A spreadsheet based derivation of the probability distribution from a random sample

Jerzy Letkowski
Western New England University

ABSTRACT

Spreadsheet programs are frequently used as an alternative to professional statistical packages. Many statistical problems can be quickly and accurately solved in a spreadsheet. All major summary measures, probability functions, tables, charts, etc., are well supported by the contemporary spreadsheet programs. Nonetheless, there are some tasks that require special attention and care in order to ensure proper outcomes. This paper presents a complete case for generating empirical frequency distributions for continuous-numeric data, using a Microsoft Excel compatible spreadsheet program. It shows how to best align the theory and spreadsheet-based practice with respect to production of frequencydistributions. This alignment is implemented by testing the goodness of fit of the empirical frequency distribution to the theoretical probability distribution. A brief survey and analysis of spreadsheet methods, used to generate of frequency distributions, is also provided.

Keywords: statistics, frequency distribution, probability distribution, data types, spreadsheet, test

LEARNING OBJECTIVES

Generally, in business, the purpose of identifying a probability distribution is to improve awareness of an uncertain decision situation. The probability distribution is the most convenient and powerful source of information about some characteristic (variable) of the uncertain situation. In order for students, taking Business Statistics courses, to be able to produce such information they should: know how to capture a sample in a spreadsheet; be aware of different options and issues related to generating frequency distributions; learn how to implement typical spreadsheet procedures for generating frequency distributions; understand strengths and weaknesses of different spreadsheet procedures; be able to select a matching probability distribution based on intuitive (visual) assessment of the [empirical] frequency distributions; and learn how to perform a Goodness-of-Fit test, involving the frequency distribution and the selected probability distribution. This paper is an attempt to satisfy these objectives.

THEORETICAL BACKGROUND

Most of the modern introductory Statistics textbooks start exploring statistics with data-centric topics, for example (Anderson at al., 2012, Black 2012, Donnelly, 2012, Larose, 2010, Levine, et al., 2011, Pelosi at al., 2003, Triola, 2007). Students first learn about the purpose of statistics, including both descriptive statistics and inferential statistics. Next they become familiar with data scope and types. In particular, in terms of the scope, they learn in general about the difference and relationship between a sample and population. Regarding the data types, the students become familiar with properties of and differences between qualitative data and quantitative data. At this stage, even so statistics deals with random data, there are few, if any, references to the probability or to the probability distributions of the data.

Arguably, in the context of statistics, any random data set is best described by its probability distribution. Any information that can be derived directly from a data set (for example, from a sample) can also be derived from the probability distribution. However, probability related topics are usually covered in later chapters, after the students have learned about how to treat data empirically.

Undoubtedly, there is a very strong connection between Statistics and Probability. One could say that Statistics is a playground for Probability or that Probability is a soul for Statistics. Like in other disciplines (such as Physics) there are many good reasons for acquiring at least a rudimentary understanding of the theory before attempting to study applications. Nonetheless, the contemporary textbooks go the other way around. It is out of the scope of this paper to argue which way is better. Suffice to say, failing to observe basic rules of the theory (Probability) may result in errors or imperfections when developing statistical applications.

Many textbooks, covering introduction to Statistics, provide detail instructions about how to construct frequency distributions, using both tabular and graphical representations (Anderson at al., 2012 p. 49-52, Black 2012, p.20-22, Donnelly, 2012, p.30-34, Larose, 2010, p.47-50, Levine, et al., 2011, p.26-30, Pelosi at al., 2003, p.64-67, Triola, 2007, p.). In the same time, students do not necessarily learn from these instructions about the important connections between the frequency distribution derived from a sample and the [usually unknown] probability distribution, representing the population from which the sample was selected.

Much like sample statistics $\bar{x}, \bar{p}$ are used to estimate population parameters μ, p, respectively, one can use cumulative frequency distribution, $F_s(x)$, to estimate cumulative

probability distribution, $F(x)$ as shown in Figure 1. Thus, it is imperative to ensure consistency of the procedure used to determine function $F_s(x)$ with the formal definition of function $F(x)$:

$$F(x) = P(X \leq x), \quad \text{(i)}$$

where X is a continuous, numeric, random variable, representing the population.
Now, consider a sample, X_s, of size n, selected randomly from the population, X:

$$X_s = \{x_1, x_2, \ldots, x_n\}, \quad X_s \subseteq X \quad \text{(ii)}$$

Since function $F(x)$ defines the probability of random variable X taking on any value less than or equal to x, function $F_s(x)$ must have the same interpretation with respect to the sample, X_s:

$$F_s(x) = \#X_s(x)/n, \qquad X_s(x) \subseteq X_s \wedge \forall_{x_i \in X_s(x)} x_i \leq x \quad \text{(iii)}$$

