AP Statistics

Review #1
Exploring Data

Name _____

AP Statistics **Exploring Data Review**

I. Exploring Data: Describing patterns and departures from patterns (20%-30%)
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.

A. 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

B. 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

C. 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

D. 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

E. 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

Tips and Hints:

- When you analyze one-variable data, always discuss shape, center, spread, and unusual characteristics. (Don't forget the context!!)
- Look for patterns in the data first, and then look for deviations from those patterns.
- When commenting on shape:
  * Symmetric is not the same as "equally" or "uniformly" distributed.
  * Do not say that a distribution "is normal" just because it looks unimodal and symmetric.
- Treat the word "normal" as a "four-letter word." You should only use it if you are really sure that it's appropriate in the given situation.
- Do not confuse median and mean. They are both measures of center, but for a given data set, they may differ by a considerable amount.
  * If distribution is positively (right) skewed, then mean is likely greater than median, however, Mean > median is not sufficient to show that a distribution is positively skewed.
  * If distribution is negatively (left) skewed, then mean is likely less than median, however, Mean < median is not sufficient to show that a distribution is skewed left.
- Don't confuse standard deviation and variance. Remember that standard deviation units are the same as the data units, while variance is measured in square units.
- Remember that the variance is the standard deviation squared and the standard deviation is the square root of the variance.
- When s=0, all of the data points are the same value.

- Know how transformations of a data set affect summary statistics..
  (a) Adding (or subtracting) the same positive number k, to (from) each element in a data set increases (decreases) the mean and median by k. The standard deviation and IQR do not change.
  (b) Multiplying all numbers in a data set by a constant k multiplies the mean, median, IQR, and standard deviation by k. For instance, if you multiply all members of a data set by four, then the new set has a standard deviation that is four times larger than that of the original data set, but a variance that is 16 times the original variance.
- Be able to create a histogram, boxplot, scatterplot, etc by hand, as well as by the TI.
- When testing for outliers, remember that typically the rule we use is that outliers are values that are beyond the "fences" of 1.5(IQR) beyond the box on either side.
- When comparing graphs, make sure you use comparative language such as higher, lower, more, less, etc. Listing the attributes of center, spread, etc will only earn you a partial score at best!!

When describing a scatterplot:
* Comment on strength, outliers, form (linear, curvilinear), and direction (association) of the relationship.
* Look for patterns in the data, and then for deviations from those patterns.

- A correlation coefficient near 0 doesn't necessarily mean there are no meaningful relationships between the two variables – just not a linear relationship!

Don't confuse correlation coefficient and slope of least-squares regression line.
* A slope close to 1 or -1 doesn't mean strong correlation.
* An r value close to 1 or -1 doesn't mean the slope of the linear regression line is close to 1 or -1.
* The relationship between β_1 (slope of regression line) and r (coefficient of correlation) is $\beta_1 = r\frac{s_y}{s_x}$ ….

  This is on the formula sheet provided with the exam.
* Remember that $r^2 > 0$ doesn't mean $r > 0$. For instance, if $r^2 = 0.81$, then $r = 0.9$ or $r = -0.9$.

- Know where to find the slope and y-intercept formulas on the formula sheet!!!
- You should know difference between a scatter plot and a residual plot. For a residual plot, be sure to comment on:
  * Whether the residuals appear to be randomly distributed
  * A good residual plot has NO pattern!!
  * A curved pattern indicates a non-linear relationship

Given a least squares regression line, you should be able to correctly interpret the slope and y-intercept in the context of the problem.
* Slope = change in the predicted y-variable as x-variable increases by APPROXIMATELY 1 unit
* Y-intercept = predicted value of the response variable when predictor = 0 (Not always useful!)

- Remember properties of the least-squares regression line:
  * Contains the point $(\bar{x}, \bar{y})$, where $\bar{x}$ is the mean of the x-values and $\bar{y}$ is the mean of the y-values.
  * Minimizes the sum of the squared residuals (vertical deviations from the LSRL)
- If working with bivariate data, make sure to turn DiagnositicOn in the 2nd: Catalog menu so that your r and r^2 will show up! (You should only have to do this once)
- Residual = (actual y-value of data point) - (predicted y-value for that point from the LSRL)
- A residual plot can give you a feel for if the line is a good fit. By plotting the x-variable against the residuals, you can determine if the linear regression is appropriate. Anytime you do a LinReg, a residual list is automatically created and stored in the RESID list found under 2nd:Stat. You can create a residual plot by going to StatPlot and choosing a scatterplot with your x-list and the RESID list as the y-list. Zoom-9 to see the graph.
- If the line is not a good fit, you have to transform the data and try again. By performing various mathematical operations on your list and running your LinReg on the transformed data and looking at the residual plots, you can find a better fit.

