Visualizing Statistical Concepts with the Aid of Technology

Tower Chen
Unit of Mathematical Sciences,
College of Natural and Applied Sciences,
University of Guam, UOG Station, Mangilao, Guam 96923.
E-mail: tchen@uguam.uog.edu

Abstract: *Most professional statistics programs only display the data, assumptions, and final numerical output. These programs that omit the intermediate steps and calculations are, from an educational perspective, not the best tools for learning statistics. The use of a spreadsheet framework will aid students in visualizing relationships between data sets and statistical formulas, in exploring data, and in interpreting data. In our new statics course, students learn to develop programs in Excel to solve homework problems and projects. Using their programs, they can review the output from each step without performing repetitive, manual calculations. This allows students to focus more on the statistical concepts rather than the details. Initial assessment results of our new statistics course which utilizes a spreadsheet framework are very positive.*

1. Introduction

This paper discusses how we use technology, specifically spreadsheets, to help students visualize statistical concepts in our new undergraduate statistics course, Bio-Statistics, for Biology majors. Students enrolled in this course are expected to have taken College Algebra or higher. The course is designed as a prerequisite course for an upper division/graduate level course in Biometrics. Statistical methods taught include describing and plotting data, probability theory, sampling theory, estimation theory, hypothesis testing, linear regression, variance analysis, and non-parametric statistics. An emphasis is placed on biological statistics examples and problems. All methods in this course are developed using a tabular approach to construct mathematical formulas to perform operations in statistics paralleled with constructing formulas to perform calculations in Microsoft Excel.

Statistical analyses often demand the processing of large amounts of data. A logical, systematic, effectual method of organizing that information is indispensable. The tabular approach is aptly suited. Calculating statistics from samples, e.g., standard deviation or variance, to make inference about population parameters is often a multi-step process requiring each individual datum to be manipulated in a similar way to the rest before proceeding to the next step. To manage a vast quantity of numbers, it is prudent to tabulate the results from each subsequent step directly adjacent to the data from which it was derived using spreadsheet framework. When employing this approach on large data sets, the greatest difficulty lies in the tedious repetition of the same type of calculation over and over again. This can be conveniently remedied by harnessing the power of computers. The rows and columns of this tabular approach parallel the cell layout in spreadsheet programs, such as Microsoft Excel.

In this course, students initially solve problems manually or with the aid of a calculator, as this method provides students with a better understanding of the basic steps. As the course progresses and students begin to delve into more complex concepts, they will develop programs in Excel to solve homework problems and projects. Students learn to write their own programs to calculate statistics by utilizing spreadsheet cells to organize data and implement the relative and

absolute addresses of cells to perform the repetitive calculations that are necessary to obtain those statistics. This facilitates the learning process by allowing students to visualize the statistical formulae and manipulate large amounts of information without time-consuming tedium. The reduction of repetitive calculations will allow students to focus on the concepts rather than the details and hopefully spur a higher level of interest in statistics.

The programs developed in Microsoft Excel will make transparent each step involved in solving different problems, e.g. to calculate mean, variance, standard deviation, z-value, t-value, confidence interval, F-value, Chi-square value, linear regression line, sum of rank for nonparametric. In addition, the ease of graphing data in Microsoft Excel will also help students visualize the relationships. From an educational perspective, the transparency and the visualizations offered in developing programs in Microsoft Excel will help students learn statistical concepts better than utilizing professional statistical software, which are more useful for commercial or research work.

The following sections explain how we utilize this spreadsheet framework, tables, and graphs teach statistical formulas, the Central Limit Theorem, and hypothesis testing.

2. Statistical Formulas

One of our main goals is to help students not only understand "why" but also "how". Thus, all statistical formulas are presented in a table for better visualization. For example, the following table is used to evaluate the sample variance formula: $s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$.

