Experiment 1 Yahtzee or validating the t-test

Learning Goals

- 1. Receive training and practice with many useful spreadsheet functions in Excel.
- 2. Review and "experience" the concepts of probability and statistics.
- 3. Appreciate the foundations on which the concepts of the confidence limits and the t-table are based.

Introduction

Confidence limits

In general, when we make a measurement, we would like to report a level of confidence in our results. Assuming the accuracy of our method has been independently established, this level of confidence depends on the precision of our method. Regardless of precision, if our method is accurate, the actual correct result would be equal to the average value of an infinite number of measurements (μ). Of course, making an infinite number of measurements is not possible. So, what if we make three repeat measurements? Could we expect the average of three measurements to be close the actual correct result? Would the average of ten results be closer? How many measurements must we make? The concept of confidence limits gives quantitative answers to these questions.

$$\mu = m \pm s_m t_{(N,CL\%)}$$

Using this expression one can calculate a range of values in which you can be certain to a preset confidence level that the true mean (μ) lies within. One would calculate the standard error of the mean (s_m) from the repeated measurements and obtain t from the t-table.

$$S_m = S_e / (N^{1/2})$$

Since the sample mean is an estimation of the true mean of the total population, this statistical factor can be understood as the *standard deviation of the error* in the sample mean relative to the true (total population) mean.

The t-table provides different values of t corresponding to different degrees of freedom (generally N-1) and a desired confidence limit.

For instance the t-value for N=10 at the 95 % CL is 2.262. One of the purposes of this exercise is to show how this value of t is connected with the 95 % CL when N=10.

The Nature of the Exercise

When you roll five, six-sided dice, the sum of the numbers that land face up range from 6-36. The distribution of these sums estimates a normal distribution with a mean of 17.5 and a standard deviation of 3.8. The probability distribution below is a dice-rolling simulation in which five dice were rolled together 10,000 times.¹ Probability distributions are histograms with percentages on the y-axis, rather than the total quantity of that observation.



Normal distributions are often referred to as Gaussian or bell curve distributions. These types of probability distributions exemplify a vast array of processes that occur in nature (from the spread of 10,000 dice rolls to the IQ scores of a large random population to the average height of 15 year old males in a given random population).

A useful observation is that roughly 95% of all data sampled for a normal distribution lies within two standard deviations of the true mean (μ). The range between +/- 2SD is referred to as the 95% confidence interval.



¹ <u>http://www.financeisland.com/tutorials/Statistics.jsp</u>

As your number of observations (degrees of freedom) increase, this rule becomes more accurate with respect to the true mean—you can see that the (t) values approach 2.0 as the degrees of freedom for the 95% CL approach infinity.²

	Confidence level (%)						
Degrees of freedom	50	90	95	98	99	99.5	99.9
1	1.000	6.314	12.706	31.821	63.657	127.32	636.619
2	0.816	2.920	4,303	6.965	9.925	14.089	31.598
3	0.765	2.353	3.182	4.541	5.841	7.453	12.924
4	0.741	2.132	2.776	3.747	4.604	5.598	8.610
5	0.727	2.015	2.571	3.365	4.032	4.773	6.869
6	0.718	1.943	2,447	3.143	3.707	4.317	5.959
7	0.711	1.895	2.365	2.998	3.500	4.029	5.408
8	0.706	1.860	2.306	2.896	3.355	3.832	5.041
9	0.703	1.833	2.262	2.821	3.250	3.690	4.781
10	0.700	1.812	2.228	2.764	3.169	3.581	4.587
15	0.691	1.753	2.131	2.602	2.947	3.252	4.073
20	0.687	1.725	2.086	2.528	2.845	3.153	3.850
25	0.684	1.708	2,060	2.485	2.787	3.078	3.725
30	0.683	1.697	2.042	2.457	2.750	3.030	3.646
40	0.681	1.684	2.021	2.423	2.704	2.971	3.551
60	0.679	1.671	2.000	2.390	2.660	2.915	3.460
120	0.677	1.658	1.980	2.358	2.617	2.860	3.373
∞	0.674	1.645	1.960	2.326	2.576	2.807	3.291

In this exercise you will illustrate the concept of confidence limits by using an Excel spreadsheet to simulate 10000 rolls of the five dice, calculate the sums, and calculate the average and standard deviation of these sums. More importantly, we will divide these 10000 rolls into *1000* sets of ten and calculate the mean (m), standard deviation (s_e), and standard deviation of the mean (s_m). We will do this as a way of investigating the statistical confidence in which one can measure the real mean (μ) by measuring a small sample size (N=10). We will calculate *t*-calc values based on rearranging the confidence limit expression and determine how many of these *t*-calc values fall within the range predicted by the *t*-table; thus validating some of the values of *t* in the table. This exercise will solidify an appreciation the origin of the t values in the t- tables and an appreciation for why these values can be used in conjunction with a small sample size to state with statistical confidence a range of values that contains the true mean (μ).