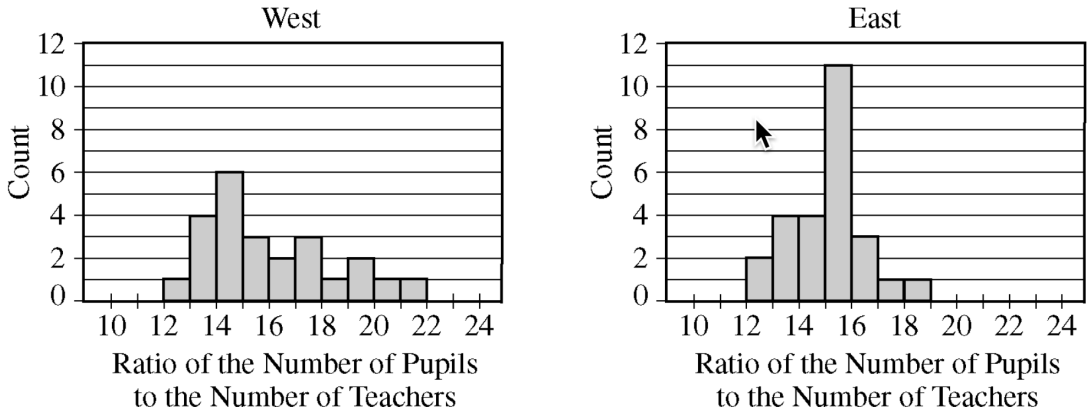
  The typical transformations we do are:
  * Logging the x-variable only -> logarithmic regression
  * Logging the y-variable only -> exponential regression
  * Logging both the x and y variables -> power regression
-

- Remember that z-scores are the number of standard deviations a value is from the mean.
- A z-score can always be found, but can only be converted to percentiles if we know the distribution is a Normal model.
- Be able to use your Normal tables to find probabilities.
- Be able to read an ogive (a cumulative frequency graph) to find percentiles.
- Some values are more resistant to the effects of outliers than others. When dealing with outliers, be careful of using values like the mean, standard deviation, and correlation, as they are all influenced by outliers.
- The properties of r: always a number between -1 and +1; the sign tells the direction; r is not affected by which variable is x or y or by changing units (like cm to inches); r only applies to linear relationships!!!
- Squaring the correlation can give you a better feel for the relationship. R2 tells the percent of the variation in the y-variable that can be explained by the regression between x and y.
- Be able to calculate the LSRL using summary statistics (the slope: b=r(Sy/Sx) and the y-intercept: a=y-bx) AND how to find the LSRL using LinReg (Stat:Calc)

2011B #1

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

1. Records are kept by each state in the United States on the number of pupils enrolled in public schools and the number of teachers employed by public schools for each school year. From these records, the ratio of the number of pupils to the number of teachers (P-T ratio) can be calculated for each state. The histograms below show the P-T ratio for every state during the 2001–2002 school year. The histogram on the left displays the ratios for the 24 states that are west of the Mississippi River, and the histogram on the right displays the ratios for the 26 states that are east of the Mississippi River.



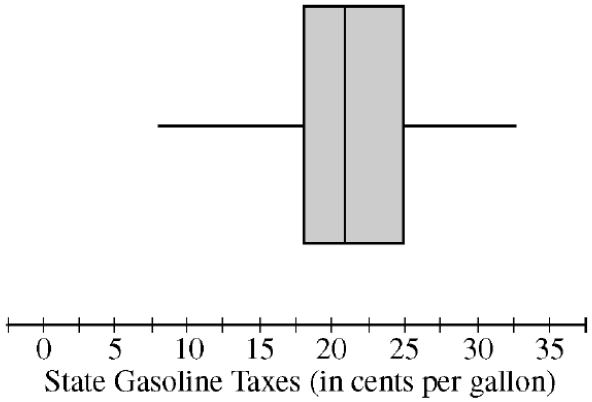
(a) Describe how you would use the histograms to estimate the median P-T ratio for each group (west and east) of states. Then use this procedure to estimate the median of the west group and the median of the east group.

