

DATA (VARIABLE) CATEGORIES

NOMINAL	Non Parametric (less precise) qualitative	Cannot analyze statistically (cannot add these, cannot find a mean).	Labels, categories, or NAMES Example: Dead/alive, blood types, ethnicity, zip codes, male/female
ORDINAL	Non Parametric (less precise) qualitative	Has order but no origin, no equal distance, cannot find a mean	Example: Likert Scale, rankings 1st-2nd-3rd-4th Small-medium-large We know that one is larger or smaller than the other but not how much.
INTERVAL	Parametric More precise quantitative	Has order and equal distance No origin, no zero Can calculate a mean Assume normal distribution	Example: Frequency counts of things that cannot be divided.
RATIO (THE BEST!)	Parametric Most precise quantitative	Has order, equal distance, absolute zero Can calculate a mean Assume normal distribution	Example: height, weight, B/P, time, temperature, length,

Parametric = based on the assumption that the scores are distributed in a normal, bell-shaped curve.

DESCRIPTIVE STATISTICS are statistics about one variable...

-MIDDLE (“central tendency”)

-SPREAD/VARIATION (“range” or “dispersion”)

These two number concepts are utilized to answer advanced statistical questions.

For example:

A survey was given to eight students asking them how far they lived from campus. The replies were: 5, 7, 9, 13, 11, 8, 10, 9.

(n=total number of responses = 8)

MEASURES OF CENTRAL TENDENCY (the MIDDLE or AVERAGE)

A measure of central tendency gives a single score as typical or representative of all the other scores.

MEAN – parametric – the average of all responses

Add all responses and divide by the total number of responses.

$$\text{Mean} = 5 + 7 + 9 + 13 + 11 + 8 + 10 + 9 / 8 ; 72 / 8 = 9$$

MEDIAN – non parametric – the middle number in the responses

Arrange all responses in numerical order from the least to the greatest.

Then, cross off the edge numbers from the list.

- If n is odd, only the median will remain in the middle.
- If n is even, stop when you have only two middle numbers remaining. The median is the average of these two numbers.

$$\begin{aligned} \text{Median} &= \underline{5}, \underline{7}, \underline{8}, \underline{9}, \underline{9}, \underline{10}, \underline{11}, \underline{13}, \\ &= 9 + 9 = 18 \\ &= 18 / 2 = 9 \end{aligned}$$

MODE – non parametric – the number that appears most often

If two or more numbers tie for the most appearances, there is “no” mode.

The number above that appears the most is 9, so 9 is the mode.

MEASURES OF VARIATION - in addition to knowing the average score or response by examining a measure of central tendency, it is useful to know the degree of spread or individual differences among the scores.

MIDRANGE = lowest number + highest number divided by 2.

The midrange for the above numbers is $5 + 13 = 18$; $18 / 2 = 9$.

RANGE = greatest number minus the least number – or -

The difference between the upper and lower limits. The range for the above numbers is $13 - 5 = 8$. Non parametric and is the least useful due to influence of extreme scores.

PERCENTILES - a score that indicates the percentage of individuals within a distribution that fall below a designated score. Percentiles are not the same as percentage of correct answers. These are used most often in the interpretation of standardized tests.

Semi-Interquartile Deviation – Divide the data from 0-100% into quarters.

$$SIQD = \frac{(Q3 - Q1)}{2}$$

Not influenced by extreme scores.

STANDARD DEVIATION = the most important measure of variation or distribution. It is a measure of the average spread among the individual scores in a set of scores. It is a measure of individual differences. In a normal distribution, six standard deviations cover the range of all of the scores – three standard deviations above and three standard deviations below the mean. The larger the standard deviation, the larger the spread among the scores. In general, when scores in large samples are normally distributed, the higher score will be about three standard deviations above the mean, and the lowest score will be about three standard deviations below the mean. It is parametric and therefore has greater precision. It uses the mean as part of the formula.

To find the standard deviation for a data set:

1. Create a chart as follows:

DATA	DATA-MEAN	SQUARE of DATA-MEAN
X	X-M	$(X-M)^2$

Collect all measurements (responses) and write them down in the first column. All of the data should have the same precision – the same number of decimal places.