Table 1 Sample variance formula

n	1	2	…	k	Total
x	x_1	x_2	…	x_k	$\sum x$
$(x-\bar{x})^2$	$(x_1-\bar{x})^2$	$(x_2-\bar{x})^2$	…	$(x_k-\bar{x})^2$	$\sum (x-\bar{x})^2$

Mean $\bar{x} = \frac{\sum x}{n}$

Variance $s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$

Standard Deviation $s = \sqrt{s^2}$

	A	B	C	D	E	F	G
1	*Sample Variance Example*						
2							
3	*n*	1	2	3	4	5	*Total*
4	*x*	2	5	7	8	9	=SUM(B4:F4)
5	$(x-\bar{x})^2$	=(B4-C7)^2	=(C4-C7)^2	=(D4-C7)^2	=(E4-C7)^2	=(F4-C7)^2	=SUM(B5:F5)
6							
7	*Mean*		=G4/F3				
8	*Variance*		=G5/(F3-1)				
9	*Standard Deviation*		=SQRT(C8)				
10							

	A	B	C	D	E	F	G
1	***Sample Variance Example***						
2							
3	*n*	1	2	3	4	5	*Total*
4	*x*	2	5	7	8	9	31
5	$(x-\bar{x})^2$	17.64	1.44	0.64	3.24	7.84	30.80
6							
7	*Mean*		6.20				
8	*Variance*		7.70				
9	*Standard Deviation*		2.77				
10							

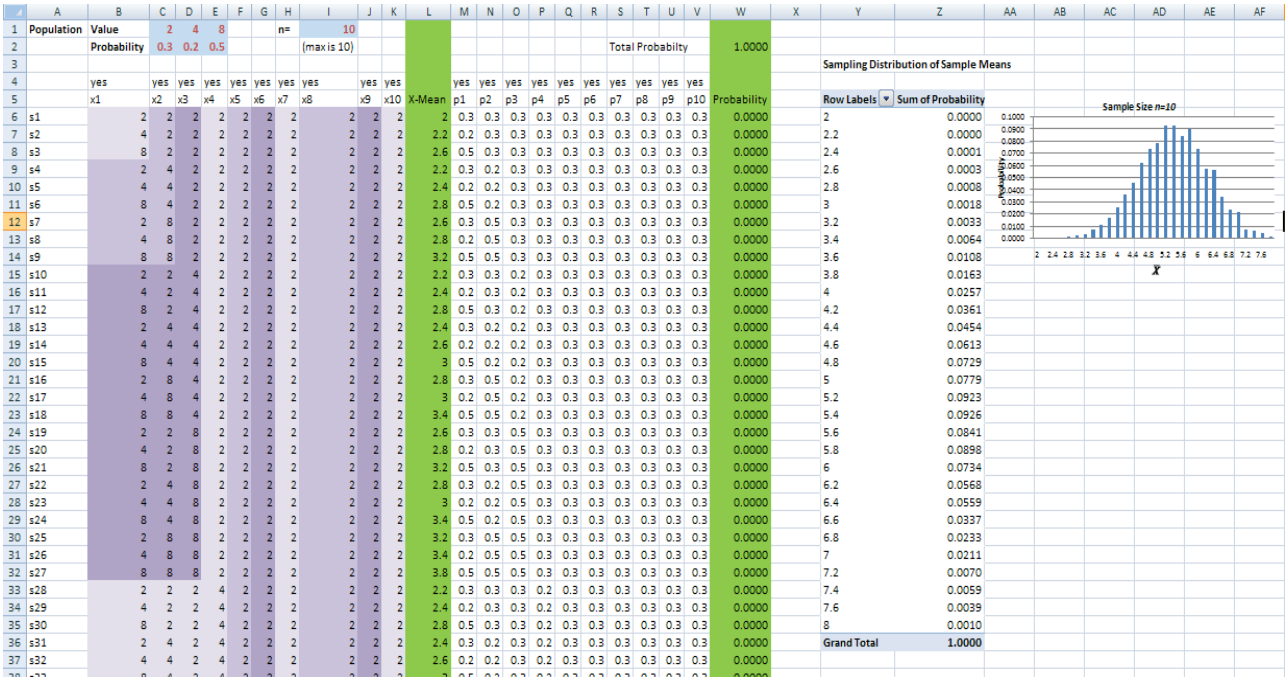
Figure 1 Screenshot of sample variance formula and cell values in Microsoft Excel

The above framework helps students understand the formula's components and the basic steps involved in evaluating the formula. After using the above framework to manually evaluate a formula, students are able to develop programs to automatically evaluate the formula, which should enhance their learning. Upon students entering new data, the program will perform all calculations and display results of each step, in addition to the final result. Students can review the output and check the calculations for each step, as if they would manually perform the calculations.

