

**Professor W. David Hoisington** (last update 1/22/12)

**Overview:** This is a data analysis project and as such you need to consider not only the data but the way that you do the analysis, how you reach your conclusions and what supportive research you have for your conclusions.

**Step one:** Choosing your 3 areas of analysis from the hospital survey (remember you need to do more to get an A).

An Excel spreadsheet will be sent to you by email when you are done with the first assignment. You will need to complete a minimum of one qualitative and two quantitative analyses. The qualitative analysis should be thematic and have 4 to 5 themes that are discovered from the word/write-in responses. The quantitative analysis **MUST** be a comparison of two data sets and at least one of those sets **MUST** have a numerical sequence (1, 2, 3 ...).

Example: For the patient response analyses, you will need to develop comparison subgroups. The subgroups might be categorized as a) hardly ever, b) sometimes, c) most of the time, d) almost always, e) always. These answers can be converted to a numerical sequence a=1, b=2, c=3, d=4, e=5. Now you can do a quantitative analysis that compares the text responses to quantitative data in the other subsets.

**Step 2:** First organize your data

Use Excel to organize your data. Once you have your analysis topics determined, go to the complete data set (it has been sent to you as an Excel document). Open that in Excel. Next copy the original data set into a second worksheet within your original workbook. **DO NOT WORK** on your original data sheet.

Rename your two worksheets. I suggest “All Data” as the original data sheet and “Analysis” for the second worksheet. The reason for this is if you inadvertently lose data from the Analysis worksheet you can always copy the All Data worksheet again.

You may wish to create separate copies of the worksheet and then eliminate columns of data to have just two columns of the comparative data sets.

For the qualitative analysis you want to make sure you can read the answers - so increase the width of the column to the point where the material can be read. Use the next column to the right of the patient responses where you will enter in the number value for the patient response. Do not use the number zero. Instead start with the number 1.

***Question: What is the best way to count the frequency of words in the column?***

The best way to count is to use a formula. Using the COUNTA formula will count the number of times the cells in a range have an entry (versus cells that have nothing in them – no number or value).

For the purposes of counting the number of specific times the cell contains a unique value, you must use the formula COUNTIF

=countif(A2:A10, "food")

There are a couple of things to understand with this formula:

- 1) There should be no spaces anywhere in the formula
- 2) Your range (in this example, A2:A10) is specific to where values exist. Your values may exist in a different column or row. Using the "A2:A10" in this example is just to demonstrate the formula.
- 3) You must have the number, letter, word or phrase you are searching on in the formula between quotation marks

If you are running formulas, place the formula identification in the cell next to the formula, for instance, you would type, "count of A".

### Step 3: Qualitative Analysis\*

- 1) All the words written for a given answer are considered as a group of data.
- 2) You are looking for a way to make categories, or themes, out of these word responses. You need to find a theme, for example "attitude of nurses" and then find comments that address this theme. You create a separate column in your spread sheet so you can code the responses that would fit this theme (and other these you decide are in the data). Maybe you use a number, or a letter, or even a word. You place that code in the new column next to the survey response.
- 3) In qualitative analysis the work is often about "coding" the text responses, identifying the response into a category. The best way to do this is by hand (sorry, the computer can't make that kind of decision).

Step 1) Copy a new data sheet.

Step 2) Label the column next to the patient responses as "Categories".

Step 3) Read over the written patient response and decide which category is a match.

Step 4) Type that category name into the "Categories" column.

Remember you can have from 4 to 7 categories.

- 4) After all of the responses have been placed into a thematic category, then you can make a frequency graph by counting the number of responses in each category.
- 5) Next is your analysis and reflection - what does the data mean? Include one reference.

### Step 4: Quantitative Analysis

Unlike the qualitative, here you are seeking to see if a hypothesis can be shown to be true. For example, you might have the hypothesis that younger women have higher birth weight babies. The answer to this hypothesis is two fold - it is a **frequency graph with the mean** (illustrating the difference, you should do this with your data) and also a calculation of the statistical significance of the difference between the means. You are doing two (2) quantitative analyses, and thus you will have two hypothesis statements.

