***Curriculum Topic I – DATA ANALYSIS***

***Chapters 1-4***

**Exploratory analysis of data makes use of graphical and numerical techniques to study patterns and departures from patterns.**

This topic in the curriculum expects students to be able to effectively describe/compare data sets, including **univariate data** (a distribution of numerical values), **bivariate data** (paired numerical values) or **categorical data** (groupings by characteristic).

In examining distributions of data, students should be able to detect important characteristics, such as shape, location, variability, and unusual values. From careful observations of patterns in data, students can generate conjectures about relationships among variables. The notion of how one variable may be associated with another permeates almost all of statistics, from simple comparisons of proportions through linear regression. The difference between association and causation must accompany this conceptual development throughout.

**Summary of important ideas**

**Univariate data: A distribution of numerical values**

Students should be able to both construct and interpret graphical displays of a distribution of values. There are five types of plots that students are expected to be able to use:

* dotplot,
* stemplot,
* histogram,
* boxplot, and
* cumulative frequency plot.

Students should be able to both construct and interpret numerical summaries of a distribution of values. There are several descriptive measures that students should understand:

* Measures of center: median, mean
* Measures of spread: range, interquartile range, standard deviation
* Measures of position: quartiles, percentiles, standardized scores (z-scores)

From either a plot or a numerical summary, the student is expected to be able to comment on the distribution’s

* shape (symmetric vs. skewed, bimodal, uniform, clusters, gaps, plausibly normal),
* center (mean, median),
* spread,
* unusual values

Students should be able to compare the distributions of two different univariate data sets, addressing the issues of shape, center, spread, and unusual values.

**Bivariate data: A set of pairs of numerical values**

Students should be able to both construct and interpret scatterplots created from a set of paired numerical values, making a reasonable choice for the response variable (*y*) and explanatory variable (*x*). For a given set of data, students should be able to use their calculator to determine the least squares regression line. They should be able to interpret the output of computer software that has been used to calculate the least squares regression line.

In the descriptive phase of the students’ understanding of bivariate data, they should be able to

* comment on direction, form and amount of scatter in a scatterplot,
* determine a least squares regression line for bivariate data,
* comment on the fit of the least squares regression line, and using a residual plot and numerical measures of fit,
* identify and describe outliers from the pattern and influential points, and
* make a prediction for the response variable for a given value of the explanatory variable.

In a scatterplot that shows a relationship that is not linear, students should be able to

* Apply transformations to either one or both of the data values to try to achieve a linear relationship.
* Apply a logarithmic transformation to the response variable, to determine if an exponential model is appropriate.
* Apply a logarithmic transformation to both the response and explanatory variable to determine if a power model is appropriate.
* Students should be able to determine predicted values from a non-linear model.

**Categorical data**

Students should be able to make and interpret frequency tables that summarize categorical data. They should be able to determine marginal and conditional frequencies in two-way tables, and use these to identify association between two categorical variables. They should be familiar with graphical displays that illustrate the distribution of categorical data, including bar graphs, pie graphs, and segmented bar charts.

**Exploring Data**   
  
When you analyze one-variable data, always discuss shape, center, and spread.   
  
Look for patterns in the data, and then for deviations from those patterns.   
  
**Don't confuse *median* and *mean*.** They are both measures of center, but for a given data set, they may differ by a considerable amount.

(a) If distribution is skewed right, then mean is greater than median.

(b) If distribution is skewed left, then mean is less than median.

Mean > median is not sufficient to show that a distribution is skewed right.

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.   
  