2. Find the MEAN for the data in the first column and write it down.

3. Subtract the mean from each number in column one and place that result in column two.
4. Calculate the square root of each number in column two and write the results in column three. Do not round.
5. Add all of the numbers in column three and write the result down. Do not round.
6. Divide the result in step 6 by the number of data pieces in column 1 of your chart and write this result down. Do not round.
7. Take the square root of your result in step 7 and round this to the same number of decimal places as the data that was collected. This is your standard deviation.

$$sd = \sqrt{\frac{\text{sum of } (X - \text{mean})^2}{\text{number of data pieces}}}$$

When you have a choice, always use parametric procedures. They are more precise and can detect smaller differences within the experimental groups. Normal distribution is assumed and a mean can be calculated. Normal distribution has nothing to do with the clinical concept of “normal”. It means the normal distribution is a bell-shaped curve and is symmetrical around the mean. The normal distribution (or bell curve) is characterized by two parameters: the mean and the standard deviation.

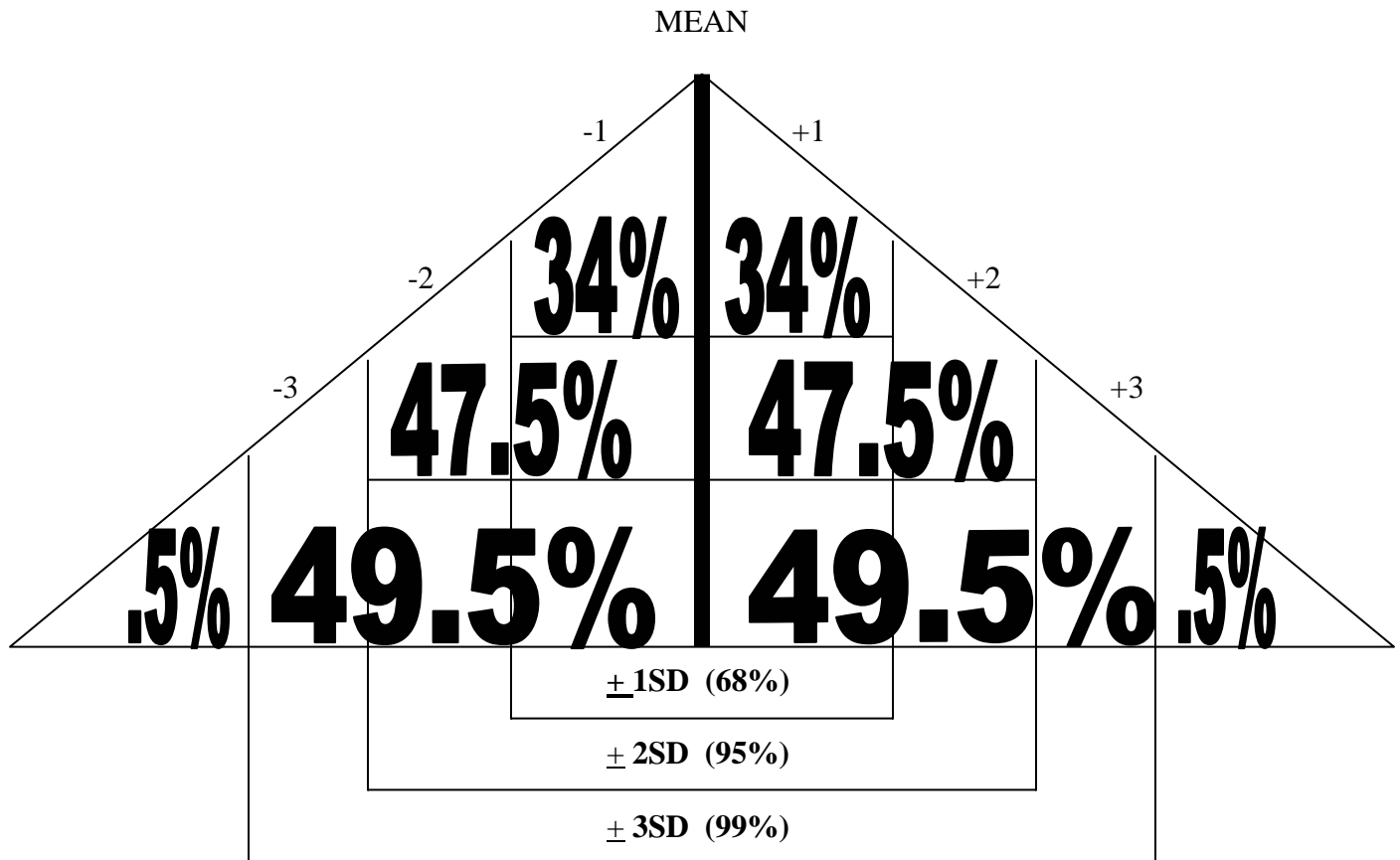
To determine whether the results of your study are valid, check you data to see if the mean is a good estimate of central tendency. If the mean is bad or data is skewed to the right or left, your statistical analysis maybe invalid.

