Descriptive Statistics Refresher

MFT Topics discussed in these review slides are in bold:

1. Measure of set operations

2. Conditional/joint probabilities

3. Counting rules

4. Measures of central tendency and dispersion

5. Distributions (including normal and binomial)

6. Sampling and estimation

7. Hypothesis testing

8. Correlation and regression

9. Time-series forecasting

10. Statistical concepts in quality control

“SOCS”

When you analyze a set of data, remember “SOCS”

Shape – what is the shape of the distribution?

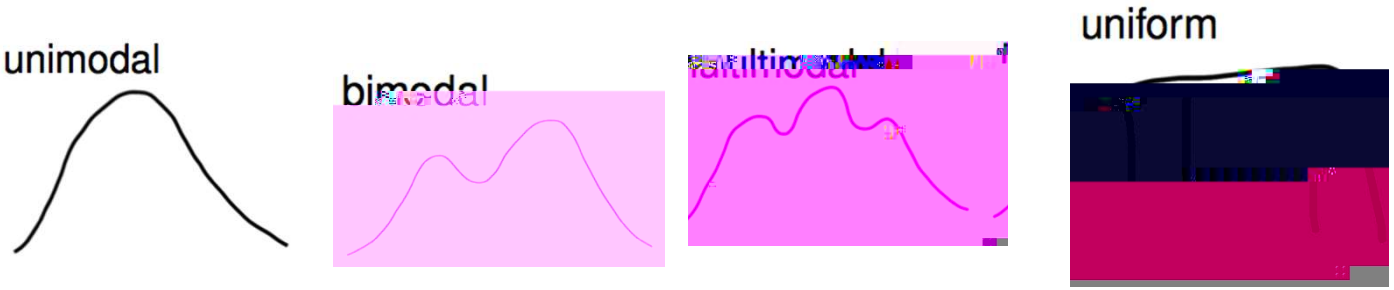
Outliers – Are there any unusual data values in the distribution?

Center – How can I best describe the typical value or center of the distribution?

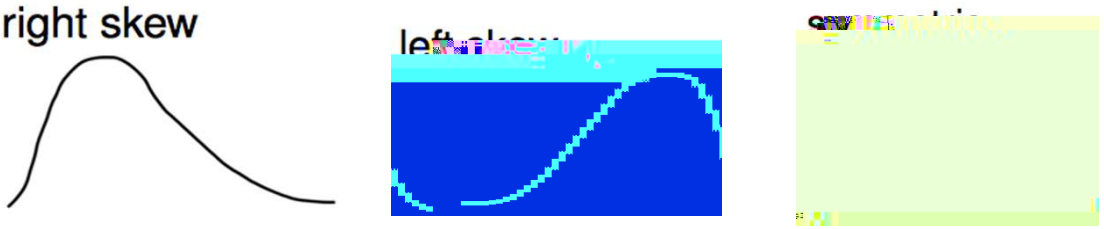
Spread – How can I describe the variation or dispersion in the data?

Commonly observed shapes

Modality:



Skewness or symmetry:



Histogram

Common graph for quantitative data. The horizontal axis is a number line broken into ranges and the vertical axis is the count or frequency.

Shape: skewed left

Outliers

An outlier is a data value that appears extreme relative to the rest of the data.

Outliers can often be identified by examining an appropriate graph of the data.

Some outliers are data entry or data collection errors that can be corrected after they are identified. Other outliers are natural features of the data.

Since outliers impact many calculations, you should inspect your data for outliers near the beginning of your analysis.

Center or central tendency

Measures of central tendency the center, or middle, or typical value of a distribution.

Common measure of center: mean and median

Mean: The sample mean, denoted as $\bar{x}$, can be calculated as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

where x_1, x_2, ..., x_n represent the observed values.

Mean

The population mean is also computed the same way but is denoted as μ. It is often not possible to calculate μ since population data are rarely available.

