In Search of the Most Likely Value

Jerzy Letkowski
Western New England University

Abstract

Descripting Statistics provides methodology and tools for user-friendly presentation of random data. Among the summary measures that describe focal tendencies in random data, the mode is given the least amount of attention and it is frequently misinterpreted in many introductory textbooks on statistics. The purpose of the paper is to provide a formal definition of the mode and to show appropriate interpretation of this measure with respect to different random data types: qualitative and quantitative (discrete and continuous). Several cases are presented to exemplify this purpose. Instructional guidelines for implementing the cases, using Google Spreadsheet are provided.

Keywords: mode, statistics, summary measures, probability distribution, data types, spreadsheet.

WHAT IS THE MODE?

Consider a random variable, *X*, having probability (density) function, *f*(*x*). The mode of the variable, if any, is a value x^* that maximizes the probability function (Spiegel, 1975, p. 84, Letkowski, 2013, p. OC13077-2 , McClave, 2011, p. 58, Sharpie, 2010, p. 115:

$$x^* : f\left(x^*\right) = \max_{x \in D_X}\{f(x)\} \; \textit{or None} \qquad \text{(i)}$$

One could refer to such a definition of the mode as a probability-distribution driven definition. In a way, it is a universal definition. For example, the maximum point for the Normal distribution (density function) is μ. Thus μ is the mode of every normally distributed random variable. It is interesting to note that random variables, whose probability-distribution functions have more than one maximum point, are sometimes referred to as bi- or multi-modal variables. For example, a Binomial variable, having $\lfloor (n+1)p \rfloor = (n+1)p$, is bi-modal: $Mode_1 = (n+1)p$, $Mode_2 = (n+1)p-1$, (Wikipedia-Binomial, 2012). This paper categorizes multi-modal variables as modeless.

In the context of a sample, the mode is frequently defined as the value that occurs most frequently. One could refer to this definition as a data driven definition. This is a narrow definition and it can lead to misrepresentation of the mode. In fact, many authors take an oversimplified approach by deriving the mode directly from the sample. Using the spreadsheet-based counting function, *CountIf*(), such a mode would be defined as:

$$x^* : \textit{CountIf}\,(\textit{Sample}, x^*) = \max_{x \in \textit{Sample}}\{\textit{CountIf}\,(\textit{Sample}, x)\} \; \textit{or None} \qquad \text{(ii)}$$

Examples of direct application of this definition are shown in many textbooks, exploring introductory Statistics (Anderson, 2012, p. 101, Black, 2012, p. 78, Donnelly, 2012, p. 84, Larose , 2010, p.92, Levine, 2011, p. 100, Triola, 2007, p. 96). Important weakness of this approach is exposed below (see section CASE 2 - THE MODE FOR QUANTITATIVE-DISCRETE RANDOM VARIABLES). Technically, definition (ii) is appropriate for qualitative random variables and for some numeric-discrete variables, featuring rather small domains.

Arguably, the main reason for misrepresentation of the mode derived from a sample can be attributed to data type and sample size effects. The following sections address this issue and show how to properly calculate the mode for two data types: qualitative and quantitative.

CASE 1 - THE MODE FOR QUALITATIVE RANDOM VARIABLES

A qualitative random variable, *Q*, has a domain, D_Q, that consists of nominal or ordinal categories, c_i (i=1,2,…,m) that occur by chance. *Performance-evaluation Level* (Excellent, Good, Acceptable, Poor, Unacceptable), *Color Preference* (Blue, Green, Red, etc.), *State of Agreement* (Yes, No) or *State of Wisdom* (bright, smart, dense, brainless) are just a few examples of such variables. The probability-distribution function of variable *Q* maps each of the categories, c_i, into probability values, p_i:

$$P(Q = c_i) = f(c_i) = p_i, \; i=1,2,\ldots,m, \qquad \text{(iii)}$$

The mode of variable *Q* is a category, having the highest frequency, if any. Interestingly, for any categorical sample, the mode can be calculated unambiguously, using either definition (i) or (ii):

$$c_k : f\left(c_k\right) = \max_{i=1,2,\ldots,m}\{f(c_i)\} \; \textit{or None}$$

$$c_k : Countif\left(Sample, c_k\right) = \max_{i=1,2,\ldots,m} \left\{CountIf\left(Sample, c_i\right)\right\} \textit{ or None}$$