3. Central Limit Theorem

The Central Limit Theorem states that for any underlying population distribution with mean μ and standard deviation σ, the sampling distribution of sample means $\bar{X}$ approaches a normal distribution with mean $\mu_{\bar{X}} = \mu$ and standard error $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, as the sample size n increases. Students are provided an intuition of the Central Limit Theorem through visualizations of the sampling distributions of sample means of different sample sizes from different underlying population distributions. Through examples created in Microsoft Excel, we illustrate the properties of the Central Limit Theorem. Our examples include populations with positive skew, negative skew and high kurtosis distributions. Using graphs, we show that, regardless of the underlying population distribution, the sampling distribution of sample means approaches a normal distribution as the sample size increases. Using tables, we show that the sampling distribution mean is equal to the population mean; and the standard error, the sampling distribution standard deviation, is equal to the population standard deviation divided by square-root of the sample size.

A screenshot of the program used to create the examples is provided below. The cell values, pivot table, and graph automatically update if the population values (cells C1:E1), population distribution (cells C2:E2), and/or sample size (cell I1) are updated. This program allows us to easily view the effects of changing the underlying population distribution or sample size on the sampling distribution of sample means.



Population	Value	2	4	8		n=	10
	Probability	0.3	0.2	0.5			(max is 10)

Total Probabilty 1.0000

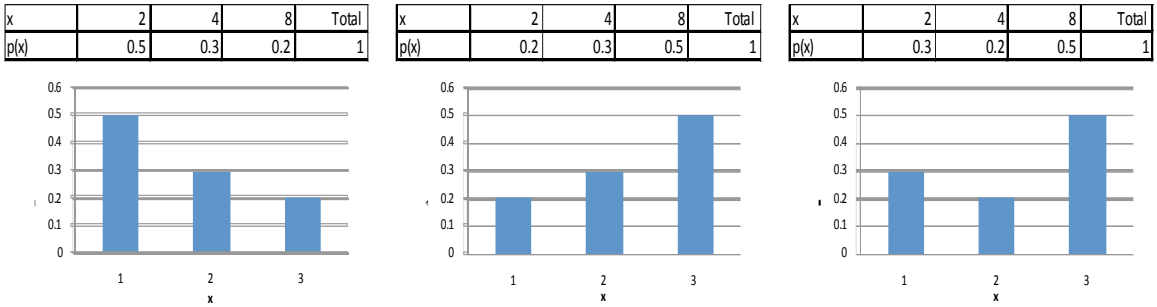
Sampling Distribution of Sample Means

Row Labels	Sum of Probability
2	0.0000
2.2	0.0000
2.4	0.0001
2.6	0.0003
2.8	0.0008
3	0.0018
3.2	0.0033
3.4	0.0064
3.6	0.0108
3.8	0.0163
4	0.0257
4.2	0.0361
4.4	0.0454
4.6	0.0613
4.8	0.0729
5	0.0779
5.2	0.0923
5.4	0.0926
5.6	0.0841
5.8	0.0898
6	0.0734
6.2	0.0568
6.4	0.0559
6.6	0.0337
6.8	0.0233
7	0.0211
7.2	0.0070
7.4	0.0059
7.6	0.0039
8	0.0010
Grand Total	1.0000

Figure 2 Screenshot of Central Limit Theorem example in Microsoft Excel

3.1. Population Distributions

Imagine a large box containing 10,000 identical balls labeled with 2, 4, and 8, representing the population. The population's probability distribution depends on the number of balls with each label. To illustrate the Central Limit Theorem, three discrete probability distributions of the above population are utilized: (a) positive skewed distribution, (b) negative skewed distribution, and (c) high kurtosis distribution. The three population distributions provided below are used as the basis for the examples in our new statistics course, but only the population with positive skewed distribution will be discussed in this paper for saving space.