The sample mean is a sample statistic, and serves as a point estimate of the population mean. This estimate may not be perfect, but if the sample is good (representative of the population),

Median

The median is the value that splits the data in half when ordered in ascending order.

$$0, 1, 2, 3, 4$$

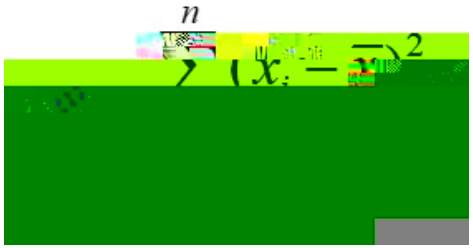
If there are an even number of observations, then the median is the average of the two values in the middle.

$$0, 1, 2, 3, 4, 5 \rightarrow \frac{2+3}{2} = 2.5$$

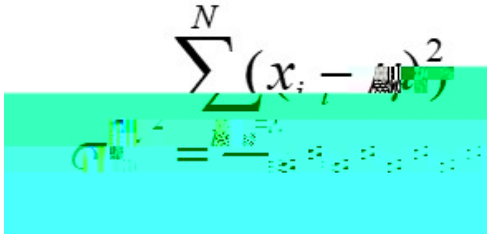
Since the median is the midpoint of the data, 50% of the values are below it. Hence, it is also the 50th percentile.

Variance

The variance is roughly the average squared deviation from the mean. Here is the formula for the sample variance:



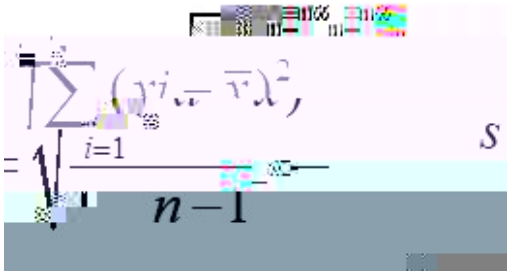
The sample variance is a point estimate of the corresponding population variance. The population variance has a slightly different formula:



Standard deviation

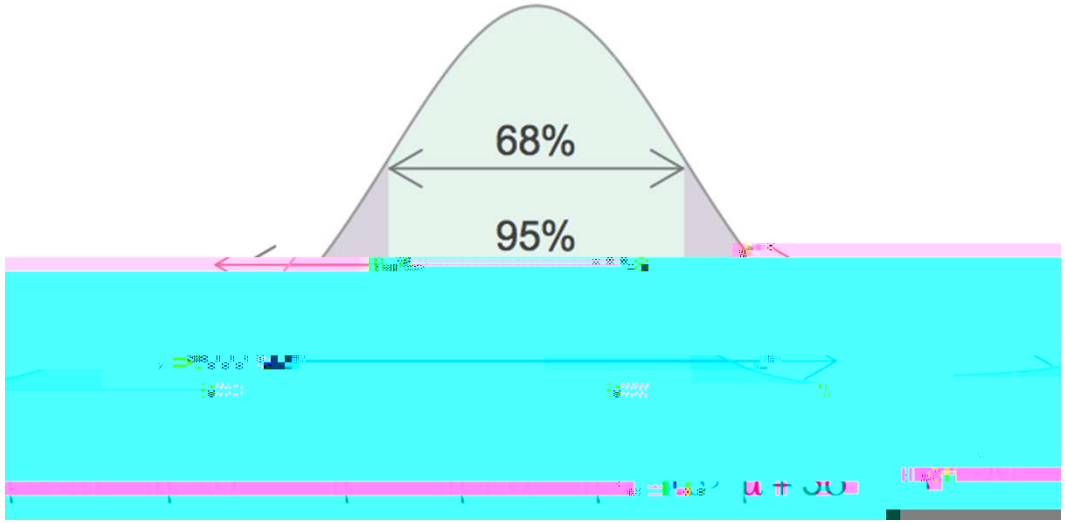
The standard deviation is the square root of the variance. The unit of measure of the standard deviation is the same as the data. This makes the standard deviation more practical than the variance to use in applications. The SD is interpreted as roughly the average distance between observations in the data set and the mean of the data.

Formula:



Empirical rule (or 68-95-99.7 rule)

In a unimodal, symmetric distribution, about 68% of the values fall within one standard deviation of the mean, about 95% of the values fall within two standard deviations of the mean, and about 99.7% of the values fall within three standard deviations of the mean.