The two formulas are equivalent since each probability (frequency), $f(c_i)$, is calculated as the ratio of the category count, *CountIf*(*Sample*, c_i) to the sample size, *n*=*CountA*(*Sample*):

$$f\left(c_i\right) = \frac{CountIf\left(Sample, c_i\right)}{CountA\left(Sample\right)}, i = 1, \cdots, m \qquad \text{(iv)}$$

Consider a qualitative sample, consisting of instances of categories (Excellent, Good, Acceptable, Poor, Unacceptable):

> Excellent, Good, Excellent, Good, Acceptable, Poor, Good, Unacceptable, Good, Unacceptable, Excellent, Good, Poor, Acceptable, Good, Acceptable, Poor, Poor, Excellent, Excellent

The sample size is 20. Grouping the sample by categories will make it easier to do the counting:

> Excellent, Excellent, Excellent, Excellent, Excellent, Good, Good, Good, Good, Good, Good, Acceptable, Acceptable, Acceptable, Poor, Poor, Poor, Poor, Unacceptable, Unacceptable

This sample has the following frequency distribution:

P(*Q*=”Excellent”) = *f*(“Excellent“) = 5/20
P(*Q*=”Good”) = *f*(“Good “) = 6/20
P(*Q*=”Acceptable”) = *f*(“Acceptable “) = 3/20
P(*Q*=”Poor”) = *f*(“Poor “) = 4/20
P(*Q*=”Unacceptable”) = *f*(“Unacceptable “) = 2/20

Since *f*(“Good “) has the highest value, 6/20, category “Good “ is an unambiguous representation of the sample’s mode.

INSTRUCTIONAL GUIDELINES FOR QUALITATIVE DATA

Google spreadsheet (Letkowski – Qualitative Mode, 2012) shows a complete solution for a large sample. It also includes detail instructional documentation that students can use to explore the mode for any other categorical sample. Figure 1 (APPENDIX) shows the final result.

It is important to note that function *Mode*(), available in both Excel and Google spreadsheet, is unable to compute the mode for qualitative samples, except when the samples’ categories are expressed numerically. Thus, in order to compute the mode of a qualitative sample, using the *Mode*() function, the sample’s categories must be first mapped (*one-to-one*) into numbers. Next, the *Mode*() function is applied to the “numerically” expressed sample and its result is mapped back to the corresponding category, if any. This procedure is shown on a separate worksheet, “Using Mode Function”, in (Letkowski – Qualitative Mode, 2012). A fragment of this worksheet is also shown in Figure 2 (APPENDIX).

CASE 2 - THE MODE FOR QUANTITATIVE-DISCRETE RANDOM VARIABLES

A quantitative-discrete random variable, *I*, has a domain, D_I, of integers. D_I can be a finite or infinite. Common examples of such variables are Binomial and Poisson. A Binomial variable, *I*(*n*,*p*), has a finite domain of nonnegative integers, $D_I = (0,1,2,\ldots n)$. A Poisson variable, $I(\lambda)$, has an infinite domain of nonnegative integers, $D_I = (0,1,2,\ldots,\infty)$. Probability distribution functions and formulas for the mode of these variables are shown in (Wikipedia-Binomial, 2012, Wikipedia-Poisson, 2012).

Consider the following example, representing a sample of random variable *ClassSize* (Anderson, 2012, p.101):

32, 42, 46, 46, 54

It is hard to apply definition (i) to determine the mode for this sample because the sample is too small in order to construct a meaningful frequency distribution. Thus applying definition (ii), one may conclude that the mode for this sample is 46. Altogether, *ClassSize*=46 has the highest frequency. Suppose however that one has collected the following sample for the same variable:

8, 16, 17, 20, 21, 22, 23, 24, 25, 26, 27, 28, 32, 33, 38, 42, 46, 46, 54

Still, according to definition (ii), the mode would be 46. However, one can question this result after constructing an interval-based frequency distribution:

(0-10]: 1
(10-20]: 2
(20-30]: 9
(30-40]: 3
(40-50]: 3
(50-60]: 1

There is one value above 0 and up to 10, 2 values above 10 and up to 20, 9 values above 20 and up to 30, and so on. If a class is selected at random, the most likely size happens to be between 20 and 30. Since the mode is a single value, it is assigned in this this case to the midpoint of the interval that has the highest frequency. Such a point is in the middle of interval (20, 30]. Thus the mode would be set to 25. So which one is the true mode? Is it 46 or 25? The probability criterion overwhelmingly points to interval (20, 30]. If someone had to make a decision about the capacity of the newly built classroom, most likely, he or she would not use 46 as a hint.

