USE OF EXCEL IN STATISTICS: PROBLEM SOLVING VS PROBLEM UNDERSTANDING

Dr.Avanti P. Sethi

Jindal School of Management,University of Texas at Dallas,Texas,USA

ABSTRACT

MS-Excel's statistical features and functions are traditionally used in solving problems in a statistics class. Carefully designed problems around these can help a student visualize the working of statistical concepts such as Hypothesis testing or Confidence Interval.

KEYWORDS

MS Excel, Data Analysis,Hypothesis Testing, Confidence Interval

1.INTRODUCTION

MS Excel is used in a majority of graduate / undergraduate introductory statistics classes, especially in business schools. Problems related to Regression, ANOVA, difference between two means, etc. are solved using Excel's Add-In *Data Analysis* tool. In addition, students are also taughtstatisticalfunctions such as AVERAGE, STDEV.S, NORM.DIST, etc.

Books such as Keller [1], Abbott [2] have been written on the use of these functions and tools. However, the typical focus in these textbooks, as is the case even in Excel-based statistic classes, ison problem-solving using Excel. For example, the students will use NORM.DIST function to solve a normal Distribution problem.

Such training is critical since in the real world, packages/ add-ons such as MS Excel, SAS, SPSS, or R are used in statistical analysis of the data.

Narges [3] has used Excel in generating tests which can be very useful for those who teach statistics classes. However, what we present here is the use of technology (MS-Excel)to teach statistics by clarifying the underlying mechanisms. We believe that there is a clear difference between using a tool to solve a problem and in using it to actually understand the problem-solving concepts. For example, typical business school students in a statistics class don't know where the probability tables come from. They can find a confidence interval for a population mean, but don't really understand whyit works. They don't fully understand why the Null Hypothesis H_0: $\mu = 50$can be rejected in favor of Alternate Hypothesis H_1: $\mu < 50$ even when the sample mean

DOI :10.5121/ijite.2015.4401

$\overline{x} = 47$ which is less than 50. Or, how does the choice of α impact the confidence level or Hypothesis testing?

In this paper, our goal is to present a set of simple exercises using MS-Excel thatteachers can use to demonstrate intuitively what goes on behind the scene when a decision such as the rejection of the Null Hypothesis is made.

2.PROBABILITY DISTRIBUTION TABLE

Most students, especially in non-technical areas such as business administration, tend to view the probability tables as entities created by some unknown force.Whether it is the binomial distribution, normal distribution, or any other, students are given pre-made probability tables and are asked to use these to find probabilities. At the same time, they also learn Excel's probability functions like BINOM.DISTorNORM.DIST. In this paper, wepresent simple manipulations of the function NORM.S.DISTwhich can be used to demystify the Normal distribution or z-table. NORM.S.DIST(z)gives the area (probability) from - ∞ to z, z being the number of standard deviation measured from the mean. As a reference, NORM.S.DIST(1.43)= 0.92364 or 92.364% which is the probability P(z ≤ 1.43), that is, the area to the left of z = 1.43.

To create our own Normal distribution table, we use NORM.S.DIST*(*$A3+B$2*)* in cell(3,2) to get 0.5000. Then we copy it across the columns to get the 1st row followed by copying the row down to fill the rest of the table. It is simple and intuitive.

A	B	C	D	E	F	G	H
1	Area from - ∞						
2	z	0	0.01	0.02	0.03	0.04	
3	0.0	0.5000	0.5040	0.5080	0.5120	0.5160	
4	0.1	0.5398	0.5438	0.5478	0.5517	0.5557	
5	0.2	0.5793	0.5832	0.5871	0.5910	0.5948	
6	0.3	0.6179	0.6217	0.6255	0.6293	0.6331	
7	0.4	0.6554	0.6591	0.6628	0.6664	0.6700	
8							

Table 1

Many textbooks such as Keller [1] provide multiple z-tables representing probabilitiesfrom various points such as -∞ (Table 1) or the mean. While teaching introductory undergraduate or graduate classes, we have realized that the students fail to understand the relationships between these tables and hence treat them as independent tables. Subtracting 0.5 from the function NORM.S.DIST($A3+B$2)will create a table (Table 2) where the area is measured from the mean as seen below. Similarly, a separate table can be created when the z-scores are negative. Such exercises will make the students realize that the tables are essentially the same, and that any table can be used to solve a problem.