where $\#X_s(x)$ is the cardinality of set $X_s(x)$ which is a subset of sample X_s such that for every number x_i in $X_s(x)$, $x_i \leq x$. A spreadsheet formula to implement (iii) uses standard functions *CountIf*() and *Count*():

$$F_s(x) = CountIf(X_s, \text{"<="\&x})/Count(X_s) \quad \text{(iv)}$$

Count(X_s) returns the sample size, n. *CountIf*(X_s, "<="&x) counts the number of sample values that are less than or equal to limit x.
From definition (i) one can derive an interval-based probability function:

$$f(l_{j-1},l_j) = F(l_j) - F(l_{j-1}) = P(l_{j-1} < X \leq l_j) \quad \text{(v)}$$

The probability of the random variable, X, to be greater than a lower limit, l_{j-1}, and less than or equal to an upper limit, l_j, is the difference between the values of the cumulative probability function for limits l_j and l_{j-1}, respectively.
Analogically, an interval-based frequency function is defined as follows:

$$f_s(l_{j-1},l_j) = F_s(l_j) - F_s(l_{j-1}) = \#X_s(l_j)/n - \#X_s(l_{j-1})/n \quad \text{(vi)}$$

It is important to note that when using any software to implement this equality (vi), it must process the sample according to (v). In other words, for all the intervals (l_{j-1} , l_j] the counting process must be exclusive with respect to the lower limits (l_{j-1}) and—inclusive with respect to the upper limits (l_j). There is a specialized spreadsheet function that fully supports such a counting procedure:

$$= Frequency(X_s, L), \quad \text{(vii)}$$

where X_s is a range, containing the sample values, and L is a range, containing intervals defined by limits (l_0, l_1, l_2, …, l_m). This is an array function as it returns an array of absolute frequencies associated with intervals ($-\infty$,l_0], (l_0, l_1], (l_1, l_2], (l_2, l_3], …, (l_{m-1}, l_m],(l_m,$+\infty$) that are expected to cover the entire domain of variable X. Thus:

$$\{ f_s(-\infty,l_0),\ f_s(l_0,l_1), f_s(l_1,l_2),\ldots, f_s(l_{m-1},l_m), f_s(l_m, +\infty)\} = Frequency(X_s, L)/n \quad \text{(viii)}$$

If the sample minimum is above l_0 ($min(X_s) > l_0$) and the sample maximum is up to l_m ($max(X_s) \leq l_m$) then the open-ended intervals, ($-\infty$,l_0], (l_m,$+\infty$), are not accounted for. The empirical frequencies associated with these intervals default to zero.

GENERATING AND EXPLORING EMPIRICAL FREQUENCY DISTRIBUTION FOR NUMERIC-CONTINUOUS DATA

There are many methods to produce an empirical frequency distribution in a spreadsheet. This case shows two methods: a *CountIf Method* and a *Frequency Method*. The former uses spreadsheet function *CountIf*() to first generate the cumulative frequency distribution (iv) and the latter—*Frequency*() to produce the [interval] frequency distribution (vi). Both the methods are

compatible with the theoretical definition of the probability distribution (as shown in the previous section).

The critical part of any method is determining the interval limits, (l_0, l_1, l_2, …, l_m). *Microsoft Excel* and *Google Spreadsheet* refer to these limits as a *Bin Range* and *Classes*, respectively. Typically, the limits are defined incrementally. Three parameters are required in order to define the intervals: m (the number of the intervals), w (the common width of the intervals) and l_0 (the left limit of the first interval). A good starting value for parameter m is a number close to $\sqrt{n}$ (Pelosi 2003, p. 64) or close to $log_2(n)$ (Donnelly 2012, p.30):

$$m \approx \sqrt{n} \quad \text{or} \quad m \approx log_2(n) \tag{ix}$$

where n is the sample size. In general it is recommended that m be not too small and not too large, for example $5 \le m \le 15$ (Black 2012, p.20, Levine et al., 2011, p. 26) or $5 \le m \le 20$ (Anderson et al., 2012, p. 48, Pelosi et al., 2003, p. 64). Overall, the final result, the frequency distribution, is the ultimate judge. If there are empty of poorly populated intervals, m should be reduced. On the other hand, if some intervals seem to be overcrowded, m should be increased.

Once parameter m has been determined, the width of the intervals, w, can be set close to the sample range divided by the number of the intervals, m:

$$w \approx [max(X_s) - min(X_s)] / m \tag{x}$$

In most situations width w works well when its value is slightly greater than $[max(X_s) - min(X_s)] / m$. Wherever possible, the interval width, w, should be set to a *friendly* or *convenient* value.