These two examples of quantitative samples show that, with respect to numeric variables, definition (i) and (ii) may produce different outcomes. One can also learn from the second example that larger samples may lead to interval-based frequency distribution for which the mode becomes one of the interval midpoints, if any.

INSTRUCTIONAL GUIDELINES FOR QUANTITATIVE-DISCRETE DATA

Google spreadsheet (Letkowski – Quantitative Mode, 2012) shows a complete solution for a large sample. It also includes detail instruction for constructing an interval-based distribution and for computing the mode. Figure 3 shows the distribution table and calculated mode.

CASE 3 - THE MODE FOR QUANTITATIVE-CONTINUOUS RANDOM VARIABLES

The last data type to be examined is quantitative-continuous. A continuous random variable, X, has a domain, D_X, of real numbers. D_X can be a bounded or unbounded.
A uniformly distributed random variable is bounded by a lower limit, a, and an upper limit, b. Since the probability (density) function is constant, $f(x) = 1/(b-a)$, this variable does not have any mode. It is a modeless variable.
The probability distribution of an exponential random variable is given by the following function (Wikipedia-Exponential, 2012):

$$f(x) = \lambda e^{-\lambda x}, x \geq 0 \qquad \text{(v)}$$

The domain of the variable is bottom-bounded and top- unbounded, $X \in [0,+\infty)$. The highest value of this function occurs at x = 0. Although this value of the mode does not seem to be useful, it represents the neighborhood, having the highest probability. In other words, taking an interval of a fixed width, the one that starts at the mode (x=0) has the highest chance to return a random value.

Each normally-distributed random variable is totally unbounded, $X \in (-\infty,+\infty)$. Its probability distribution is defined by the following bell-shaped function (Wikipedia-Normal, 2012):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < +\infty \qquad \text{(vi)}$$

This function takes on the highest value at $x = \mu$. Thus the mode of this variable is equal to μ. In fact, for this and for any other symmetric distribution, if the mode exists, then the mean, median and mode are the same.

It is important to note that for a continuous random variable, the mode indicates the most likely neighborhood but itself may never occur since for any x, $P(X = x) = 0$. One can also say that the probability function shows the highest density close to the value of the mode (if it exists).

INSTRUCTIONAL GUIDELINES FOR QUANTITATIVE- CONTINUOUS DATA

So how does one find out the mode empirically, given a sample of values $(x_1,x_2,\ldots,x_n)$ selected from a continuous population? Since selecting two or more identical values, from a continuous domain, is unlikely, determining the mode is supposed to be done by means of definition (i). Thus before one can identify the mode, the sample must be processed to produce a frequency distribution. To this end one generates interval limits $(c_0,c_1,c_2,\ldots,c_m)$. Typically, the number, m, of the intervals is expected to be close to the square root of the sample size, $m \approx\sqrt{n}$. It is convenient if the left limit of the first interval, c_0, is slightly smaller than the sample minimum and the interval width, c_{i-1}-c_i, i=1,2,...,m, is slightly larger than the sample range divided by the number of class intervals. Excel or Google spreadsheet can easily generate the absolute frequency distribution, using the *Frequency*() function:

$$\{f(c_0,c_1), f(c_1,c_2),\ldots, f(c_{m-1},c_m)\} = Frequency(x_1,x_2,\ldots,x_n, c_0,c_1,c_2,\ldots,c_m) \qquad \text{(vii)}$$

Frequency() is an array function. It returns multiple values in a range of spreadsheet cells. Its implementation along with the necessary settings and formulas are all presented in a Google spreadsheet (Letkowski– Quantitative Mode, 2012). It is based on the following quantitative-continuous sample:

500.41	399.72	328.38	623.29	438.02	400.09	**255.57**	586.35	511.90	**255.57**	434.00	595.46
602.86	463.49	511.63	475.72	542.70	406.69	368.52	654.34	602.90	556.63	570.40	347.79
500.56	686.21	495.70	526.55	481.08	608.13	532.05	588.92	330.88	463.75	443.42	581.70

As shown in Figure 3, the frequency distribution clearly shows interval (471,543] having the highest absolute frequency, 10. Thus the mode of this sample is the midpoint of this interval, 507. As pointed out by (Pelosi, 2003, p. 121), a continuous sample may have the modal interval rather than the modal value. A midpoint value is simply a reasonable representative of this interval. It is important to note that the spreadsheet function, *Mode*(),when applied directly to the sample, =*Mode*(x_1,x_2,…,x_n), produces a value, 255.57, which occurs in this sample twice. However this value belongs to the least populated interval, (255,327], and as such should not be confused with the actual mode.

CONCLUSIONS

The three "amigos", mean, median and mode, are important statistical summary measures. With respect to the mean, one could say "show me where to compromise". The median says "here is your 50/50 chance value". Being the only controversial measure —the mode attempts to suggest "what to bet on". Together, the measures are used to explain the shape of the probability (frequency) distribution. If the [unique] mode exists and it is equal to both mean and median, the distribution is perfectly symmetric. The farther the mode from the mean, the more skewed the distribution. The Exponential distribution is an example for an extreme skewness.

There are no problems with determining the mode for qualitative (categorical) random variables. Both definition (i) and definition (ii) lead to the same result. One has to be very careful when attempting to find out the mode for a numeric variable. Special attention must be given to the spreadsheet function *Mode*() which rarely provides the correct value. Ideally, the mode should be derived from the probability (frequency) distribution function. Even so the mode is defined as a specific point value, it is important to remember that it represents the neighborhood having the highest concentration of data. Again, definition (i) is the best way to determining the mode, if any!

REFERENCES

Anderson, D. R., Sweeney, D. J., Williams, T. A., (2012), Essentials of Modern Business Statistics with Microsoft® Excel. Mason, OH: South-Western, Cengage Learning.

Black, K. (2012) Business Statistics For Contemporary Decision Making. New York, NY: John Wiley and Sons, Inc.

De Veaux, R. D., Velleman, P. F., Bock D.E. (2006) Intro to Stats. Boston, MA: Addison Wesley, Pearson Education, Inc.

Doane, D., Seward, L. (2010) Applied Statistics in Business and Economics, 3rd Edition, Mcgraw-Hill, 2010.

Donnelly, Jr., R. A. (2012) Business Statistics. Upper Saddle River, NJ: Pearson Education, Inc.

Larose, D.T. (2010) Discovering Statistics. New York: W. H. Freeman Company.

Letkowski, J. (2013) Exploring the Mode - the Most Likely Value? Proceedings of the 2013 Academic and Business Research Institute Conference, Orlando, January 3-5, 2013

Letkowski, J. – Qualitative Mode (2012) Exploring the Mode for Qualitative Data. Retrieved from: https://docs.google.com/spreadsheet/ccc?key=0AsmhQG4y08HcdHVKTWdhNkJRSjdnNTVFVG1wSE1RSmc

Letkowski, J. – Quantitative Mode (2012) Exploring the Mode for Quantitative Data. Retrieved from: https://docs.google.com/spreadsheet/ccc?key=0AsmhQG4y08HcdGxlSXM3Qm5jNzk4aEV6NkxQVzhvTEE

Levine, D. M., Stephan, D.F., Krehbiel, T.C., Berenson, M.L. (2011) Statistics for Managers Using Microsoft® Excel, Sixth Edition. Boston, MA: Prentice Hall, Pearson Education, Inc..

McClave, J. T., Benson, P. G., Sincich, T. (2011) Statistics for Business and Economics, 11th Edition. Boston, MA: Prentice Hall, Pearson Education, Inc.

Pelosi, M. K., Sandifer, T.M. (2003) Elementary Statistics. New York, NY: John Wiley and Sons, Inc.

Sharpie, N. R., De Veaux, R. D., Velleman, P. F. (2010) Business Statistics. Boston, MA: Addison Wesley, Pearson Education, Inc..

Spiegel, M. R., (1975) Probability and Statistics, Schaum's Outline Series in Mathematics. New York, NY: McGraw-Hill Book Company.