A	B	C	D	E	F	G	H
1	Area from Mean						
2	z	0	0.01	0.02	0.03	0.04	
3	0	0.0000	0.0040	0.0080	0.0120	0.0160	
4	0.1	0.0398	0.0438	0.0478	0.0517	0.0557	
5	0.2	0.0793	0.0832	0.0871	0.0910	0.0948	
6	0.3	0.1179	0.1217	0.1255	0.1293	0.1331	
7	0.4	0.1554	0.1591	0.1628	0.1664	0.1700	
8							

Table 2

The students can even be assigned homework problems where they have to use Excel's built-in functions to create such tables, and then use these tables to solve general normal distribution problems. Similarly, function BINOM.DIST can be used to create binomial distribution table for desired number of trials and probability of success. Likewise, T.DISTcan be used to create T-table.

3.SAMPLING DISTRIBUTION

Excel's Add-In *Data Analysis* can be used to make students visualize concepts such as Central Limit Theorem, Confidence Interval, or Hypothesis Testing. *Data Analysis*, a collection of many useful statistical tools built into Excel, is normally hidden from Excel's menu as most users do not need it. It can be turned on by going into Excel's Option menu.

The instructor can create a multistep homework assignment where the 1st step is to generate a distribution-based population. Here are some suggestions.

3.1.Generate a population

A uniform distribution can be generated by simply using Excel's RANDBETWEEN function. For example, to generate a population between 120 and 180, use RANDBETWEEN(120, 180) and copy it down to say, 2000 times, to get a population of size N = 2000 and mean $\mu \approx 160$.
A normal distribution population can be generated in a similar fashion by using NORM.INV(RAND(),160,15) which will give a mean $\mu \approx 160$ and standard deviation $\sigma \approx 15$.

Since newer versions of Excel no longer have a Random Seed, these 2000 random numbers will be continuously changing. One can use Excel's special copying feature to copy the values only which will not change anymore.

Alternatively, one can generate a population of desired distribution by using Excel's Add-In *Data Analysis* bygoing to *Data Analysis*>*Random Number Generation* and puttingin the necessary parameters.

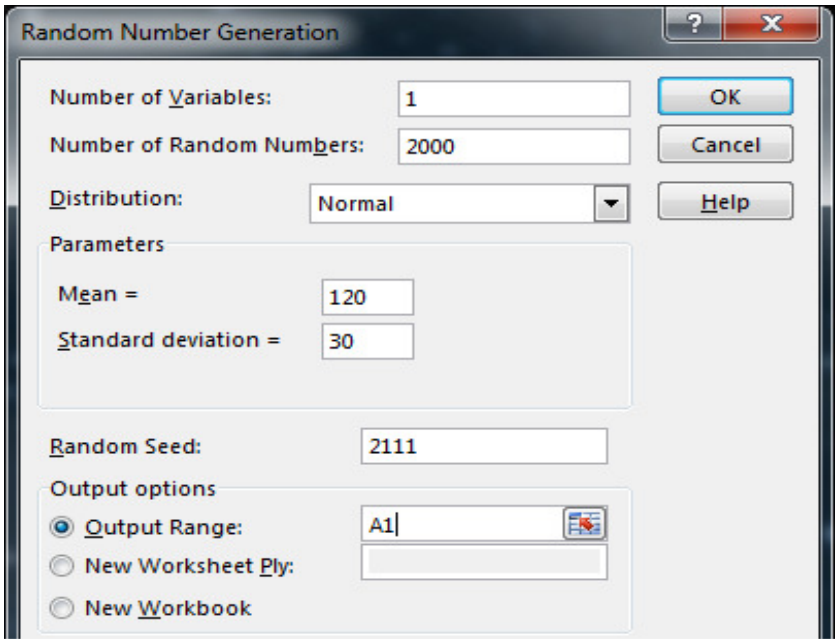

Fig 1. Generate a population

Sometimes it is helpful to gray-out the population numbers in the Excel column to give the students the impression that the population, in general, is not accessible in its entirety.

