Statistics is the science of conducting studies to collect, organize, summarize, analyze, and draw conclusions from data.

Descriptive statistics consists of the collection, organization, summarization, and presentation of data.

Inferential statistics consist of generalizing form samples to populations, performing hypothesis tests, determining relationships among variables, and making predictions.

Measures of Center:

A measure of center is a value at the center or middle of a data set.

The arithmetic mean of a set of values is the number obtained by adding the values and dividing the total by the number of values. (Commonly referred to as the mean)

$\bar{x} = \frac{\sum x}{n}$ (sample mean) $\mu = \frac{\sum x}{N}$ (population mean)

The median of a data set is the middle value when the original data values are arranged in order of increasing magnitude. Find the center of the list. If there are an odd number of data values, the median will fall at the center of the list. If there is an even number of data values, find the mean of the middle two values in the list. This will be the median of this data set. The symbol for the median is $\tilde{x}$.

The mode of a data set is the value that occurs with the most frequency. When two values occur with the same greatest frequency, each one is a mode and the data set is bimodal. Use M to represent mode.

Measures of Variation:

Variation refers to the amount that values vary among themselves or the spread of the data.

The range of a data set is the difference between the highest value and the lowest value.

The standard deviation of a set of sample values is a measure of variation of values about the mean.

$s = \sqrt{\frac{n\sum(x^2) - \left(\sum x\right)^2}{n(n-1)}}$ (sample standard deviation shortcut formula)

$$\sigma = \sqrt{\frac{\sum x - \mu}{N}}$$ (population standard deviation)

Round off rule—Carry one decimal place more than the original data set for measure of central tendency, two places for variation. (standard deviation is a minimum of 2 decimal places)

Pictures of Data:

Distribution is the nature or the shape of the distribution of the data. Terms like bimodal, skewed, bell-shaped, uniform are types of distributions.

A distribution of data is skewed if is not symmetric and if it extends more to one side than the other. (think histogram)

Data skewed to the left are negatively skewed; generally the mean is to the left of the median. Data skewed to the right are positively skewed; generally the mean is to the right of the median.

A histogram is a bar graph in which the horizontal axis represents classes and the vertical axis represents frequencies. The heights of the bars correspond to the frequency values, and the bars are drawn adjacent to each other (without gaps unless a class has a frequency of zero).

To construct a histogram, the date must be organized into a frequency table. A frequency table lists classes of values along with their frequency (count) of the number of values in each class.

Lower class limits are the smallest numbers that belong to each class.

Upper class limits are the largest numbers that belong to each class.

Class boundaries are the number used to separate classes, but do so without gaps. (usually add 0.5 to upper and subtract 0.5 from lower)

$$\text{Class width} \approx \frac{highvalue - lowvalue}{\#ofclasses}$$

$$\text{Class midpoint} = \frac{lowerclass\,\text{limit} + upperclass\,\text{limit}}{2}$$

Guide for constructing frequency tables:

1) Be sure that are classes are mutually exclusive (values belong to only 1 class)

2) Include all classes, even if frequency is zero.

3) Try to use the same width for classes.

4) Select convenient numbers for class limits.

5) Use between 5 and 20 classes

6) The sum of the class frequencies must equal the number of original data values.

Probability Distributions:

A continuous random variable has infinitely many values, and those values are often associated with measurements on a continuous scale with no gaps or interruptions.

If a continuous random variable has a distribution with a graph that is symmetric and bell-shaped, we say it has a normal distribution.

The standard normal distribution is a normal probability distribution that has a mean of 0 and standard deviation of 1.

Z-score (or standard score): the distance along the horizontal scale of the standard normal distribution.

A z- score represents the number of deviations from the mean the value is located. If values are converted to standard scores, then the procedure for working with all normal distributions are the same as those for the standard normal distribution.

$$z = \frac{x - \overline{X}}{\sigma} \quad \text{(sample)}$$

$$z = \frac{x - \mu}{\sigma} \quad \text{(population)} \qquad x = \mu + (z \bullet \sigma)$$

If we convert values to standard scores (z scores), then the procedure for working with all normal distributions are the same as those for the standard normal distribution.

Although a z-score can be negative, the area under the curve (or the corresponding probability) can NEVER be negative.

The probability of the z-score is the area under the curve or the chance the event will happen.

Notation:

$P(a < z < b)$ denotes the probability that the z score is between a and b.

$P(z > a)$ denotes the probability that the z score is greater than a.

$P(z < a)$ denotes the probability that the z score is less than z.

Using a table:

1) Draw the bell-shaped curve, draw the centerline and identify the shaded region under the curve.
2) Using the given x, find the corresponding z to find the probability that represents the shaded area.
3) Find the probability from the table (remember it represents from zero to the value) and perform the correct operation to find the required shaded region.

Central Limit Theorem:

Given;

1) The random variable x has a distribution (which may or may not be normal) with a mean μ and standard deviation σ.
2) Samples all of the same size η are randomly selected from the population of x values. (The samples are selected so that all possible samples are equally likely)

Conclusions:

1) The distributions of sample means x will as the sample size increase, approach a normal distribution.
2) The mean of the sample means will approach the population mean μ.
3) The standard deviation of the sample means will approach $\sigma / \sqrt{n}$.