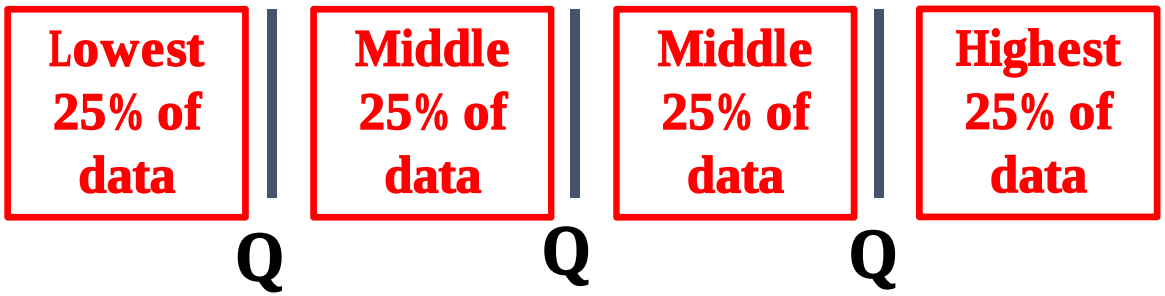
Quartiles

Quartiles are special percentiles that divide a data set into four sections, each containing 25% of the data set.

Q1 = first quartile = 25th percentile

Q2 = second quartile = 50th percentile

Q3 = third quartile = 75th percentile

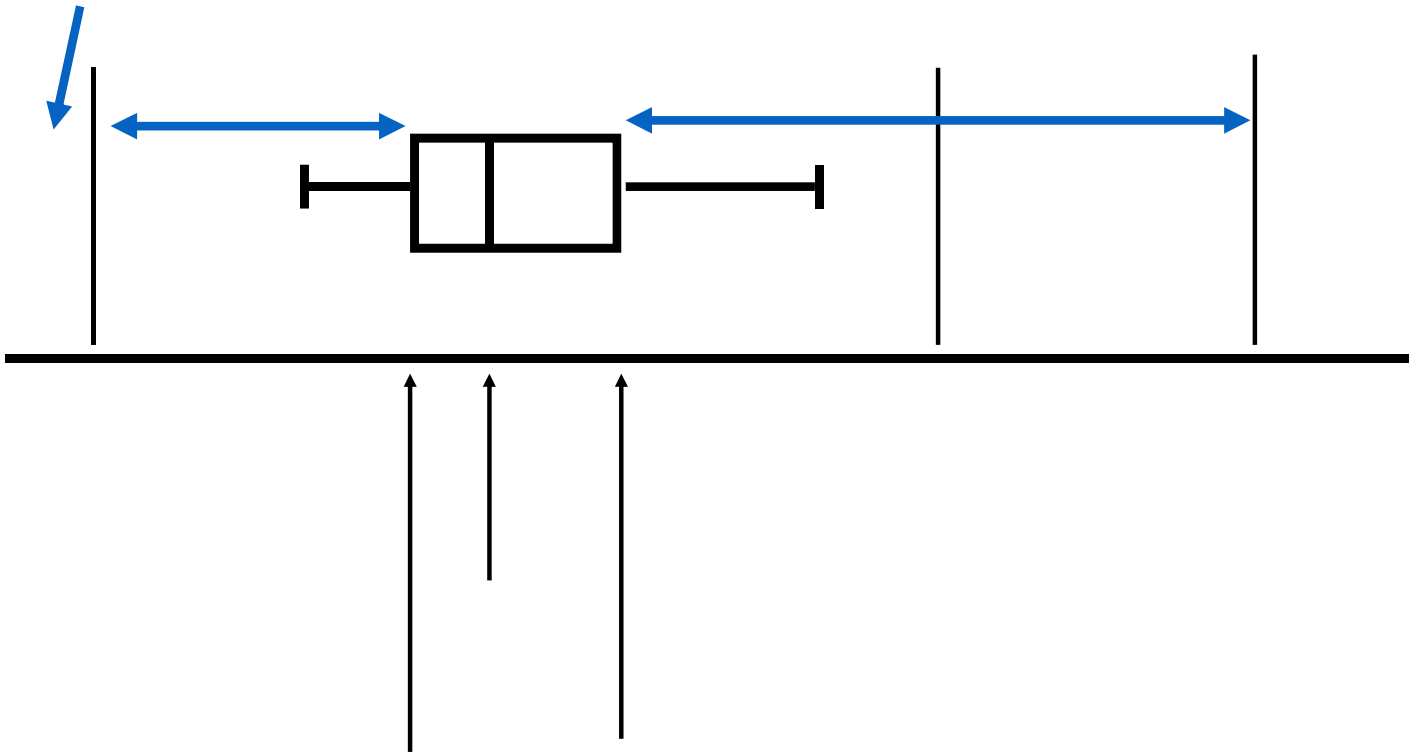


Boxplots

Boxplots (or box-and-whisker plots) are graphical displays built from the five-number summary. The five-number summary consists of the min, Q1, median, Q3, and max.

The box extends from Q1 to Q3. A line is drawn inside the box at the value of the median. The "whiskers" extend to the values of the min and max.

In addition, boxplots are often modified to incorporate outlier detection rules based on distances beyond the quartiles and either 1.5×IQR, for potential outliers, or 3×IQR for probable outliers.



Z-scores

A standardized value, commonly called a z-score, provides a relative measure of the distance an observation is from the mean, which is independent of the units of measurement.

Subtracting the mean from all data values centers the data set at 0. Dividing all of the centered values by the standard deviation scales the values to a new standard deviation of 1.

The proces Ú * a D h inion o

Impact of outliers

Outliers pull the mean .

Outliers the value of the range, variance, and standard deviation.

(Note: measures of variation always get larger when outliers are present.)