There are several different ways to use Excel to do analysis on groups of cells. What you will need to do in a T-test on one of your hypotheses and then one additional test on your other hypothesis. There is an Excel help document on the CSJ page, and there are great Excel tutorials out there (including the Excel help).

5) Do frequency graphs for each of your subgroups (you can place them on one graph with different colors). Note that Excel will make these for you. Having the mean on the graph and the N (number) of each bar is required.

6) Discuss the statistical test used, what were your expectations (hypothesis), and then what did you find. Provide an analysis and reflection. Here one reference can be used.

#### **Step 4: Writing the report**

I suggest that you read chapter 13 in "Questionnaire Research". There is an easy to understand format for your report. Feel free to add more if you wish.

The format of the paper will be as follows:

1. Title
2. Abstract
3. An overview of the survey (see the website for information)

The following procedures are to be headings in the paper

#### **A. Qualitative analysis.**

- a) Analysis statement: describe the intent of the question and your expectation for analysis.
- b) Procedure
- c) Results (include frequency graphs) - support references would be added
- d) Conclusions

#### **B. Quantitative analysis**

- a) Hypothesis statement: the relationship between the independent and dependent variables followed by clarification in layman's terms and research documentation as necessary.
- b) Procedure
- c) Results - to include mean, mode, standard deviation and mean (You should have a mean and the associated value or N for every data set).
- d) Conclusions

One or more of the following: bar graph, scatter plot, line graphs, correlation coefficient.

Two or more of the following: The T-test, Z-test, Chi-square test, ANOVA.

4. **Summary and Conclusions** - draw conclusions from all the data gathered by your group, also include in your final conclusions information from the graphs posted on the website.
5. **References Citations** (at least one for every analysis).

## T-Tests and Z-Tests

T-test is best applied if you have a limited sample size ( $n < 30$ ) as long as the variables are approximately normally distributed and the variation of scores in the two groups is not reliably different. It is also good to use if you do not know the populations' standard deviation.

If the standard deviation is known, then, it would be best to use another type of statistical test, the Z-test. The Z-test is also applied to compare sample and population means to know if there's a significant difference between them. Z-tests always use normal distribution and also ideally applied if the standard deviation is known. Z-tests are often applied if the certain conditions are met; otherwise, other statistical tests like T-tests are applied in substitute.

Z-tests are often applied in large samples ( $n > 30$ ). When T-test is used in large samples, the t-test becomes very similar to the Z-test. There are fluctuations that may occur in T-tests sample variances that do not exist in Z-tests. Because of this, there are differences in both test results.

### Summary:

1. A T-test is appropriate when you are handling small samples ( $n < 30$ ) while a Z-test is appropriate when you are handling moderate to large samples ( $n > 30$ ).
  2. T-test is more adaptable than Z-test since Z-test will often require certain conditions to be reliable. Additionally, T-test has many methods that will suit any need.
  3. T-tests are more commonly used than Z-tests.
2. Consider an experiment where you've measured values in two samples, and the means are different. How sure are you that the population means are different as well? There are two possibilities:
  - The populations have different means.
  - The populations have the same mean, and the difference you observed is a coincidence of random sampling.

The P value is a probability, with a value ranging from zero to one. It is the answer to this question: If the populations really have the same mean overall, what is the probability that random sampling would lead to a difference between sample means as large (or larger) than you observed?

The P value is a fraction. In many situations, the best thing to do is report that number to summarize the results of a comparison. If you do this, you can totally avoid the term "statistically significant", which is often misinterpreted.