(b) Write a few sentences comparing the distributions of P-T ratios for states in the two groups (west and east) during the 2001–2002 school year.

(c) Using your answers in parts (a) and (b), explain how you think the mean P-T ratio during the 2001–2002 school year will compare for the two groups (west and east).

2009B #1

1. As gasoline prices have increased in recent years, many drivers have expressed concern about the taxes they pay on gasoline for their cars. In the United States, gasoline taxes are imposed by both the federal government and by individual states. The boxplot below shows the distribution of the state gasoline taxes, in cents per gallon, for all 50 states on January 1, 2006.



(a) Based on the boxplot, what are the approximate values of the median and the interquartile range of the distribution of state gasoline taxes, in cents per gallon? Mark and label the boxplot to indicate how you found the approximated values.

(b) The federal tax imposed on gasoline was 18.4 cents per gallon at the time the state taxes were in effect. The federal gasoline tax was added to the state gasoline tax for each state to create a new distribution of combined gasoline taxes. What are approximate values, in cents per gallon, of the median and interquartile range of the new distribution of combined gasoline taxes? Justify your answer.

2001 #1

1. The summary statistics for the number of inches of rainfall in Los Angeles for 117 years, beginning in 1877, are shown below.

N	MEAN	MEDIAN	TRMEAN	STDEV	SE MEAN
117	14.941	13.070	14.416	6.747	0.624

MIN	MAX	Q1	Q3
4.850	38.180	9.680	19.250

(a) Describe a procedure that uses these summary statistics to determine whether there are outliers.

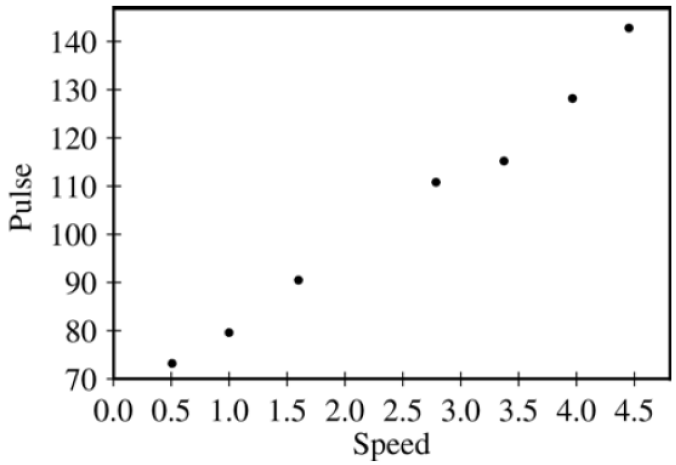
(b) Are there outliers in these data? _____

Justify your answer based on the procedure that you described in part (a).

(c) The news media reported that in a particular year, there were only 10 inches of rainfall. Use the information provided to comment on this reported statement.

2005B #5

5. John believes that as he increases his walking speed, his pulse rate will increase. He wants to model this relationship. John records his pulse rate, in beats per minute (bpm), while walking at each of seven different speeds, in miles per hour (mph). A scatterplot and regression output are shown below.



Regression Analysis: Pulse Versus Speed

Predictor	Coef	SE Coef	T	P
Constant	63.457	2.387	26.58	0.000
Speed	16.2809	0.8192	19.88	0.000

S = 3.087 R-Sq = 98.7% R-Sq (adj) = 98.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	3763.2	3763.2	396.13	0.000
Residual	5	47.6	9.5		
Total	6	3810.9			

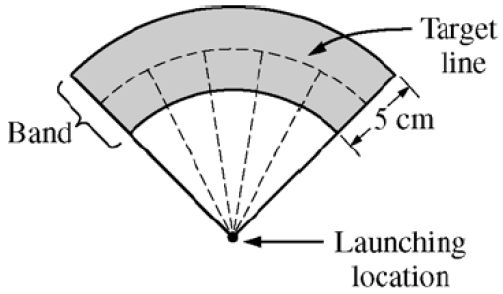
(a) Using the regression output, write the equation of the fitted regression line.

(b) Do your estimates of the slope and intercept parameters have meaningful interpretations in the context of this question? If so, provide interpretations in this context. If not, explain why not.