The left limit, l_0, of the first interval is expected to be set close to the sample minimum, $min(X_s)$. If the limit is equal to or greater than the minimum, then the first interval will become open-ended, $(-\infty,l_0]$. Such a solution is recommended particularly if the resulting frequency distribution is going to be scrutinized by a [theoretical] probability distribution. Otherwise, setting the limit to a friendly or convenient value below $min(X_s)$ may work just fine.

Having the three parameters (m, w and l_0) defined, the limits of the intervals can be defined recursively as follows:

$$l_j = l_{j-1} + w, \quad j=1,2,\ldots,m \tag{xi}$$

A Google spreadsheet application, accessible via (Letkowski, 2013) shows how to process a sample to construct and analyze a frequency distribution, using both the *Frequency Method* and *CountIf Methods*. Additionally, the two methods utilize different interval setups:

Setup 1: (l_0, l_1], (l_1, l_2], (l_2, l_3], …, (l_{m-1}, l_m]

Setup 2: $(-\infty,l_0]$, (l_0, l_1], (l_1, l_2], (l_2, l_3], …, (l_{m-1}, l_m],(l_m,$+\infty$)

The second setup includes two open-ended intervals. It covers the entire domain of a Normal probability distribution, $(-\infty,+\infty)$.

Implementing the Frequency Method With Setup 1

Figure 2 shows a fragment of the sheet labeled as *FrequencyMethod*. Range A2:A626 contains a sample selected randomly from a Normal distribution with μ=1000 and σ=160. The following instruction provides all necessary formulas. It is derived from a procedure, involving the *Frequency*() function, presented in (Pelosi et al., 1998, p. 103-117).

Step 1: Setting up the intervals:

a) Cell C2: =Count(A2:A626) -- sample size, *n*.
b) Cell C4: =Log(C2,2) -- suggested number of intervals
c) Cell C5: 10 -- accepted number of intervals, *m*
d) Cell C7: =Min(A2:A626) -- sample minimum
e) Cell C8: =Max(A2:A626) -- sample maximum
f) Cell C9: =C8-C7 -- sample range
g) Cell C10: =C9/C5 -- suggested interval width, *w*
h) Cell C11: 80 -- accepted interval width
i) Cell C13: 600 -- left limit of the first interval, l_0

Note that the suggested number of the intervals is 9.29 (cell C4). It is reasonable to use $m = 10$ intervals. The suggested width of the intervals is 76.3195. The selected value $w = 80$ is both close and friendly. Finally, the starting point (left limit of the first interval) is chosen as $l_0 = 600$, a value slightly smaller than the sample minimum (630.4197). Since the right limit of the last interval, $l_m = 600 + 10 * 80 = 1{,}400$, is greater than the sample maximum (1,393.6150), the entire sample is covered by the intervals. Thus, with such a setup, the open-ended intervals, $(-\infty,l_0]$, $(l_m,+\infty)$, are empty. It is important to note that the resulting intervals do not cover the entire theoretical domain of population (variable) *X*. Thus, they are not fully compatible with the theoretical domain.

Step 2: The three parameters, *m*, *w*, l_0, defined in the previous step, are used to construct the interval limits:

a) Range E4:E14, labeled as *j*, contains a sequence of indexes, $j = 0, 1, 2, \ldots, m$, in this case: 0,1,2,3,4,5,6,7,8,9,10.
b) Range F4:F14, labeled as *bin*, defines the interval limits, l_j, $j = 0, 1, 2, \ldots, m$,
   Cell F4: =C13 (l_0)
   Cell F5: =F4+$C$11 ($l_0 + w$)
   Range F6:F14:
   Copy the formula in cell F5 and paste it to range F6:F14.
c) Range G5:G14, labeled as *interval*, contains the intervals
   $(l_0, l_1]$, $(l_1, l_2]$, $(l_2, l_3]$, …, $(l_{m-1}, l_m]$
   Cell G5: ="("&F4&","&F5&"]" ($(l_0, l_1]$)
   Range G6:G14:
   Copy the formula in cell G5 and paste it to range G6:G14.

Step 3: It is now a good time to generate the frequency distribution, using formula (vii) by first computing the absolute frequencies:

a) Select range H4:H15, labeled as *fsn*() *absolute frequency*, type formula =FREQUENCY(A2:A626,F4:F14), hold down keys Shift+ Ctrl and press Enter. This array-based formula will fill the H4:H15 range with absolute frequencies associated with all the intervals, including the open-ended intervals, $(-\infty,l_0]$ and $(l_m,+\infty)$. As expected the open-ended intervals contain no data, since $l_0 > min(X_s)$ and $l_m < max(X_s)$. The absolute frequency distribution, $f_{sn}()$, is then used to define the frequency distribution in the range I4:I15, labeled as *fs*() *frequency*.
b) Cell I4: =H4/$C$2 ($f_{sn}(-\infty,l_0)$/n)

c) Range I5:I14:
Copy the formula in cell I4 and paste it to range I5:I14.