x	2	4	8	Total
p(x)	0.5	0.3	0.2	1

x	2	4	8	Total
p(x)	0.2	0.3	0.5	1

x	2	4	8	Total
p(x)	0.3	0.2	0.5	1

Figure 3 Population distributions: (a) positive skewed, (b) negative skewed, (c) high kurtosis

Using the information provided in Figure 1, we calculate the population mean and standard deviation to be as followings: (a) positive skewed distribution $\mu = 3.8$ and $\sigma = \sqrt{5.16} = 2.2716$, (b) negative skewed distribution $\mu = 5.6$ and $\sigma = \sqrt{6.24} = 2.4980$, (c) high kurtosis distribution $\mu = 5.4$ and $\sigma = \sqrt{7.24} = 2.6907$.

3.2. Sampling Distributions

Figure 4 illustrates that the sampling distribution of sample means $\overline{X}$ from a population with a positive skewed distribution, approaches a normal distribution as the sample size, n, increases. The samples were drawn from the population presented in Figure 3a.

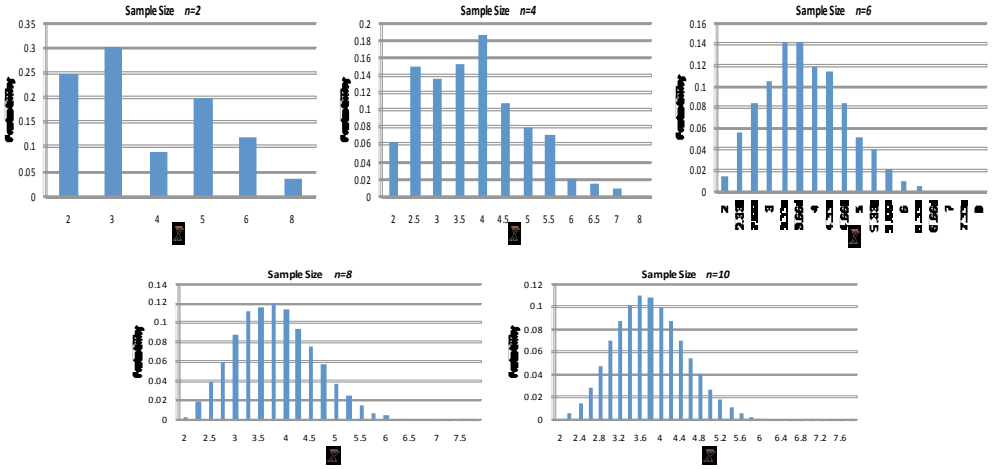


Figure 4 Sampling distributions from positively skewed population distribution in Figure 3a

3.3. Sampling Distribution Mean and Standard Error

The following examples illustrate that the sampling distribution mean is indeed equal to the population mean, and that the sampling distribution standard error is equal to the population standard deviation divided by the square root of the sample size n. Table 2 below provides the sampling distributions of sample means of varying sample sizes drawn from the positive skewed population distribution described in Figure 3a. Table 3 below provides the associated means and standard errors of the sampling distributions.

Table 2 Sampling distributions of sample means of varying sample sizes drawn from the positive skewed population distribution presented in Figure 3a.

n=2

$\overline{X}$	$P(\overline{X})$
2	0.25
3	0.30
4	0.09
5	0.20
6	0.12
8	0.04
Total	1.00

n=4

$\overline{X}$	$P(\overline{X})$
2.0	0.0625
2.5	0.1500
3.0	0.1350
3.5	0.1540
4.0	0.1881
4.5	0.1080
5.0	0.0816
5.5	0.0720
6.0	0.0216
6.5	0.0160
7.0	0.0096
8.0	0.0016
Total	1.0000

n=6

$\overline{X}$	$P(\overline{X})$
2.000	0.0156
2.333	0.0563
2.667	0.0844
3.000	0.1050
3.333	0.1429
3.667	0.1423
4.000	0.1192
4.333	0.1143
4.667	0.0839
5.000	0.0524
5.333	0.0409
5.667	0.0216
6.000	0.0103
6.333	0.0072
6.667	0.0022
7.000	0.0010
7.333	0.0006
8.000	0.0001
Total	1.0000