Outliers also impact other statistics, such as the correlation, coefficients of regression models, etc.

Some statistics are resistant to the effects of outliers, like the median and IQR.

Identifying outliers

There are several common rules of thumb for identifying outliers.

1) Values above Q3 + 1.5×IQR or below Q1 – 1.5×IQR, which are called the "inner fences," are potential outliers.

2) Values above Q3 + 3×IQR or below Q1 – 3×IQR, which are called the "outer fences," are probable outliers/extreme values.

3) Values with z-scores above +3 or below –3 are potential outliers.

Choosing appropriate measures

The mean and standard deviation are the most popular measures of center and variation. If the data is roughly symmetric in shape and contains no obvious outliers, these measures are acceptable.

The median and IQR, which are both resistant to the impact of outliers, should be strongly considered when the data contains outliers or is strongly skewed in shape.

Categorical data

Categorical data is fundamentally different than quantitative/numeric data.

Averages, standard deviations, and other summary statistics often make no sense for categorical data.

Sample proportion

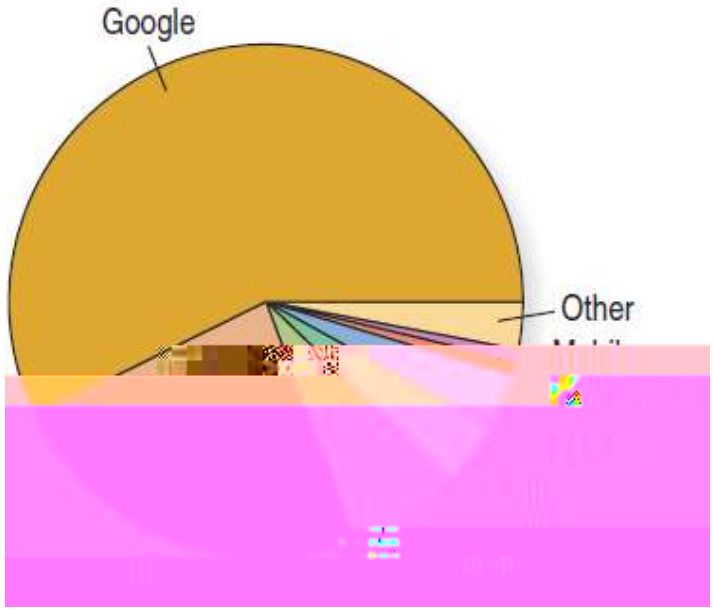
The sample proportion, denoted by or

Frequency distribution

A frequency distribution displays the values of a categorical variable and one or more measures derived from the count of how often each category occurs in the data.