d) Range J4:J14:
Based on formula (vi) one can develop formulas for the cumulative frequency,
$F_s(l_0) = f_s(-\infty,l_0)$
$F_s(l_j) = F_s(l_{j-1}) + f_s(l_{j-1},l_j)$, for $j = 1, 2, \ldots, m$.
Cell J4: =I4 ($F_s(l_0) = f_s(-\infty,l_0)$)
Cell J5: =J4+I5 ($F_s(l_1) = F_s(l_0) + f_s(l_0,l_1)$)
Range J6:J14:
Copy the formula in cell J6 and paste it to range J6:J14.

Examining visually the frequency distribution shown in Figure 2, one can formulate a hypothesis (*Null* Hypothesis, H_o) that the sample, based on which the distribution was derived, came from a Normal population. The alternative hypothesis (Hypothesis, H_A) would assert that the sample does not come from the Normal population.

Testing the Goodness of Fit (Setup 1)

Indeed, as mentioned, the sample used in this case comes from a Normal population with $\mu = 1000$ and $\sigma = 160$. Can the [empirical] frequency distribution confirm it? This question can be answered using the Pearson's Chi-squared test (Wikipedia, 2012). This test's statistic combines the total of relative squared differences between observed absolute frequencies and the expected frequencies, including the open-ended intervals in which no sample data exist:

$$\chi^2 = \frac{(f_{sn}(-\infty,l_0) - f_n(-\infty,l_0))^2}{f_n(-\infty,l_0)} + \frac{(f_{sn}(l_0,l_1) - f_n(l_0,l_1))^2}{f_n(l_0,l_1)} + \ldots + \frac{(f_{sn}(l_{m-1},l_m) - f_n(l_{m-1},l_m))^2}{f_n(l_{m-1},lm)} + \frac{(f_{sn}(l_m,\infty) - f_n(l_m,\infty))^2}{f_n(l_m,\infty)} \quad \text{(xii)}$$

The observed (empirical) frequencies, $f_{sn}(l_{j-1},l_j)$, $j = 0, 1, 2, \ldots, m+1$, come from the frequency distribution (this is column $f_{sn}()$ in the frequency distribution table shown in Figure 2). The expected (theoretical) frequencies come from the Normal distribution. Both Microsoft Excel and Google Spreadsheet calculate the Normal probabilities using the =*NormDist*(x, μ, σ,*True*) function. Thus:

$f_n(-\infty,l_0) = n\,f(-\infty,l_0) = nP(X \le l_0)$ = n*NormDist(l_0, μ, σ,*True*)
$f_n(l_{j-1},l_j) = n\,f(l_{j-1},l_j) = nP(l_{j-1} < X \le l_j)$ =
=n*(NormDist(l_j, μ, σ,*True*)- NormDist(l_{j-1}, μ, σ,*True*)), j=1,2, … ,m,
$f_n(l_m,+\infty) = n\,f(l_m,+\infty) = nP(X > l_m)$ = n*(1-NormDist(l_m, μ, σ,*True*)) (xiii)

Using a confidence level of 1-α = 95% (or significance level of α = 5%), if the test statistic, χ^2, does not exceed the critical value, χ_c^2, then there is no reason to reject the *Null* Hypothesis, H_o. Since Google Spreadsheet does not support the Chi-square distribution, the critical value is imported from the following PHP service:

http://doingstats.com/srv/chsqr.php?df="&`ref1`&"&alpha="&`ref2` (xiv)

where `ref1` is equal to the number of the intervals minus 1 (the degrees of freedom, *df*) and `ref2` is the significance level (α). In Microsoft Excel, this critical value can be computed directly, using one of the following formulas:

=*CHISQ.INV*(1-α,*df*) or =*CHIINV*(α,*df*) (xv)

Figure 3 and the following instruction show how to implement the test.

Step 1: Compute the test statistic, χ^2, value:

a) Range C17:C18:
Enter values 1000 and 160 for the mean, μ, and the standard deviation, s, of the Normal distribution, respectively.

d) Range E18:E29, labeled as *j*, contains a sequence of indexes, $j = 0, 1, 2, \ldots, m,m+1$, in this case: 0,1,2,3,4,5,6,7,8,9,10,11.

e) Range F18:F29:
Link (copy by linking) this range to the original absolute frequencies, H4:H15,
Cell F18: =H4
Copy the formula in cell F18 and paste it to range F19:F29.