n=8

$\overline{X}$	$P(\overline{X})$
2.000	0.003906
2.250	0.018750
2.500	0.039375
2.750	0.059750
3.000	0.087937
3.250	0.111510
3.500	0.117103
3.750	0.120575
4.000	0.114978
4.250	0.093682
4.500	0.076370
4.750	0.058565
5.000	0.038056
5.250	0.025872
5.500	0.016209
5.750	0.008288
6.000	0.004939
6.250	0.002419
6.500	0.000932
6.750	0.000538
7.000	0.000161
7.250	0.000051
7.500	0.000031
8.000	0.000003
Total	1.000000

Table 3 Sampling distribution mean and standard deviation for samples drawn from the positive skewed population distribution presented in Figure 3a

Sample Size	Sampling Distribution Mean	Sampling Distribution Standard Error
$n=2$	$(\mu_{\bar{X}} = 3.8) = (\mu = 3.8)$	$(\sigma_{\bar{X}} = 1.6062) = \left(\frac{\sigma}{\sqrt{n}} = \frac{\sqrt{5.16}}{\sqrt{2}} = 1.6062\right)$
$n=4$	$(\mu_{\bar{X}} = 3.8) = (\mu = 3.8)$	$(\sigma_{\bar{X}} = 1.1358) = \left(\frac{\sigma}{\sqrt{n}} = \frac{\sqrt{5.16}}{\sqrt{4}} = 1.1358\right)$
$n=6$	$(\mu_{\bar{X}} = 3.8) = (\mu = 3.8)$	$(\sigma_{\bar{X}} = 0.9274) = \left(\frac{\sigma}{\sqrt{n}} = \frac{\sqrt{5.16}}{\sqrt{6}} = 0.9274\right)$
$n=8$	$(\mu_{\bar{X}} = 3.8) = (\mu = 3.8)$	$(\sigma_{\bar{X}} = 0.8031) = \left(\frac{\sigma}{\sqrt{n}} = \frac{\sqrt{5.16}}{\sqrt{8}} = 0.8031\right)$

4. Hypothesis Testing

We also utilize a spreadsheet framework to teach hypothesis testing. The following is an example of hypothesis testing for the equality of four population means. The null hypothesis is H_o: $\mu_1 = \mu_2 = \mu_3 = \mu_4$. The alternate hypothesis is H_a: the mean of at least one population differs from the rest. To determine if the null hypothesis is rejected at the 95% confidence level, we test if the test statistic F-score is greater than 3.13. The computation of F-score is shown in Figure 5 below.

	A	B	C	D	E	F	G	H	I	J	K
1	x_{ij}	Student1	Student2	Student3	Student4	Student5	Student6	Student7	Total (T_i)	Count (n_i)	T_i^2 / n_i
2	Group1	65	87	73	79	81	69		=SUM(B2:H2)	=COUNT(B2:H2)	=I2^2/J2
3	Group2	75	69	83	81	72	79	90	=SUM(B3:H3)	=COUNT(B3:H3)	=I3^2/J3
4	Group3	59	78	67	62	83	76		=SUM(B4:H4)	=COUNT(B4:H4)	=I4^2/J4
5	Group4	94	89	80	88				=SUM(B5:H5)	=COUNT(B5:H5)	=I5^2/J5
6								Total	=SUM(I2:I5)	=SUM(J2:J5)	=SUM(K2:K5)
7											
8	x_{ij}^2	Student1	Student2	Student3	Student4	Student5	Student6	Student7	Total		
9	Group1	=B2^2	=C2^2	=D2^2	=E2^2	=F2^2	=G2^2	=H2^2	=SUM(B9:H9)		
10	Group2	=B3^2	=C3^2	=D3^2	=E3^2	=F3^2	=G3^2	=H3^2	=SUM(B10:H10)		
11	Group3	=B4^2	=C4^2	=D4^2	=E4^2	=F4^2	=G4^2	=H4^2	=SUM(B11:H11)		
12	Group4	=B5^2	=C5^2	=D5^2	=E5^2	=F5^2	=G5^2	=H5^2	=SUM(B12:H12)		
13								Total	=SUM(I9:I12)		
14											
15		mean=	=I6/J6								
16		TSS=	=I13-(J6*(C15)^2)								
17											
18	ANOVA Table										
19	Source	d.f.	SS		MS		F				
20	Treatment	=COUNTA(A2:A5)	SST=	=K6-(J6*C15^2)	MST=	=D20/(4-1)	=F20/F21				
21	Error	=J6-COUNTA(A2:A5)	SSE=	=C16-D20	MSE=	=D21/(23-4)					
22											