Practical Rules:

1) For sample of size η larger than 30, the distribution of the sample means can be approximated reasonably well by a normal distribution. As sample size increases, the better the approximation becomes.
2) If the original population is itself normally distributed, then the sample means will be normally distributed regardless of η.

Notation:

If all possible random samples of size η are selected from a population with mean μ and standard deviation σ, the mean of the sample means is denoted by μ_x

$\mu_x = \mu$

The standard deviation of them sample means is denoted by σ_x,

$\sigma_x = \dfrac{\sigma}{\sqrt{n}}$ is also known as the standard error of the mean

Hypothesis Testing:

In statistics, a hypothesis is a claim or statement about a property of a population. A statistical hypothesis is a conjecture about a population parameter.

*This conjecture or claim may or may not be true.

Two types of hypotheses:

The null hypothesis, H_0, is a statistical hypothesis that states that there is no difference between a parameter and a specific value or that there is no difference between two parameters.

***The null hypothesis MUST contain a condition of equality!

The alternative hypothesis, H_1, is a statistical hypothesis that states a specific difference between a parameter and a specific values or states that there is a difference between two parameters.

A test statistic uses the data obtained from a sample to make a decision about whether or not the null hypothesis should be rejected.

A decision is made to reject or fail to reject the null hypotheses on the basis is the value obtained from the test statistic. If the difference is significant, the null hypothesis is rejected, otherwise, fail to reject the null hypothesis.

A type I error occurs if one rejects the null hypothesis when it is true.

A type II error occurs if one fails to reject the null hypothesis when it is false.

***The decision to reject or fail to reject the null hypothesis DOES NOT PROVE ANYTHING. The only way to prove anything statistically is to use the entire population.

Decisions are made based on the difference between the probabilities of the mean obtained and the hypothesized mean.

The level of significance is the maximum probability of committing a type I error. (α)

$\alpha = 0.05$ implies there is 5% chance for rejecting a true null

$\alpha = 0.01$ implies there is a 1% chance of rejecting a true null

Z – test

Finding the critical values:
The critical value(s) separate the critical region from the non-critical region.

The critical region or rejection region is the range of values of the test value that indicates there is a significant difference and that the null hypothesis should be rejected.

The non-critical region or non-rejection region is the range of values of the test values that indicates the difference was probably due to chance and the null hypothesis should not be rejected.

Testing a Claim about a mean: large sample

Assumptions:

1) The sample is a simple random sample.
   Careless samples are useless
2) The sample is large (n > 30).
   Allows us to use the central limit theorem
   No requirement for population to be approximately normally distributed
3) The sample standard deviation is used of the population standard deviation is unknown.

Traditional Method

(test statistic)

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Reject the null hypothesis if the test statistic is in the critical region.

Fail to reject the null hypothesis if the test statistic is not in the critical region.

P-Value

A P-value (or probability value) is the probability of getting a value of the sample test statistic that is at least as extreme as the one found from the sample data, assuming the null hypothesis is true.

Keep in mind that you are deciding whether a sample result is unusual.

Small P-values (smaller that given α) give unusual sample results and have a significant difference. Reject the null hypothesis if the P-value is less that or equal to the significance level (α).

Correlation and Regression:

A correlation exists between two variables when one of them is related to the other in some way.

Assumptions:

1) The sample of paired data (x, y) is a random sample.
2) The pairs of (x, y) data have a bivariate normal distribution. (each x corresponds with a y to give a normal curve)

A scatter plot is a graph in which the paired sample data, (x, y) are plotted with x as the horizontal axis and y as the vertical axis.

Linear Correlation Coefficient:

The linear correlation coefficient, r, measures the strength of the linear relationship between the paired x and y values of a sample.

Pearson Product Moment Correlation Coefficient

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

n = number of pairs of data
r = linear correlation coefficient for a sample
ρ = linear correlation coefficient for a population

Properties of the Linear Correlation Coefficient:

1) The value of r is always between -1 and 1 inclusive.
2) The value of r does not change if all the values of either variable are converted to a different scale.
3) The value of r is not affected by the choice of x or y
4) R is this form measures the strength of a linear relationship only!

Formal Hypothesis Testing

Using Method 1

$H_0 : \rho = 0$
$H_1 : \rho \neq 0$ linear correlation when $\rho \neq 0$

Test statistic for Linear Correlation

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

If the absolute value of the test statistic exceeds the critical values, reject the null hypothesis. Otherwise fail to reject the null and there is not a linear correlation.

Regression

Given a collection of paired sample data, the regression equation

$$\hat{y} = b_0 + b_1 x$$

Algebraically describes the relationship between the two variables. The graph of the regression equation is called the regression line or the line of best fit.

$x =$ independent variable or predictor variable

$\hat{y} =$ dependent variable or response variable

$$b_1 = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} \quad \text{(slope)} \qquad b_0 = \bar{y} - b_1\bar{x} \quad \text{(y-intercept)}$$

Given a collection of paired data (x, y), $(\bar{x}, \bar{y})$ is the centroid. (Find the average of the x's and the average of the y's)

The regression line best fits the sample data points.