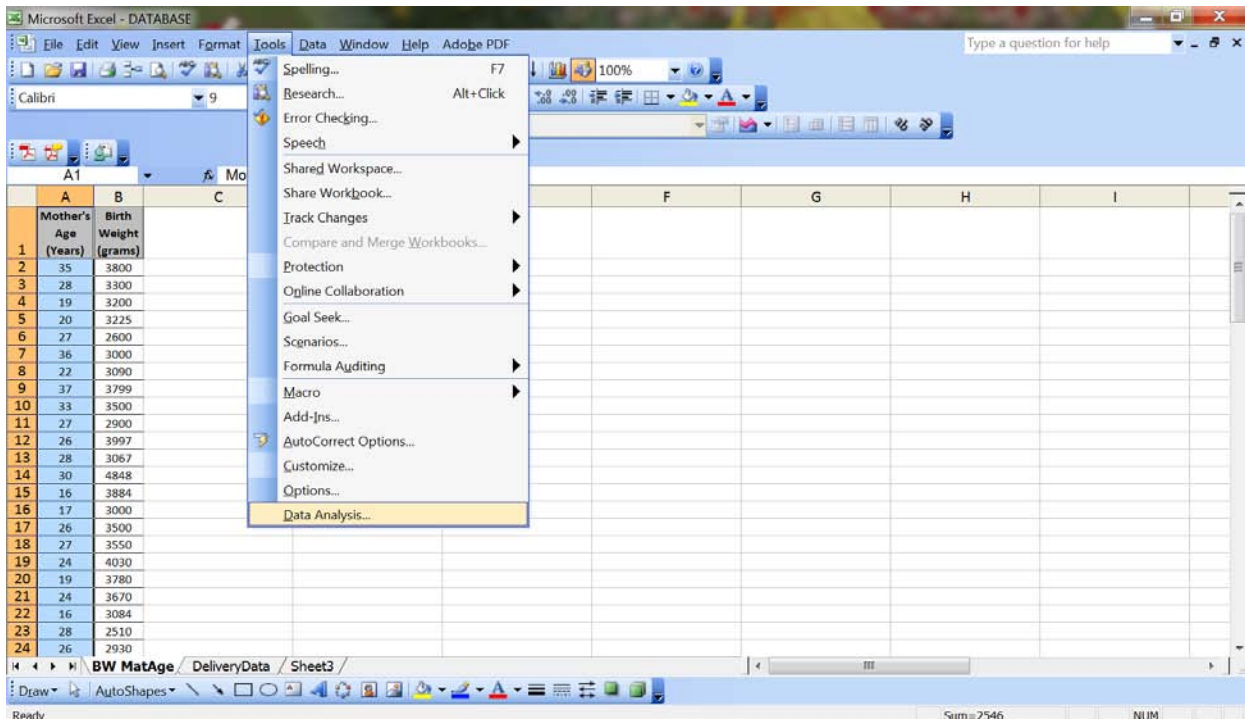
In other situations, you'll want to make a decision based on a single comparison. In these situations, follow the steps of statistical hypothesis testing.

1. Set a threshold P value before you do the experiment. Ideally, you should set this value based on the relative consequences of missing a true difference or falsely finding a difference. In fact, the threshold value (called alpha) is traditionally almost always set to 0.05.
2. Define the null hypothesis. If you are comparing two means, the null hypothesis is that the two populations have the same mean.
3. Do the appropriate statistical test to compute the P value.