	A	B	C	D	E	F	G	H	I	J	K	L
1	x_{ij}	Student1	Student2	Student3	Student4	Student5	Student6	Student7	Total (T_i)	Count (n_i)	T_i^2 / n_i	
2	Group1	65	87	73	79	81	69		454	6	34,352.67	
3	Group2	75	69	83	81	72	79	90	549	7	43,057.29	
4	Group3	59	78	67	62	83	76		425	6	30,104.17	
5	Group4	94	89	80	88				351	4	30,800.25	
6								Total	1,779	23	138,314.37	
7												
8	x_{ij}^2	Student1	Student2	Student3	Student4	Student5	Student6	Student7	Total			
9	Group1	4,225	7,569	5,329	6,241	6,561	4,761	-	34,686			
10	Group2	5,625	4,761	6,889	6,561	5,184	6,241	8,100	43,361			
11	Group3	3,481	6,084	4,489	3,844	6,889	5,776	-	30,563			
12	Group4	8,836	7,921	6,400	7,744	-	-	-	30,901			
13								Total	139,511			
14												
15		mean=	77.35									
16		TSS=	1,909.22									
17												
18	ANOVA Table											
19	Source	d.f.	SS		MS		F					
20	Treatment	4	SST=	712.59	MST=	237.53	3.77					
21	Error	19	SSE=	1,196.63	MSE=	62.98						
22												

Figure 5 Screenshot of the test statistic F-score formula and cell values in Microsoft Excel

The above framework helps students understand the formula's components and the basic steps involved in evaluating the formula. It also helps them organize the outputs from their calculations. In addition, students can easily change one or more data inputs (cells B2:H5) and view the effects on the values in the ANOVA table without performing additional calculations. By only adjusting the data inputs and viewing the program outputs, they can determine what range of values would lead to a rejection of the null hypothesis.

5. Conclusion

We utilize a spreadsheet framework along with tables and graphs to help students visualize and understand statistical concepts better. We believe that this framework will aid students in understanding relationships between data sets and statistical formulas, in exploring data, and in interpreting data. It is our hope that after completing a statistics course taught using this methodology, students will have built confidence in tackling statistical problems and have gained an interest in more advanced statistical topics. Initial assessments results from both faculty members and students of our new statistics course are very positive. In a recent survey, the students taking this Bio-Statistics course found the course very useful and would recommend the course to their friends.

Acknowledgements Development of the Bio-Statistics course and curriculum materials, including *Applied Statistics Using Spreadsheet Framework* authored by Tower Chen and Grazyna Badowski, was funded by the National Institute of General Medical Science, Minority Opportunity in Research Division RISE grant to U. Guam (R25GM063682). The authors would like to thank Dr. Chris Lobban, Program Director, for his assistance and coordination.

References

[1] Bluman, A. (2008). *Elementary Statistics: A Step By Step Approach*. New York, NY: McGraw-Hill Science/Engineering/Math.

[2] Rice, J. (2006). *Mathematical Statistics and Data Analysis*. Belmont, CA: Duxbury Press.
[3] Aczel, A. (1995). *Statistics: Concepts and Applications.* Chicago, ILL: Irwin Press.
[4] Mc Clave, J., Dietrich II, F. (1992). *A First Course in Statistics.* New York, NY: Macmillan Publishing Company.
[5] Woodbury, G. (2002). *Introduction to Statistics.* Pacific Grave, CA: Duxbury Press.
[6] Watkins, A., Scheaffer, R., Cobb, G. (2004). *Statistics in Action*. Emeryville, CA: Key Curriculum Press.