f) Range G18:G29:
Implement formulas (xiii) in order to compute the expected frequencies,
Cell G18: =C2*NORMDIST(F4,C17,C18,TRUE)
Cell G19: =$C$2*(NORMDIST(F5,$C$17,$C$18,True)-NORMDIST(F4,$C$17,$C$18,True))
Range G20:G28:
Copy the formula in cell G19 and paste it to range G20:G28,
Cell G29: =C2*(1-NormDist(F14,C17,C18,True))

g) Range H18:H29:
Calculate the relative squared difference between the observed frequencies and the expected frequencies,
Cell H18: =(G18-F18)^2/G18
Range H19:H29:
Copy the formula in cell H18 and paste it to range H19:H29.

h) Cell G31: Calculate the total relative squared differences, χ^2.
=SUM(H18:H30)

Step 2: Perform the test:

a) Cell C24:
Determine the degrees of freedom ($df = m+2-1$). Notice that two open-ended intervals (rows) have been added and *m* is here interpreted as the number of bounded intervals, covering the entire sample. The value of *df* is the same as the number of applied frequencies minus 1:
=Count(F18:F29)-1

b) Cell C25:
Enter the significance level, α,
0.05

c) Cell C26:
Calculate the critical value, χ_c^2 (Critical χ^2),
=ImportData("http://doingstats.com/srv/chsqr.php?df="&C24&"&alpha="&C25)
Note that, in Excel, you would use, for example, this formula:
=*CHIINV*(C25,C24)

d) Cell C27:
Link-copy the test statistic value, χ^2(Observed χ^2),
=H31

e) Range B29:B30:
If χ_c^2 (Critical χ^2) is below χ^2 then fail to reject H_o.

As one can see this is a close call. The [empirical] test statistic, χ^2 = 19.5938 and the critical value χ_c^2=19.6751. Nonetheless there is no reason to reject the *Null* Hypothesis. A similar analysis will be performed with a slightly different setup shown in the next section.

Implementing the CountIf Method with Setup 2

This instruction is very similar to the previous one. Significant differences include the construction of the intervals and the way the absolute frequencies are calculated. The Goodness of Fit test is identical. Figure 4 shows the setup for the tabular representation of the [empirical] frequency distribution.

Step 1: Setting up the intervals:

a) Cell C2: `=Count(A2:A626)` -- sample size, *n*.
b) Cell C4: `=Log(C2,2)` -- suggested number of intervals
c) Cell C5: `10` -- accepted number of intervals, *m*
d) Cell C7: `=Min(A2:A626)` -- sample minimum
e) Cell C8: `=Max(A2:A626)` -- sample maximum
f) Cell C9: `=C8-C7` -- sample range
g) Cell C10: `=C9/C5` -- suggested interval width, *w*
h) Cell C11: `80` -- accepted interval width
i) Cell C13: `700` -- left limit of the first interval, l_0

Note that the suggested number of the intervals is 9.29 (cell C4). It is reasonable to use *m* = 10 intervals. The suggested width of the intervals is 76.3195. The selected value *w* = 80 is both close and friendly. Finally, the starting point (left limit of the first interval) is chosen as l_0 = 700, a value significantly larger than the sample minimum (630.4197). Since the right limit of the last interval, l_m = 700 + (10-2) * 80 = 1,340, is below the sample maximum (1,393.6150), the entire sample is not covered by the closed intervals. Thus, the open-ended intervals, $(-\infty,l_0]$, $(l_m,+\infty)$, are not empty. It is important to note that all the intervals cover the entire theoretical domain of population (variable) *X*. Thus, they are fully compatible with the theoretical domain.

Step 2: The three parameters, *m*, *w*, l_0, defined in the previous step, are used to construct the interval limits. All the formulas are identical to those shown in the previous instruction (**Step 2** for section **Implementing the Frequency Method**). One small difference is related to the number of rows in the frequency table. Since the open-ended intervals are here explicitly included, the total number of rows should be equal to *m*, all indexed as 0, 1, 2, …, *m*-1. Notice also the two open-ended intervals shown in the ***interval*** column: **(-inf, 700]** and **(1340,+inf)**, where **inf** stands for infinity (∞). They are to be entered manually.

Step 3: Using the *CountIf*() function, it is more convenient to first compute the absolute, cumulative frequencies. Therefore this instruction start with column labeled as ***Fsn***(), representing the absolute cumulative frequencies. All other frequencies can be derived from ***Fsn***().