Triola, M. F. (2007) Elementary Statistics Using Excel®. Boston, MA: Addison Wesley, Pearson Education, Inc.

Wikipedia-Binomial (2012) Binomial distribution.
Retrieved from: http://en.wikipedia.org/wiki/Binomial_distribution

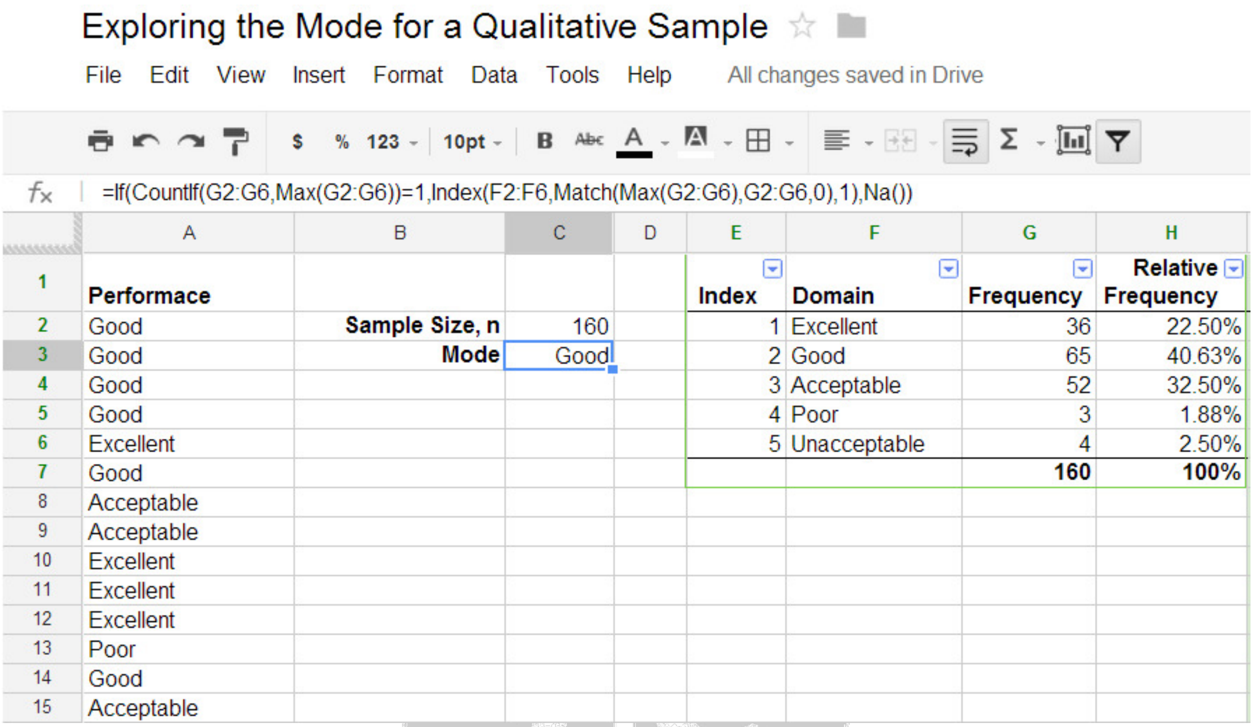
Wikipedia-Exponential (2012) Exponential distribution.
Retrieved from: http://en.wikipedia.org/wiki/Exponential_distribution

Wikipedia-Mode (2012) Mode (statistics).
Retrieved from: http://en.wikipedia.org/wiki/Mode_(statistics)

Wikipedia-Normal (2012) Normal distribution.
Retrieved from: http://en.wikipedia.org/wiki/Normal_distribution

Wikipedia-Poisson (2012) Poisson distribution
Retrieved from: http://en.wikipedia.org/wiki/Poisson_distribution