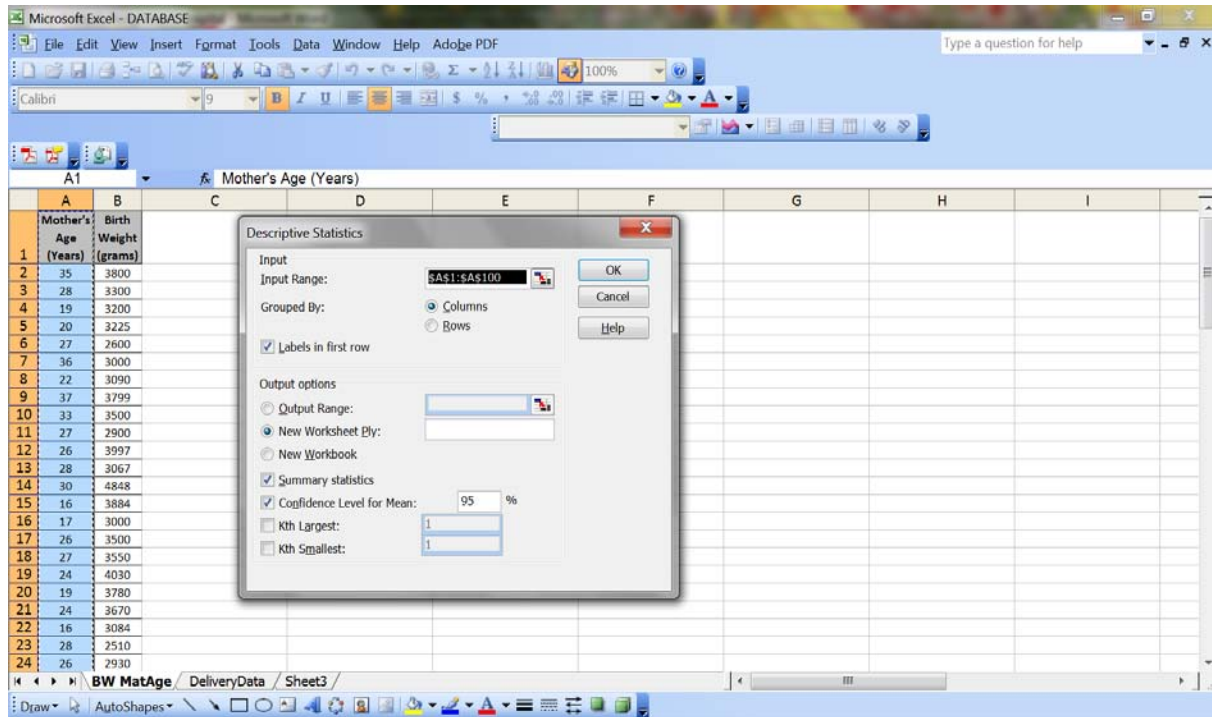
4. Compare the P value to the preset threshold value. If the P value is ***less*** than the threshold, state that you "reject the null hypothesis" and that the difference is "statistically significant". If the P value is ***greater than*** the threshold, state that you "do not reject the null hypothesis" and that the difference is "not statistically significant".

Using Excel will save some time and provide statistical calculations. In this example, two populations are sampled, Mother's Age and Birth Weight. The individual "scores" for each are set up in columns on Excel. These examples use Excel 2003. Some steps may be different in Excel 2007.

Perform a data analysis using Tools>Data Analysis>Descriptive statistics. Use this function on one group (column) of data at a time.



The resulting descriptive statistics is created in a separate spreadsheet in the Excel workbook. You need to compare if the variances between the two groups, Mother's Age and Birth Weight, are equal or unequal. Equal in this case would not mean exactly numerically equal, but somewhat equal.



The “wizard” window will ask you to set the parameters for the data analysis.

Be sure to select the input range of the column, in this case, B1:B100.

You can also check the “labels in the first row” to allow Excel not to use the first row labels as a data source.

Check also the “summary statistics” box, and then “OK”. A separate worksheet is created with the information.

Microsoft Excel - DATABASE

File Edit View Insert Format Tools Data Window Help Adobe PDF

Calibri 11 B I U

B20

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	<i>Mother's Age (Years)</i>														
2															
3	Mean	25.7													
4	Standard Error	0.5													
5	Median	26													
6	Mode	24													
7	Standard Deviation	5.1													
8	Sample Variance	26.1													
9	Kurtosis	-0.5													
10	Skewness	0.2													
11	Range	21													
12	Minimum	16													
13	Maximum	37													
14	Sum	2546													
15	Count	99													
16	Confidence Level(95.0%)	1.0													
17															
18															
19															
20															
21															

MatAge Stats BW MatAge / DeliveryData / Sheet3

Draw AutoShapes

Ready NUM

Repeat the process for the second column of data, in this case, “Birth weight”.