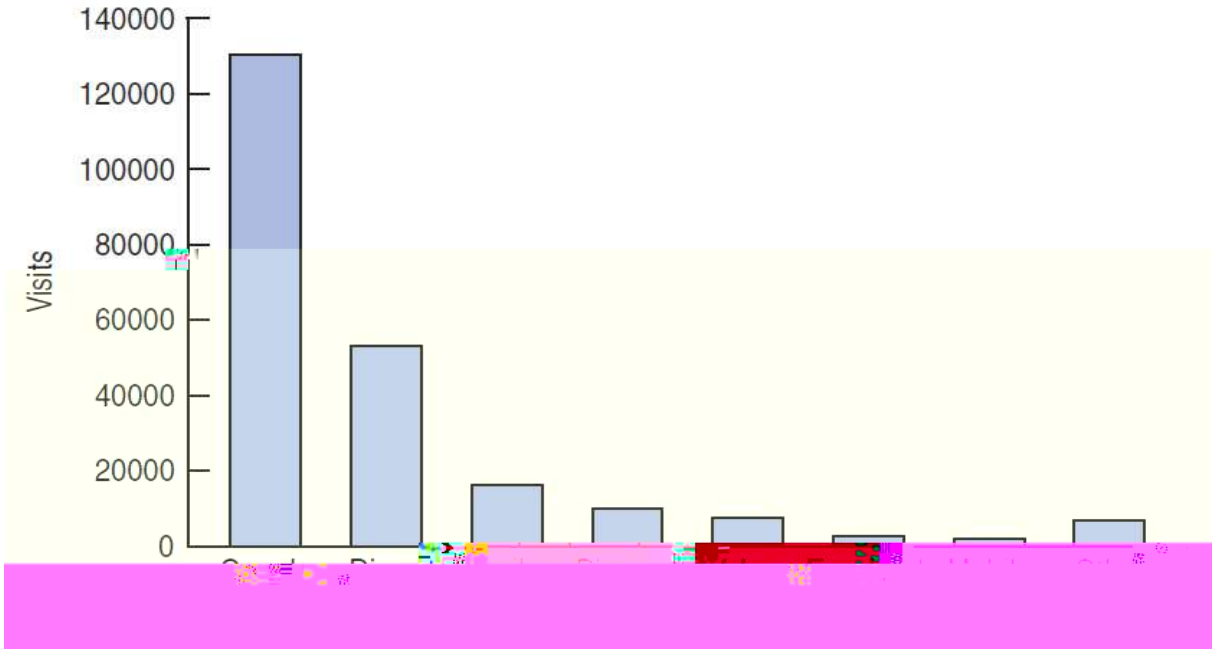
Pie chart

Pie charts show the whole group of cases as a circle sliced into pieces with sizes to the fraction of the whole in each category.



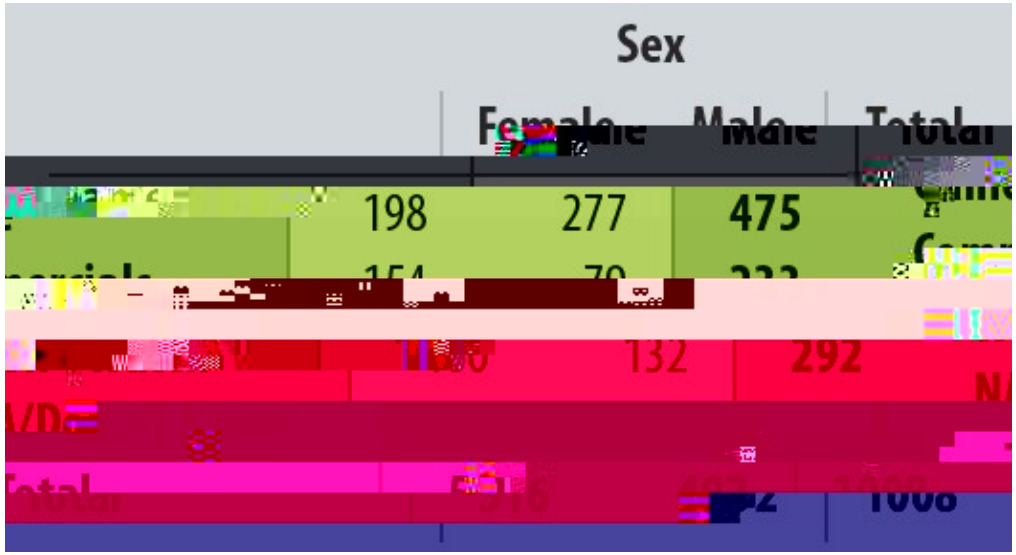
Bar chart

A bar chart displays the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison.