3.2.Standard deviation σ vs. Standard error of the sample mean $\sigma/\sqrt{n}$:

Now that a population is available (in Column 1), *Data Analysis*>*Sampling* can be used (Fig 2) to generate, as an example,12 samples of size n = 30 from this population in columns B through M.

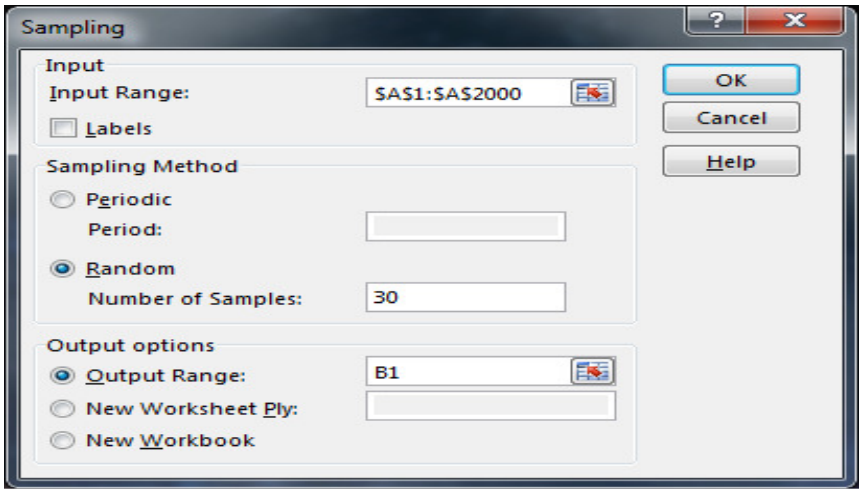

Fig 2: Select a Random Sample

The students are asked to find the mean of each sample(as shown in the Table 3) using AVERAGE. What they will see is that the sample mean is a good estimate of the population mean. Now, they have to find the standard deviation of the sample means using STDEV.S function and compare this with the standard deviation of the population (STDEV.P). For our dataset, the corresponding numbers are 2.88(standard error of sample mean) and 15.26 (population standard deviation).

A look at the table will allow the students visualize why standard error (deviation) of the sample mean is smaller than the standard deviation of the population. They can then be asked to compare this calculated value of 2.88 to the theoretical value of the standard error$\sigma/\sqrt{n} = 15.26/\sqrt{30}$ which happens to be 2.79. The difference between these two valuesof 2.88 and 2.79 will become smaller / larger as the sample size n increases / decreases.

Mean	160.53	158.50	164.97	156.73	157.63	160.20	158.07	158.00	157.07	163.43	158.20	154.63
1	155	186	173	131	164	128	171	146	153	166	167	180
2	200	165	160	165	124	163	147	158	164	149	153	175
3	170	146	151	160	162	200	164	176	165	176	146	151
4	155	149	148	163	141	160	141	152	120	141	152	148
5	205	160	153	153	145	147	168	189	160	162	166	159
6	138	146	172	153	164	163	165	154	176	183	178	168
7	163	139	159	167	169	136	131	151	158	133	134	155
8	158	133	161	134	154	185	149	160	142	167	159	149
9	165	167	184	175	194	147	163	175	162	180	184	128
10	155	145	162	160	162	165	158	164	151	164	156	136
11	130	175	169	164	164	170	150	169	158	152	185	168
12	-	-.	-.									
13	-	-	-									

Table 3

3.3.Confidence Interval

Confidence interval, associated with a Confidence level, is the interval in which the population parameter, such as the mean μ, is expected to fall. To see how this works,the students are asked to calculate a Confidence interval around each of the sample mean in our table with a Confidence interval of 90%. The population mean is expected to fall in most of these intervals (90% to be exact). Thus, with z = 1.645, our interval will be given by $\overline{x} \pm 1.645 \times 15.26/\sqrt{30}$ where $\overline{x}$ is the mean of each sample and $15.26/\sqrt{30}$ represents the standard error of the sample mean $\sigma/\sqrt{n}$.