68% of the area under the graph is found within plus or minus one standard deviation from the mean

95% of the area under the graph is found within plus or minus two standard deviations from the mean

99% of the area under the graph is found within plus or minus three standard deviations from the mean

If the distribution is bell-shaped, you can set up confidence intervals to determine the uncertainty of where the true mean would be.



Remember that working hypotheses are usually stated as NULL HYPOTHESES – that there is NO significant difference between two groups. If you reject the null hypothesis, you accept the alternative hypothesis – that there is a significant difference between the two groups.

Experiments are performed to **test** a hypothesis not to prove a hypothesis. To test a hypothesis, determine the mean of each group. Then select a level of probability or acceptable chance of being wrong (the probability of making an error in the conclusion). In medical research, the probability level that is usually chosen is 5% OR 1%, depending on what is being tested and is expressed as $<.05$ or $<.01$.

The alpha level is the minimum level of significance that is acceptable (usually .05 in the medical profession). This means the difference is large enough between the two groups to be significant –that they are really two different groups and not just one large group.

The degrees of freedom = the total number of observations or subjects in the two groups minus two.

$$df = Na + Nb - 2$$

Use the appropriate statistical method for testing the hypothesis to determine the test statistic. Then determine the probability associated with the test statistic using the appropriate table (found in a textbook). Accept or reject the null hypothesis at the chosen probability level and draw the appropriate conclusion about the population.

There are many software programs available to utilize for the statistical analysis portion of your project. Using a software program will be more efficient with less chance of mathematical errors than relying on the longhand method.

T-TEST – a statistical analysis to determine the probability that the difference between the means of two sets of scores is the result of mere chance fluctuations in the scores. Used when the standard deviation of the population is unknown. You must assume that the distribution of the population is normal (a bell-shaped curve).

Hypothesis: There is no significant difference between the results of the test group and the control group after treatment.

EXAMPLE: Suppose you ran a t-test with 32 subjects (16 in each group) and got a t-score of 2.30. To determine if this result is significant at the .05 level of significance, look down the left-hand column of the table of significance – labeled df for degrees of freedom until you come to 30 ($32-2 = 30$). Next, look under the alpha level of .05 written across the top of the table. Then, go down the .05 column to the place where the .05 alpha level intersects with the row for 30 degrees of freedom. The number that appears at the intersection is 2.042. Ask yourself: Is 2.30 greater than 2.042? The answer is yes, and therefore the result is statistically significant at the .05 level of significance and the Null Hypothesis is rejected. There IS a significant difference!

In general, almost all tables of significance are interpreted in this manner. Some tables are more complex, but the strategy of interpretation is the same

ANOVA-Analysis of Variance is used to examine the significance of the differences among two or more groups.

Chi square – estimates the probability that an obtained pattern of scores is merely a chance variation of a theoretical or expected pattern.

ACTIVITIES

See “STAT PRACTICE SHEET”

ASSESSMENT

For each question above, determine the category of the data collected, whether the data is parametric or non parametric, qualitative or quantitative, and if a mean can be calculated.

STAT PRACTICE SHEET

Student Name: _____ **Date:** _____

Determine the following statistics by creating a data collection instrument, administering the survey, and then graph the results for each question separately utilizing a software program – (bar, line or pie graphs)

1. How many students in the class are males and how many are females?
2. How many of each eye color are there?
3. Determine each student's age in months.
4. Determine each student's height in inches.
5. Determine the number of brothers and sisters each student has.
6. Determine each student's favorite choice from colors below:
 - _____ Black
 - _____ Blue
 - _____ Green
 - _____ Red
 - _____ Other