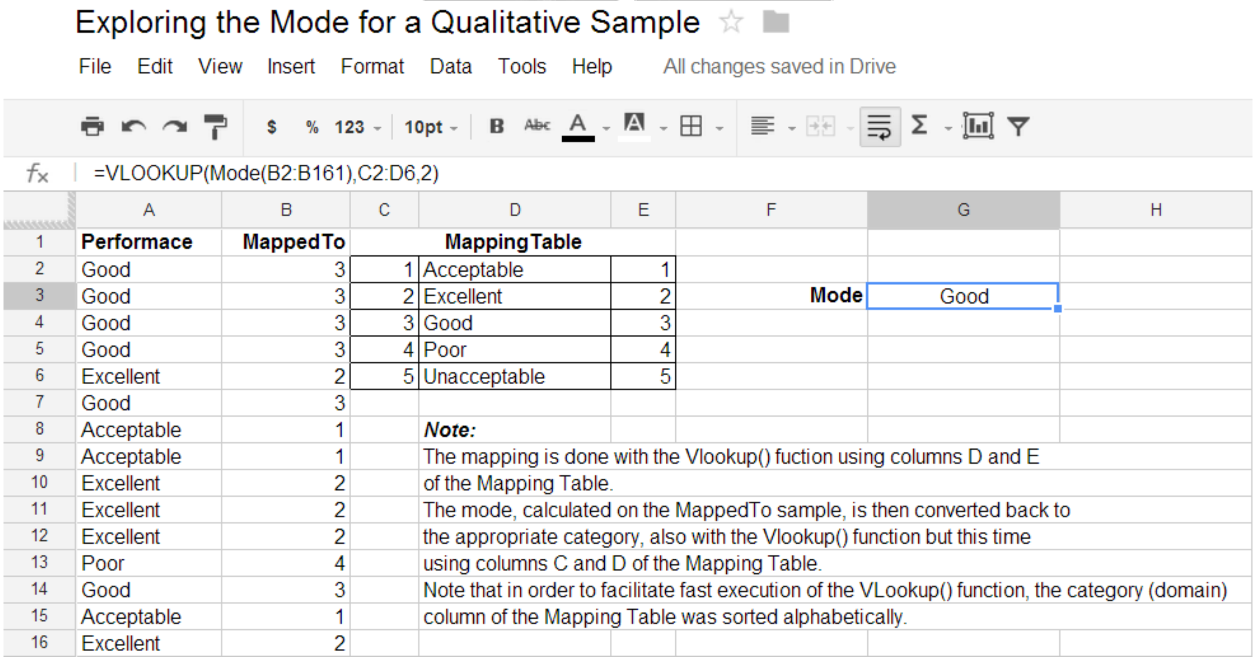
APPENDIX


Exploring the Mode for a Qualitative Sample
File Edit View Insert Format Data Tools Help
All changes saved in Drive
$ % 123 10pt B Abc A Σ

fx =If(CountIf(G2:G6,Max(G2:G6))=1,Index(F2:F6,Match(Max(G2:G6),G2:G6,0),1),Na())

	A	B	C	D	E	F	G	H
1	Performace				Index	Domain	Frequency	Relative Frequency
2	Good	Sample Size, n	160		1	Excellent	36	22.50%
3	Good	Mode	Good		2	Good	65	40.63%
4	Good				3	Acceptable	52	32.50%
5	Good				4	Poor	3	1.88%
6	Excellent				5	Unacceptable	4	2.50%
7	Good						160	100%
8	Acceptable							
9	Acceptable							
10	Excellent							
11	Excellent							
12	Excellent							
13	Poor							
14	Good							
15	Acceptable							