The Exercise

Step 1: Open the Excel file on the desktop named "ex 1 t-test template".

Step 2: Probabilities

In columns B and C there is a table in cells B1:C7 labeled dice roll and probability. The probability of rolling a 1 on any single roll of the die is 1/6 or 0.1666667. There is the same probability of rolling any one of the other numbers 2-6. Using the mouse click on C2 under the probability column and type "=1/6"

² Harris, Daniel C. <u>Quantitative Chemical Analysis, Fifth Edition</u>. New York: W.H. Freeman and Company

and enter. If the formula returns a date instead of a number, click on the "number" drop down box and select "general". Highlight all cells this may apply to and apply the same method. Click again on the cell and this time select "copy". Then click on C3 holding down the button on the mouse and scrolling down to C7. C3 through C7 should be highlighted. Select paste. 0.1666667 should now appear in cells C2 through C7.

Step 3: Random Number Generator

In this step we will simulate the outcomes of rolling five dice 10000 times

For Office 2003 Select "tools", "Data Analysis" and "Random Number Generator".

For Office 2007

On the upper set of tabs, click on "Data". Now click on the "Data Analysis" in the box to the far right. Select "Random Number Generator".

If you don't already have the "Data Analysis" button, you will need to add it. Click on the multicolored Microsoft "Orb" in the top left of the screen to enable the drop down menu. Click on "Excel Options" in the bottom right corner. Click on the "Add-Ins" tab on the left side. In the drop down tab in the bottom, middle of the screen, click Excel Add-ins. Click "Go". In the Add-Ins box select "Analysis ToolPak". Click "OK". If the program prompts you to install the add-in, click "Yes".

For Office 2010

On the upper set of tabs, click on "Data". Now click on the "Data Analysis" in the box to the far right. Select "Random Number Generator".

If you don't already have the "Data Analysis" button, you will need to add it. Click on the "File" tab. Click on (1)"Options". Click on the "Add-Ins" tab on the left side. In the drop down tab in the bottom, middle of the screen, click Excel Add-ins. Click (2)"Go". In the Add-Ins box select (3)"Analysis ToolPak". Click "OK".



You should now have the Data Analysis button available in the far right of the Data tab. Follow the above directions for selection of the Random Number Generator.

On the random number generator Window enter "5" for the number of variables, enter "10000" for the number of Random Numbers and for distribution select "Discrete". Click on the "Value and Probability Range" space and click, hold and drag the mouse to select cells B2:C7. Next select output range and click in the "output range" space and then click, hold and drag the mouse to select cells E2:I10001. Be patient. This is not easy and may take some practice. It may take a while to generate 60000 numbers! Click "OK" If done correctly, there should now be integers ranging from 1-6 in each of the cells in E2:I10001.

Step 4: Calculating the sum of each of the outcomes

Each row in the E2:I10001 set of cells represents the result of rolling five dice. In column K we will sum up the result of each trial. We anticipate that these sums follow a normal distribution about some mean sum. We can use probability to postulate that the mean sum should be close to 17.5, which would be the mean if we performed an infinite number of throws. Click on cell "K2". Type "= sum(E2:I2)" and enter. Click again on cell "K2" and this time select "copy". Then click on K3 holding down the button on the mouse and scrolling down to K10001. K3 through K10001 should be highlighted. Select paste. This is not easy and may take some practice. If done correctly, the sum of each of the 10000 trials should be shown in column K.

Step 5: Calculating the average, s_e, s_m and t_{calc} for sets of ten rolls Eventually we would like to break the 10000 sums calculated in Step 4 into 1000 sets of ten and calculate the mean, s_e, s_m and t_{calc} for each set. There is more then one way to tackle this but the easiest way seems to be calculating running results for subsequent sets of ten and using the "LOOKUP" function to select the appropriate 1000 sets. So, here goes nothing. Click on cell "N2" and type "=average(K2:K11)" and enter. Click on cell "O2" and type "=stdev(K2:K11)" and enter. Click on cell "P2" and type "=O2/(10)^.5" and enter. Click on cell "Q2" and type "=(17.5-N2)/P2" and enter. Now in cells N2:Q2 you should have the values for the average, s_e , s_m and t_{calc} for the first set of ten. Highlight N2:Q2, select copy, click and hold on N3:Q3 and drag down to N10001:Q10001, and select paste. Now, you should have a running set of values for the average, s_e , s_m and t_{calc} . However, you only want to select every tenth set, so that each roll is only used in the calculation once. This is where the "Lookup" function comes in handy.