Microsoft Excel - DATABASE

File Edit View Insert Format Tools Data Window Help Adobe PDF

Calibri 11 B I U

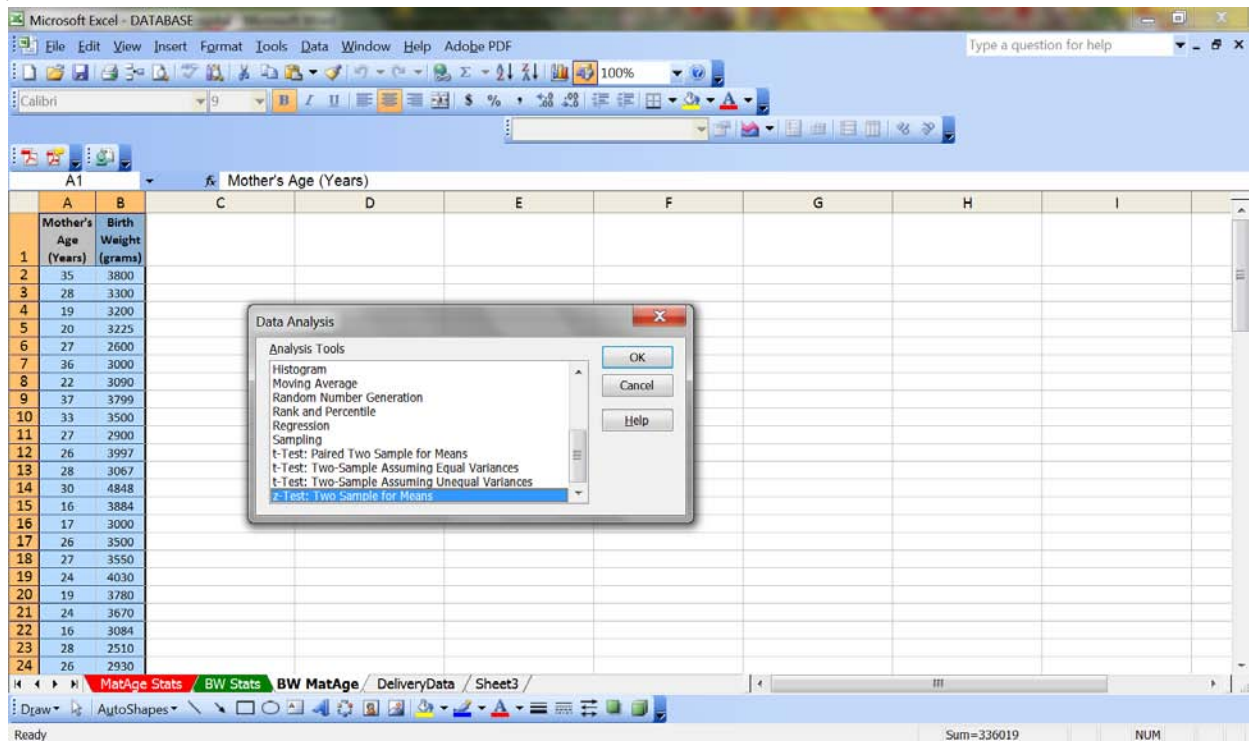
F16

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	<i>Birth Weight (grams)</i>														
2															
3	Mean	3368.4													
4	Standard Error	59.5													
5	Median	3400													
6	Mode	2900													
7	Standard Deviation	591.6													
8	Sample Variance	350004.4													
9	Kurtosis	4.4													
10	Skewness	-1.0													
11	Range	4233													
12	Minimum	615													
13	Maximum	4848													
14	Sum	333473													
15	Count	99													
16	Confidence Level(95.0%)	118.0													
17															
18															
19															
20															
21															