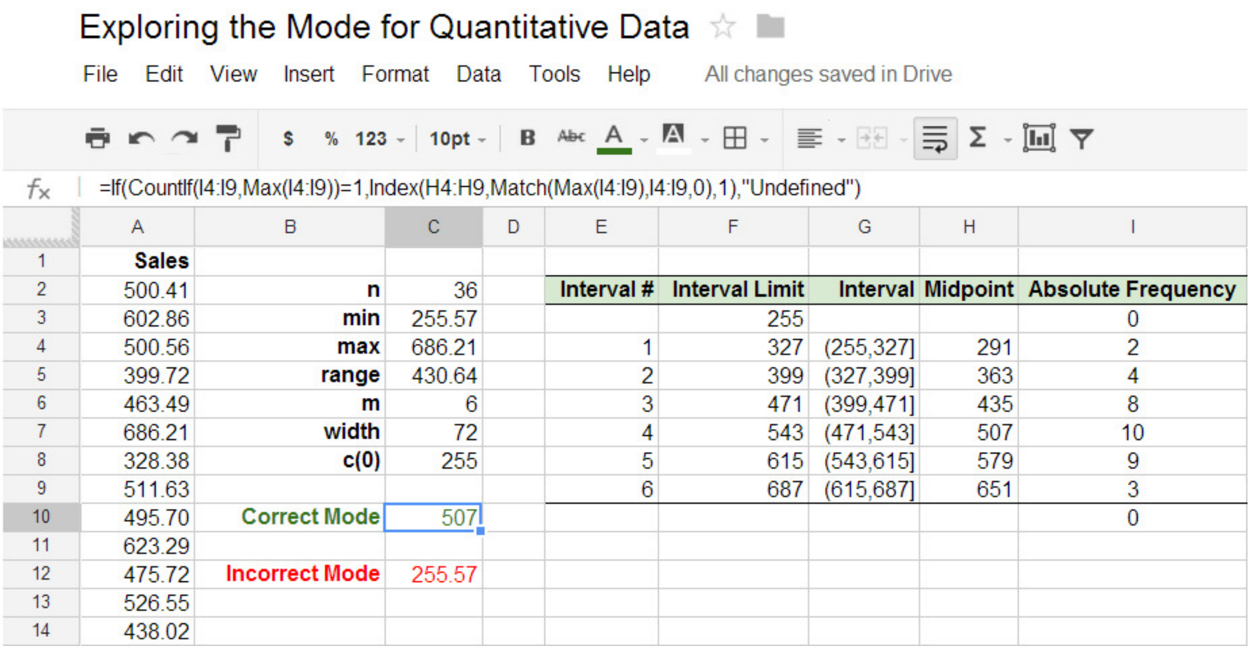
Figure 1 Getting the mode for a qualitative sample via the frequency distribution (Letkowski - Qualitative, 2012).


Exploring the Mode for a Qualitative Sample
File Edit View Insert Format Data Tools Help
All changes saved in Drive
$ % 123 10pt B Abc A Σ

fx =VLOOKUP(Mode(B2:B161),C2:D6,2)

	A	B	C	D	E	F	G	H
1	Performace	MappedTo		MappingTable				
2	Good	3	1	Acceptable	1			
3	Good	3	2	Excellent	2	Mode	Good	
4	Good	3	3	Good	3			
5	Good	3	4	Poor	4			
6	Excellent	2	5	Unacceptable	5			
7	Good	3						
8	Acceptable	1		*Note:*				
9	Acceptable	1		The mapping is done with the Vlookup() fuction using columns D and E				
10	Excellent	2		of the Mapping Table.				
11	Excellent	2		The mode, calculated on the MappedTo sample, is then converted back to				
12	Excellent	2		the appropriate category, also with the Vlookup() function but this time				
13	Poor	4		using columns C and D of the Mapping Table.				
14	Good	3		Note that in order to facilitate fast execution of the VLookup() function, the category (domain)				
15	Acceptable	1		column of the Mapping Table was sorted alphabetically.				
16	Excellent	2						

Figure 2 Getting the mode for a qualitative sample, using the Mode() function (Letkowski - Qualitative, 2012).



Exploring the Mode for Quantitative Data

=If(CountIf(I4:I9,Max(I4:I9))=1,Index(H4:H9,Match(Max(I4:I9),I4:I9,0),1),"Undefined")

	A	B	C	D	E	F	G	H	I
1	**Sales**								
2	500.41	**n**	36		**Interval #**	**Interval Limit**	**Interval**	**Midpoint**	**Absolute Frequency**
3	602.86	**min**	255.57			255			0
4	500.56	**max**	686.21		1	327	(255,327]	291	2
5	399.72	**range**	430.64		2	399	(327,399]	363	4
6	463.49	**m**	6		3	471	(399,471]	435	8
7	686.21	**width**	72		4	543	(471,543]	507	10
8	328.38	**c(0)**	255		5	615	(543,615]	579	9
9	511.63				6	687	(615,687]	651	3
10	495.70	**Correct Mode**	507						0
11	623.29								
12	475.72	**Incorrect Mode**	255.57						
13	526.55								
14	438.02								

Figure 3 Getting the mode for a quantitative sample via the frequency distribution (Letkowski - Quantitative, 2012).