a) Range J4:J13:
In this range, all the formulas, except for the last one, have the same structure, referring to the [common] sample, *X*s, contained in range $A$2:$A$626, and to the variable limits defined in column ***bin***. All the formulas in J4:J12 do the counting of sample values up to the interval limits, l_j, j=0,1,2, …, *m*-1, in column ***bin***. The last formula (in J13) is supposed to combine all the counting up to the last limit, l_{m-1}, plus the counting result about that limit.
Cell J4: =CountIf($A$2:$A$626,"<="&F4)
Range J5:J12:
Copy the formula in cell J4 and paste it to range J5:J12,
Cell J13: =J12+CountiIf($A$2:$A$626,">"&F12)

b) Range H4:H13:
This range defines the absolute [interval] frequencies, labeled as ***fsn***(). The first entry, H4, representing the count of sample value in interval $(-\infty,l_0]$ is the same as column ***Fsn***(). Each subsequent count is the difference between two consecutive cumulative counts. It implements a formula that is similar to (vi):
$f_{sn}(l_{j-1},l_j) = F_{sn}(l_j) - F_{sn}(l_{j-1}) = \#X_s(l_j) - \#X_s(l_{j-1})$
Cell H4: =J4
Cell H5: =J5-J4
Range H6:H13:
Copy the formula in cell H5 and paste it to range H6:H13.

c) The remaining columns, ***fs***() and ***Fs***(), representing the frequency and cumulative frequency distributions, respectively, are calculated in a similar way to those on sheet *FrequencyMethod*.

The resulting frequency distribution for this version of the intervals appears to be tighter and smoother.

Testing the Goodness of Fit (Setup 2)

The hypotheses remain the same as in the previous version. The outcome is, however, quite different. Figure 5 shows the details. This time the measure of the difference between the frequency distribution $f_s()$ and the Normal distribution is quite small (Observed χ^2= 4.2154). When confronted with the critical value (Critical χ^2 = 16.9190) it gives very strong evidence to support Normality of the sample.

OTHER SPREADSHEET METHODS

The *CountIf*() and *Frequency*() functions are not the only spreadsheet means for generating frequency distributions. The *CountIf*() function provides the most capable but, in the same time, the most complicated solution for this job. It can handle all data types and all interval settings for the numeric data types. The *Frequency*() function delivers probably the most elegant and straightforward solution. However, it can only handle numeric data types.

Microsoft Excel is equipped with two other tools that can be used to produce frequency distributions: the Data Analysis ad-in's *Histogram* command and the *Pivot Table* command.

The *Histogram* command can be used to generate frequency distributions for quantitative data. If the class intervals are not supplied, this command will define its own intervals, using an

open-ended setup, where the number of intervals is close to the square root of the sample size. This command can automatically generate the histogram (column-chart).

The *Pivot Table* command is cool but it is often misused. It must not be used for processing quantitative data as it is not compatible for the definition of the probability distribution. Figure 7 shows, side-by-side, outcomes for processing the same sample using the Pivot Table (Anderson at al., 2012, p. 51) and applying the *Frequency*() function, where the latter gives correct results. The sample consists of the following numbers: 12,13,14,14,15,15,16,17,18,18,18,19,20,21,22,22,23,27,28,33, representing Audit Time. Even so it is a continuous variable, the Pivot table generated a strange sequence of class interval: [10-14], [15-19], [20-24], [25-29], [30-34]. All the interval limits are inclusive and there are gaps between upper limit of the preceding interval and the lower limits of the succeeding intervals. Such a setup is not compatible with the definition of the probability distribution. The Pivot Table command should be avoided when processing quantitative data. It does a good job with handling qualitative data.

CONCLUSIONS

The case presented in this paper uses a randomly generated sample derived programmatically from a Normal population. The sample size is quite large (n = 625). One would expect without any formal testing that the observed frequency distribution is Normal. Figure 6 shows two frequency histograms obtained from the two frequency distributions. One can clearly see that the histogram for Setup 2, including the open-ended intervals, is almost perfectly bell-shaped. It may be a coincident. Nonetheless, one obvious advantage of Setup 2 is that it is 100% consistent with the theoretical domain of the [Normal] population $(-\infty,+\infty)$.

Since different interval settings provide different [empirical] frequency distributions, an obvious question is: which one is of the best quality? Many Statistics textbooks suggest visual assessment of this quality, focusing mainly on *smoothness* of the related histograms, avoiding empty intervals, breaking-down overcrowded intervals, etc. As shown in this paper, the ultimate judgment in assessing this quality can be based on the observed value of the χ^2 measure. Smaller measures of χ^2 provide better fit to the [theoretical] probability distribution. Whether or not a full optimization of the interval settings is performed, it is important to remember that each study of a frequency distribution should consider many different settings of the intervals. Ideally, the final selection should be based on the Goodness of Fit test.

If the test supports the fit, then assessing relevant probabilities can be simplified by utilizing the theoretical probability distribution. In other word, the distribution can serve as a good probability model.

When choosing a spreadsheet tool to generate the frequency distribution, leading the way to the probability distribution, one should consider applying: function *CountIf*(), function *Frequency*() or command *Histogram*.

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.