Step 6: Using the Lookup function to select the appropriate 1000 sets from the M2:P10001 array.

Click on cell T2 and type "=lookup(S2,SM2:SM10001,N2:N10001)" and enter. The value in N2 should now also appear in T2. What this function does is it looks for the value in cell S2 in the M2:M10001 array, finds it and calls up the value in the N2:N10001 array that is in the same row. Copy T2 to cells U2:W2. The value in cells O2, P2 and Q2 should appear in cells U2, V2 and W2. Note that the purpose of the \$ is so the lookup value and lookup array do not change when you copy the cells to the three other columns. Only the result array shifts. E.g.: If you write "=E5+F5" in a cell, that cell will show the addition of E3 and F3. If you copy and paste to the cell below that, the equation will automatically change into "=E6+F6", also, if you copy that into the cell to the right of the first one, equation will change into "=F5+G5". The absolute sign (\$) locks the cells you specify into the formula you are creating, thereby keeping the value within the cell you specify, a constant. Using \$ appropriately can often be a big time saver.

Now, select cells T2:W2, select copy, highlight cells T3:W3, continue to hold and drag the mouse to highlight "T3:W1001" and select paste. Now, in cells T2:W1001 you should have the values for the average, s_e , s_m and t_{calc} for 1000 sets of ten sums from the roll of the five dice.

Step 7: Calculating the average s_e , s_m , s_{pop} , mean_{pop}

In cell X2 calculate the average s_e value from column U. In cell X4 calculate the average s_m from column V. In cell X6 calculate the standard deviation of the entire population from column K, "=stdevp(K2:K10001)". In cell X8 calculate the mean of the entire population from column K, "=average(K2:K10001)".

Step 8: Investigating the nature of the t-test.

Expression for t-test

$$\mu = m \pm s_m t_{(N=10, 95 \% CL)}$$

If you take a small sample, such as a set of ten, you can say that you are 95% certain that the true value lies within the range dictated by the measured mean (m), s_m for the data set and the t value for the 95 % CL—dictated by the number of measurements. From a probability argument we can reason that the true value is 17.5. This is also confirmed by the result in cell X8. We can use this true value (μ), the mean of the data set (m), the standard deviation of the mean (s_m) for the

data set, to calculate a value for t. We did this and it is in column W for the 1000 data sets. We would expect that only about 5 % of these values lie outside the range confined by 2.262 and -2.262 [2.262 = t(95%, n=10) from the t-table]. Let's check this and see if it works. Click on the "W" on top of column W. The entire column should be highlighted when you do this. Select "copy". Now, highlight column Z and click on "paste special" and select "values" and enter. This will copy the values in column W, without the underlying formulas, which will enable you to sort this data by decreasing value. Highlight column Z. Select the Data tab at the top and select Sort.

For Office 2010: If a "Sort Warning" prompt arises, click "Continue with the current selection" option and click "OK". Choose "tcalc" for the "Sort by" column, "Values" for "Sort On", and "Largest to Smallest" for the Order.

Click on header row and descending and hit OK. The t_{calc} values should now be sorted in decreasing order. Count how many values are above 2.262 and how many values are below -2.262 and add them together. Then, divide this sum by 1000 and multiply by 100. This value should be reasonably close to 5, since one would expect about 5% of the t_{calc} values to be outside the 2.262 to -2.262 range. Try a similar exercise for the 99 % CL and the 90 % CL.

Step 9 Performing the experiment with sets of 3 Attempt to repeat the whole exercise to look at sets of three. This means you should end up with 1000 sets of three, which suggest you only need 3000 rolls of the dice. Good luck!!!

Step 10 Compare the average s_e and s_m values for the sets of ten and the sets of three and discuss the implications.

Lab Report

Data:

Produce a table that provides the number of sets out of 1000 that fell between t and -t (from the t-table) for the 90 %, 95 % and 99 % CLs for the N=10 data sets.

Produce a table that provides the number of sets out of total that fell between t and -t (from the t-table) for the 90 %, 95 % and 99 % CLs for the N=3 data sets. The valid total is generally a bit less than 1000 because a handful of datasets will have a standard deviation of zero and can not be further processed. Give the valid total in the figure caption.

Discussion Points:

Were the numbers in the t-table validated? What criteria could we use for making this judgment? What if 6 % of the data sets fell outside the 2.262 and -2.262? Is this close enough to 5 % to suggest validation? What could you do with the procedure to get even closer values or to provide an even better validation?

Did you expect the results from the N=10 datasets to match more closely with the expected results than the results from the N=3 datasets. Why or why not?