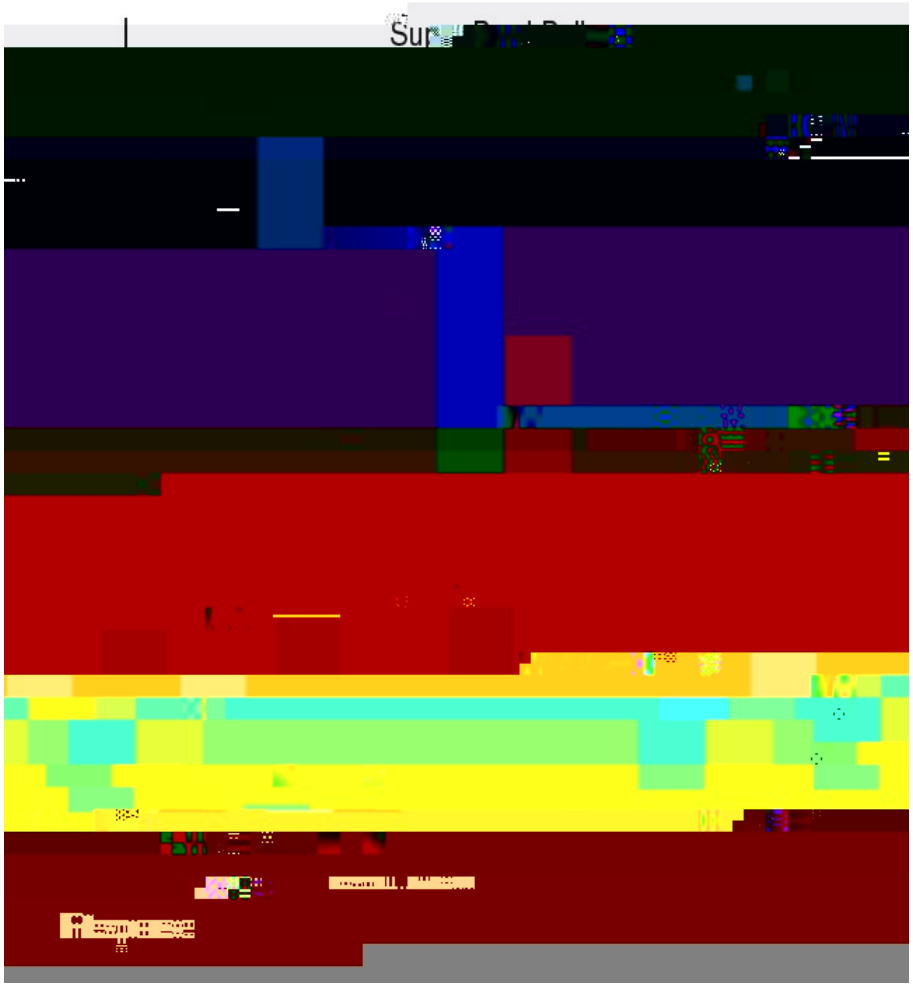
Contingency table

The frequencies of categorical variables can be summarized and displayed simultaneously using a contingency table (or crosstabulation):



Other bar charts

More elaborate bar charts, such as clustered or stacked bar charts can be created from contingency tables:



Practice

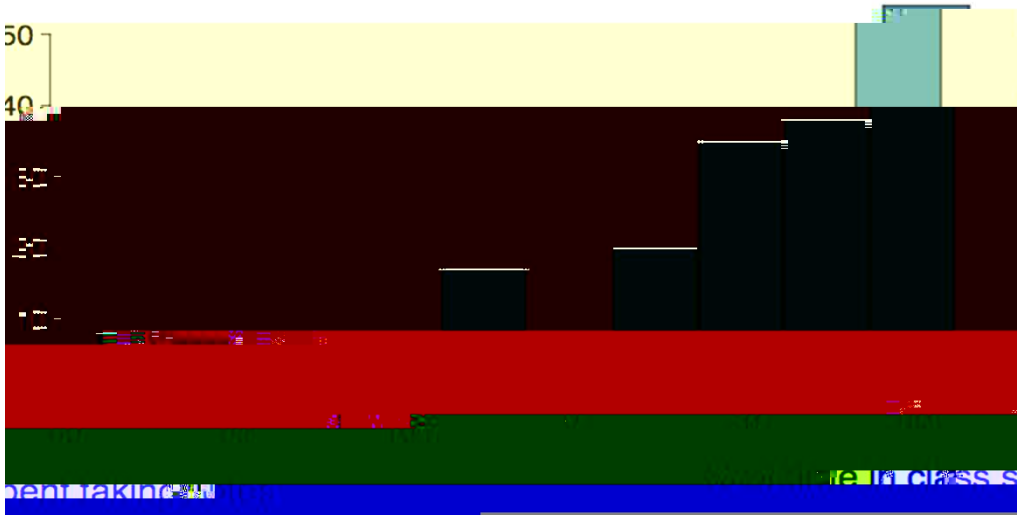
A) Which is most likely true for the distribution of "percentage of time actually spent taking notes in class," which is displayed in the histogram?

(a) mean > median

(b) mean ~ median

(c) mean < median

(d) impossible to tell



Practice

B) Which of these variables do you expect to be uniformly distributed?

(a) weights of adult females

(b) salaries of a random sample of people from North Carolina

(c) house prices

(d) birthdays of classmates (day of the month)

Practice

Practice

D) A community college school board is negotiating a new contract with the college faculty. The distribution of faculty salaries is skewed right by several faculty members who make over $100,000 per year. If the school board wants to give the community the impression that the faculty are already overpaid, should they adverəse the mean or median of the faculty salaries?

The school board should use the mean to make their argument. The mean will be higher than the median since it will be influenced by the few high salaries.

The school board should use the median to make their argument. The median will be lower than the mean since the mean is influenced by the few high salaries.

The school board should use the mean to make their argument. The mean will be lower than the median since the median is influenced by the few high salaries.

Practice

E) A company advertises a mean lifespan of 1000 hours for a particular type of light bulb. If you were in charge of quality control at the factory, would you prefer that the standard deviation of the lifespans for the light bulbs be 5 hours or 50 hours? Why?

50 hours would be preferable since a larger standard deviation indicates a longer average lifespan for the light bulbs.

5 hours would be preferable since a smaller standard deviation indicates more consistency.

50 hours would be preferable since a larger standard deviation indicates more consistency.

5 hours would be preferable since a smaller standard deviation indicates a longer average lifespan for the light bulbs.

Practice solution

A) Which is most likely true for the distribution of "percentage of time actually spent taking notes in class," which is displayed in the histogram?

(c) mean < median

The distribution is skewed to the left and the data values in the left tail pull

Practice solution

B) Which of these variables do you expect to be uniformly distributed?

(d) birthdays of classmates (day of the month)

C) If someone's gross annual income has a z-score of +2.3, what can be concluded?

Their income is 2.3 standard deviations above the mean income.

Practice solution

C) A community college school board is negotiating a new contract with the college faculty. The distribution of faculty salaries is skewed right by several faculty members who make over $100,000 per year. If the school board wants to give the community the impression that the faculty are already overpaid, should they adverəse the mean or median of the faculty salaries?

The school board should use the mean to make their argument. The mean will be higher than the median since it will be influenced by the few high salaries.

Practice solution

E) A company advertises a mean lifespan of 1000 hours for a particular type of light bulb. If you were in charge of quality control at the factory, would you prefer that the standard deviation of the lifespans for the light bulbs be 5 hours or 50 hours? Why?

5 hours would be preferable since a smaller standard deviation indicates more consistency.