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) Frequency Distribution For Bullet Speed. Retrieved from: https://docs.google.com/spreadsheet/ccc?key=0AsmhQG4y08HcdEx2MTlEc0RKSzJiNHdqN0hsU3I1Y3c&usp=sharing
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..
Pelosi, M. K., Sandifer, T.M. (2003) Elementary Statistics. New York, NY: John Wiley and Sons, Inc.
Pelosi, M. K., Sandifer, T.M., Letkowski, J. (1998) Doing Statistics with Excel 97,Software Instruction and Exercise Activity Supplement. New York, NY: John Wiley & Sons, Inc.
Triola, M. F. (2007) Elementary Statistics Using Excel®. Boston, MA: Addison Wesley, Pearson Education, Inc.
Wikipedia-Normal (2013) Pearson's chi-squared test. Retrieved from: http://en.wikipedia.org/wiki/Goodness_of_fit

APPENDIX

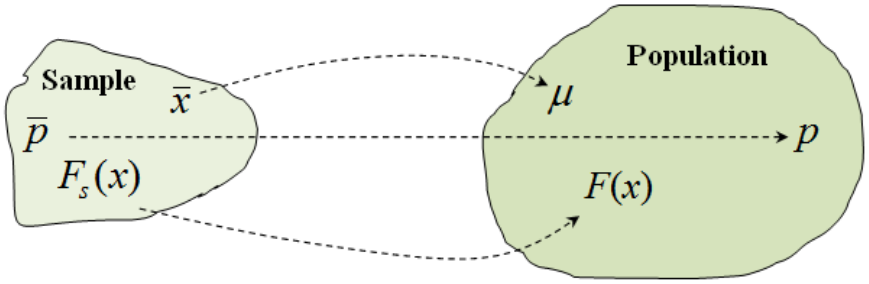


Figure 1. From a sample to a population.

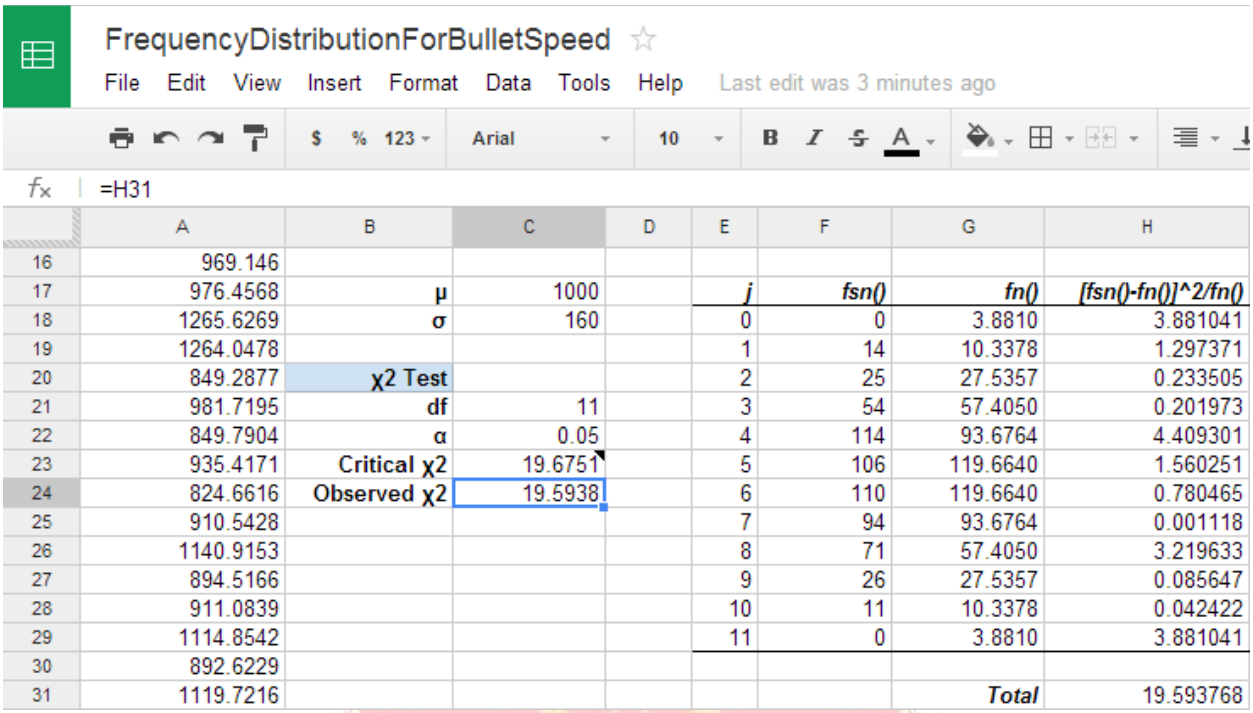
FrequencyDistributionForBulletSpeed

File Edit View Insert Format Data Tools Help All changes saved in Drive

=ArrayFormula(FREQUENCY(A2:A626,F4:F14))