MatAge Stats Sheet4 BW MatAge / DeliveryData / Sheet3

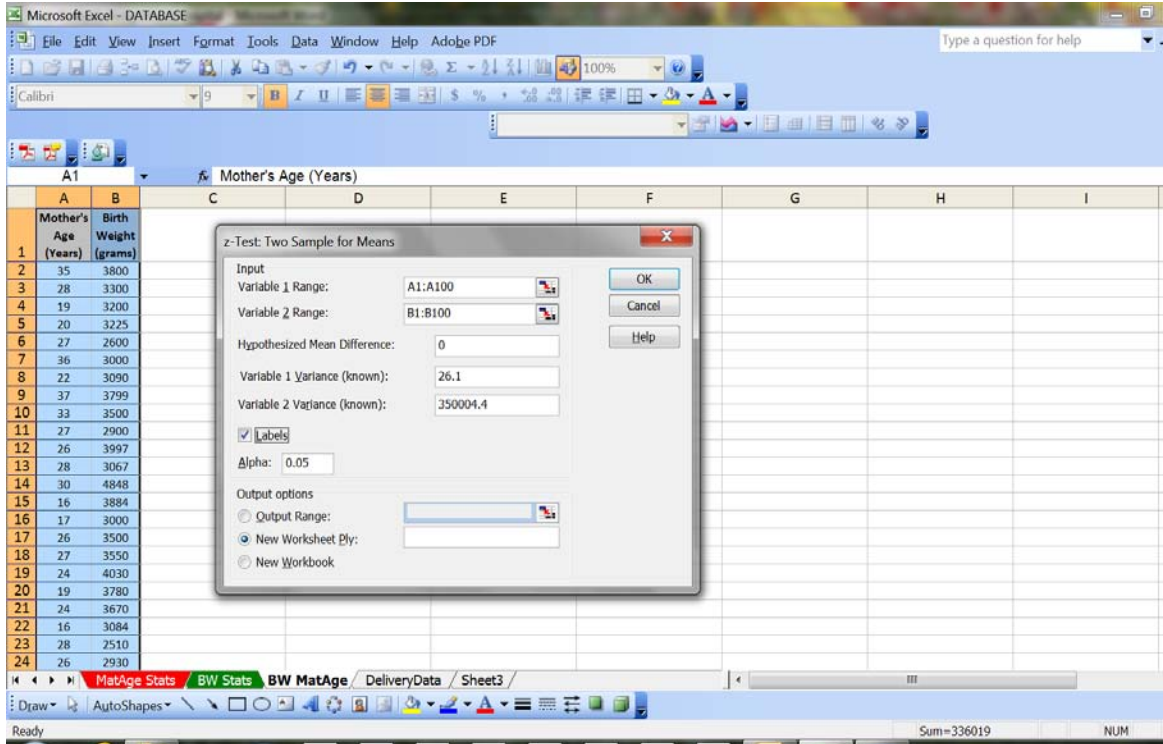
Draw AutoShapes

Now that you have compared the variances, you can perform the next step.



Highlight both columns of data and return to Tools>Data Analysis> and select the z-test.

The selection process in this example sets the variable 1 range (Mother's Age) with data in rows a2:a100, and the second variable (Birth weight) as b2:b100. You can select a1 or b1, to start the variable range with the label but you must check the "labels" checkbox to exclude that information as "data".



Enter the “hypothesized mean difference” as zero (0). That indicates that the null hypothesis assumes there is no difference between the means (statistically).

Enter in 0.05 as the alpha. This calculation indicates that you are seeking a 95% probability (traditional threshold value) that the calculations are statistically significant.

Click “OK”. A separate spreadsheet will be generated in Excel to provide the t-test analysis on the two variables you have selected.

Compare the P value to the preset threshold value of 0.05.

	A	B	C	D	E	F	G	H	I	J	K	L
1	z-Test: Two Sample for Means											
2												
3		<i>Mother's Age (Years)</i>	<i>Birth Weight (grams)</i>									
4	Mean	25.71717172	3368.414141									
5	Known Variance	26.1	350004.4									
6	Observations	99	99									
7	Hypothesized Mean Difference	0										
8	z	-56.21622267										
9	P(Z<=z) one-tail	0										
10	z Critical one-tail	1.644853627										
11	P(Z<=z) two-tail	0										
12	z Critical two-tail	1.959963985										
13												
14												
15												
16												
17												
18												
19												
20												
21												

If the ***P value is less*** than the threshold, state that you "reject the null hypothesis" and that the difference is "statistically significant".

If the ***P value is greater than*** the threshold, state that you "do not reject the null hypothesis" and that the difference is "not statistically significant". Statistically speaking, it is more correct to say "do not reject" than "accept".

In the example above, the P ( $Z \leq z$ ) one-tail test is 1.64 which is more than the threshold of 0.05, therefore indicating that we "do not reject the null hypothesis that there is no difference between the means of the two populations, Mother's Age and Birth weight, and that the difference between the two populations is not statistically significant".

Some more helpful tutorials (videos):

<http://www.youtube.com/watch?v=JlfLnx8sh-o>

<http://www.youtube.com/watch?v=SHOSNYLrhus>