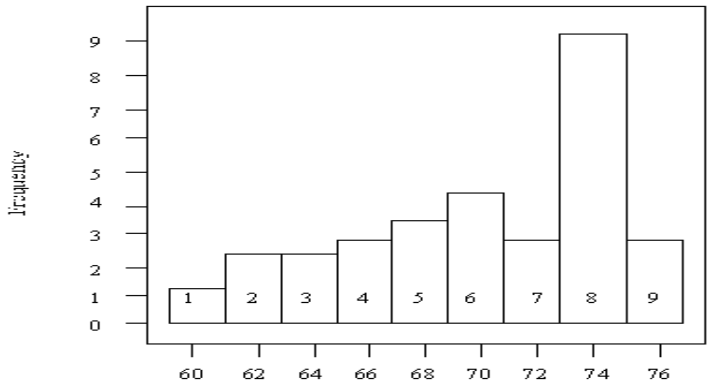
16. A distribution of 6 scores has a median of 21. If the highest score increase 3 points, what will be the value of the median?

(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) 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.

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

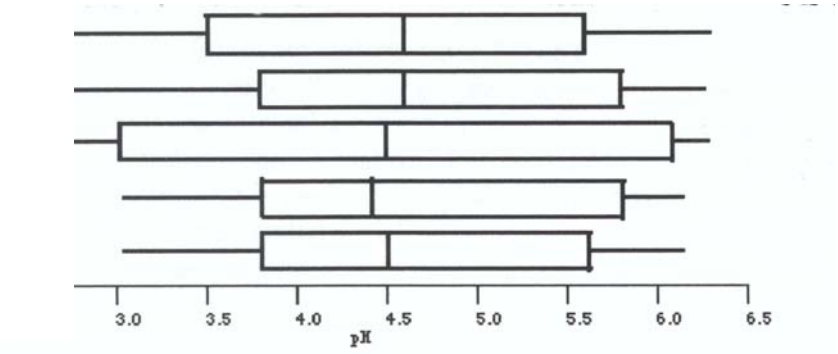
4|2 = 4.2 pH

20. What is the median pH reading?

(a) 4.2 (b) 4.4 (c) 4.5 (d) 4.6 (e) Average of 15 and 16

21. Which boxplot represents the data in the stemplot?

(a)
(b)
(c)
(d)
(e)



22. What is the interquartile range?

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

23. 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.

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

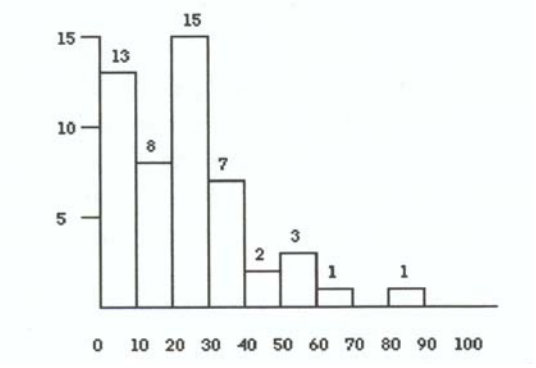
Statistics Class Total Points

```
11 | 6 8
12 | 1 4 8
13 | 3 7
14 | 2 6          14|2 = 142 points
15 |
16 |
17 | 9
```

What is this stemplot most similar to?

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

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.



Closing Price of 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

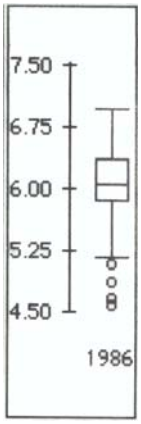
26. In the histogram above, what best describes the shape of the distribution?

(a) symmetric like (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

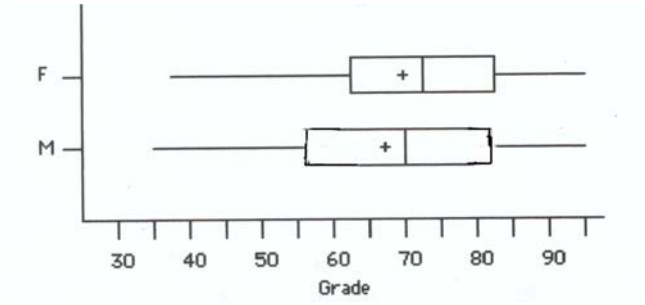
(b) The 25th percentile is about 5.9.

(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.