(c) John wants to provide a 98 percent confidence interval for the slope parameter in his final report. Compute the margin of error that John should use. Assume that conditions for inference are satisfied.

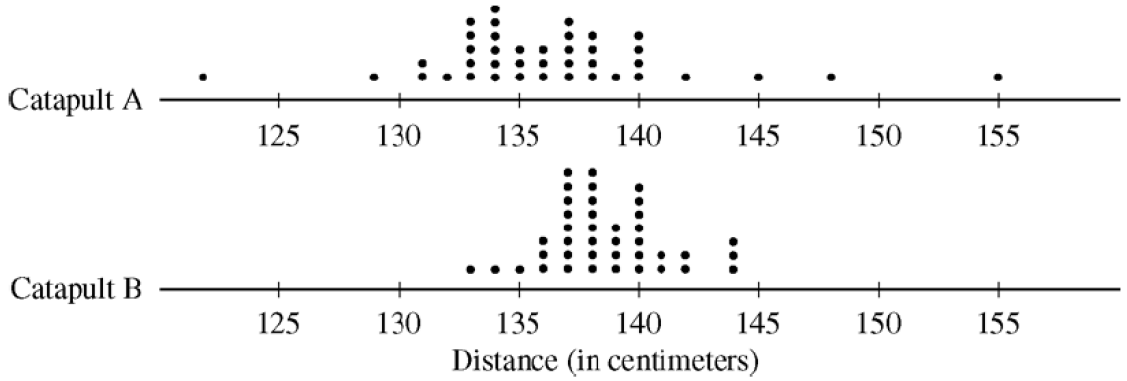
2006 #1

1. Two parents have each built a toy catapult for use in a game at an elementary school fair. To play the game, students will attempt to launch Ping-Pong balls from the catapults so that the balls land within a 5-centimeter band. A target line will be drawn through the middle of the band, as shown in the figure below. All points on the target line are equidistant from the launching location.



If a ball lands within the shaded band, the student will win a prize.

The parents have constructed the two catapults according to slightly different plans. They want to test these catapults before building additional ones. Under identical conditions, the parents launch 40 Ping-Pong balls from each catapult and measure the distance that the ball travels before landing. Distances to the nearest centimeter are graphed in the dotplots below.



(a) Comment on any similarities and any differences in the two distributions of distances traveled by balls launched from catapult A and catapult B.

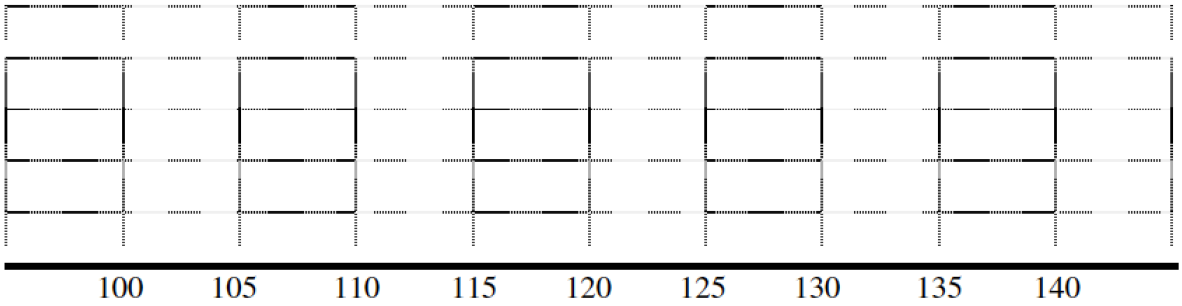
(b) If the parents want to maximize the probability of having the Ping-Pong balls land within the band, which one of the two catapults, A or B, would be better to use than the other? Justify your choice.

(c) Using the catapult that you chose in part (b), how many centimeters from the target line should this catapult be placed? Explain why you chose this distance.

5. 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 Above 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.

(b) Is it reasonable to use the sample of mandible lengths of modern Thai dogs to construct an interval estimate of the mean mandible length for the population of modern Thai dogs? Justify your answer. (Note: You do not have to compute the interval.)

(c) Is it reasonable to use the sample data of mandible lengths of modern Thai dogs and the sample data of mandible lengths of golden jackals to perform a two-sample *t*-test for the difference in mean mandible lengths for the two types of dogs? Justify your answer. (Note: You do not have to conduct the test.)