The following table gives the calculated upper (UCL) and the lower (LCL) levels.

UCL	165.28	163.25	169.71	161.48	162.38	164.95	162.81	162.75	161.81	168.18	162.95	159.38
LCL	155.79	153.75	160.22	151.99	152.89	155.45	153.32	153.25	152.32	158.69	153.45	149.89

Mean	160.53	158.50	164.97	156.73	157.63	160.20	158.07	158.00	157.07	163.43	158.20	154.63

So, if we took the first sample, we expect the population mean to be somewhere between 155.79 and 165.28. But if our sample happened to be the last one, the corresponding interval will be (149.89, 159.38). Which of these is correct? Now we can see the dilemma – for each interval, we are 90% confident that the population mean μ will fall in that interval, but the intervals themselves are so different!

Since in our example, we have access to the whole population, we can calculate its mean μ which in this case is 159.43. By scanning the 12 confidence intervals, we find that interval number 3 and number 12 do NOT include the population mean.Do we expect that? Yes,because 90% Confidence interval implies that we expect the mean NOT to fall in 1 out of 10 intervals.

In a real life scenario,where normally we draw only one sample at any given time, which sample could we be drawing? If we drew sample number 1, then we would be right about the interval estimation of the population mean. But if we drew sample number 3, we would come to a wrong conclusion about the population mean. Such is the life of a statistician!

We can increase the confidence level to 95%, use z = 1.96, but then our interval size increases making it less useful. As a parallel statement, we can be 100% confident that the outside temperature in Septemberin Boston will be between 0 and 100 degrees, but what good is that?

3.4 Hypothesis Testing

A two-tail hypothesis testing is another form of Confidence Interval. Since we know the population mean here, we can set up a Hypothesis test as shown below:

$$H_0: \mu = 159.43$$
$$H_1: \mu \neq 159.43$$

Let us keep our level of significance = α = 10% which gives us z = 1.645. Now, our lower and upper cut-off points can be calculated to be $159.43 \pm 1.645 \times 15.26/\sqrt{30}$ = (154.84, 164.01)

In other words, if the sample mean is less than 154.84 or more than 164.01, we will reject the null hypothesis claiming that the true population mean is not equal to 159.43. With our α = 10% , we are taking a 10% risk of rejecting the true null – that is, of committing a Type I error. Now, if we look at our sample means, sample number 3 (mean = 164.97) and number 12 (mean = 154.63) will make us reject the true null and make us commit a Type 1 error.

4.CONCLUSION

MS-Excel and Add-In *Data Analysi*s are commonly used in solving statistic problems in a classroom setting. The purpose of this paper is to show how the same tools can be used to make the students actually visualize the underlying concepts.

REFERENCES

[1] Keller, Gerald, Statistics for Management and Economics, Cengage Publishing, 10th Edition, ISBN-10: 1285425456

[2] Abbott, Martin, Understanding Educational Statistics Using Microsoft Excel and SPSS, Wiley, ISBN: 978-0-470-88945-9

[3] NargesAbbasi, ShahramDokoohaki, An Initiative in Making Tests for Statistics Lessons, Asian Journal of Education and eLearning, Vol 1, No. 5, 2013

Author

Dr. Avanti Sethi, a faculty member at Jindal School of Management at UT Dallas, received his MS and Ph. D. in Operations Research from Carnegie-Mellon University in Pittsburgh, USA.