**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.   
Simple examples:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | Original data set | Mean | St. Dev. | Variance | Median | IQR | Range | | {1,2,3,4,5} | 3 | 1.414 | 2 | 3 | 3 | 4 | |

Add 7 to each element of the original data set:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | New data set | Mean | St. Dev. | Variance | Median | IQR | Range | | {8,9,10,11,12} | 10 | 1.414 | 2 | 10 | 3 | 4 | |

Multiply each element of the original data set by four:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | New data set | Mean | St. Dev. | Variance | Median | IQR | Range | | {4,8,12,16,20} | 12 | 5.6569 | 31 | 12 | 12 | 16 | |

Multiply elements of the original data set by four, then add seven:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | New data set | Mean | St. Dev. | Variance | Median | IQR | Range | | {11,15,19,23,27} | 19 | 5.6569 | 32 | 19 | 12 | 16 | |

**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 symmetric and unimodal.

**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.   
  
**When describing a scatterplot:**

* Comment on the direction, shape, and strength 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.** Consider the following data points:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | X | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | | Y | 6 | 30 | 8 | 50 | 10 | 70 | 12 | 90 | 14 | 110 | 16 | |

In this case, *r* = .38, indicating fairly weak correlation, but a scatterplot displays something quite interesting. Moral of the story: Always plot your data.   
  
**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 *b* (slope of regression line) and *r* (coefficient of correlation) is

  
This is on the formula sheet provided with the exam.

* Remember that r2 > 0 doesn't mean r > 0. For instance, if r2 = 0.81, then r = 0.9 or r = -0.9.

**You should know difference between a scatter plot and a residual plot.**   
  
**For a residual plot, be sure to comment on:**

* The balance of positive and negative residuals
* The size of the residuals relative to the corresponding y-values
* Whether the residuals appear to be randomly distributed

**Given a least squares regression line, you should be able to correctly interpret the slope and y-intercept in the context of the problem.**   
  
**Remember properties of the least-squares regression line:**

* Contains the point , where is the mean of the *x*-values and is the mean of the *y*-values.
* Minimizes the sum of the squared residuals (vertical deviations from the LSRL)

**Residual** = (**actual y-value** of data point) - (**predicted y-value** for that point from the LSRL)   
  
**Realize that logarithmic transformations can be practical and useful.** Taking logs cuts down the magnitude of numbers. Also, if there is an exponential relationship between *x* and *y* (*y=abx*), then a scatterplot of the points {(*x,log y*)} has a linear pattern.   
**Example:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | *x* | *y* | *log y* | | 1 | 24 | 1.3802 | | 2 | 192 | 2.2833 | | 3 | 1,536 | 3.1864 | | 4 | 12,188 | 4.0859 | | 7 | 6,290,000 | 6.7987 | | 8 | 49,900,000 | 7.6981 | |

An **exponential fit** to (x,y) on the TI-83 yields y = 3.002(7.993x), with *r* = 0.9999. When *x* = 9, this model predicts y = 399,901,449.2.   
  
A **linear fit** to (*x,log y*) on the TI-83 yields *log y* = 0.477395 + 0.9027286*x*, with *r* = .9999. If *x* = 9, then log y = 0.477395 + 0.9027286(9) = 8.601952978. Hence y = 108.601952978 = 399,901,449.2.   
  
If the relationship between *x* and *y* is described by a power function (*y=axb*), then a scatterplot of (log *x*, log *y*) will have a linear pattern.   
  
**Example:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | | x | y | log x | log y | | 1 | 8 | 0 | .90309 | | 2 | 64 | .30103 | 1.8062 | | 3 | 216 | .47712 | 2.3345 | | 4 | 512 | .60206 | 2.7093 | | 7 | 2744 | .8451 | 3.4384 | | 8 | 4096 | .90309 | 3.6124 | |

A **power fit** to (*x,y*) on the TI-83 yields *y*=8*x*³ with *r*=1. When *x*=9, this model predicts *y*=8(9)³ =5832.  
  
  
A **linear fit** to (log *x*, log *y*) on the TI-83 yields log *y* = .90309 + 3 log *x* with *r* = 1. When *x* = 9, this model predicts log *y* = .90309 + 3 log(9) = 3.76582  
Hence, *y* = 103.76582 = 5832.