	A	B	C	D	E	F	G	H	I	J
1	Xs							*fsn()*	*fs()*	*Fs()*
2	870.2568	n	625					*absolute*		*cumulative*
3	863.9796				*j*	*bin*	*interval*	*frequency*	*frequency*	*frequency*
4	1035.9525	log2(n)	9.29		0	600		0	0	0
5	1351.4137	m	10		1	680	(600,680]	14	0.0224	0.0224
6	706.8938				2	760	(680,760]	25	0.0400	0.0624
7	758.6824	min	600.3044		3	840	(760,840]	54	0.0864	0.1488
8	1082.109	max	1,399.8166		4	920	(840,920]	114	0.1824	0.3312
9	1189.2267	range	799.5122		5	1000	(920,1000]	106	0.1696	0.5008
10	982.5707	range/m	79.95122		6	1080	(1000,1080]	110	0.1760	0.6768
11	1038.0772	w	80		7	1160	(1080,1160]	94	0.1504	0.8272
12	1175.8826				8	1240	(1160,1240]	71	0.1136	0.9408
13	1074.3864	lo	600		9	1320	(1240,1320]	26	0.0416	0.9824
14	1150.3335				10	1400	(1320,1400]	11	0.0176	1
15	1278.9678							0		

Figure 2 Computing a frequency distribution in a Google Spreadsheet, using function *frequency*().



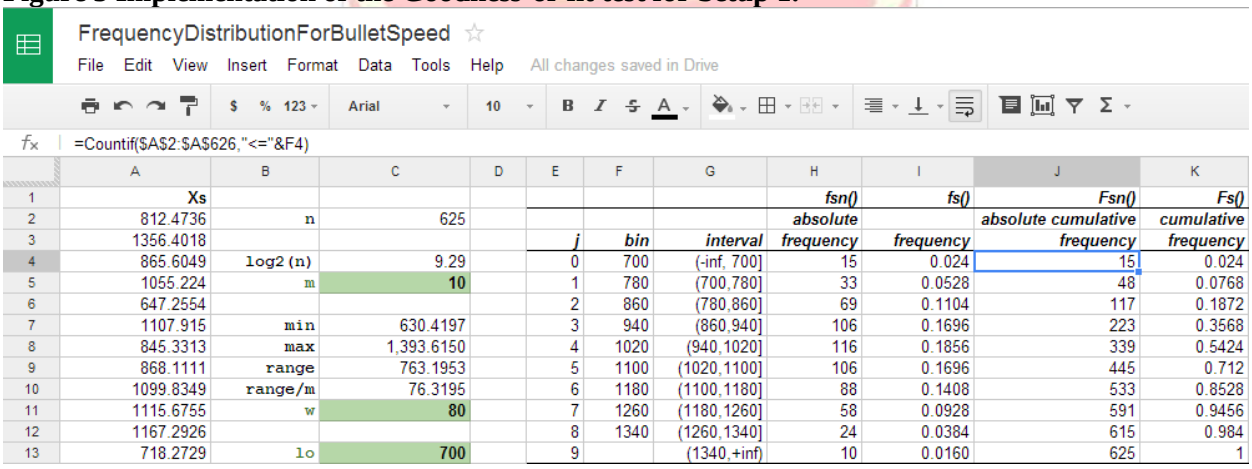
FrequencyDistributionForBulletSpeed

File Edit View Insert Format Data Tools Help Last edit was 3 minutes ago

=H31

	A	B	C	D	E	F	G	H
16	969.146							
17	976.4568	μ	1000		*j*	*fsn()*	*fn()*	*[fsn()-fn()]^2/fn()*
18	1265.6269	σ	160		0	0	3.8810	3.881041
19	1264.0478				1	14	10.3378	1.297371
20	849.2877	χ2 Test			2	25	27.5357	0.233505
21	981.7195	df	11		3	54	57.4050	0.201973
22	849.7904	α	0.05		4	114	93.6764	4.409301
23	935.4171	Critical χ2	19.6751		5	106	119.6640	1.560251
24	824.6616	Observed χ2	19.5938		6	110	119.6640	0.780465
25	910.5428				7	94	93.6764	0.001118
26	1140.9153				8	71	57.4050	3.219633
27	894.5166				9	26	27.5357	0.085647
28	911.0839				10	11	10.3378	0.042422
29	1114.8542				11	0	3.8810	3.881041
30	892.6229							
31	1119.7216						*Total*	19.593768

Figure 3 Implementation of the Goodness-of-fit test for Setup 1.



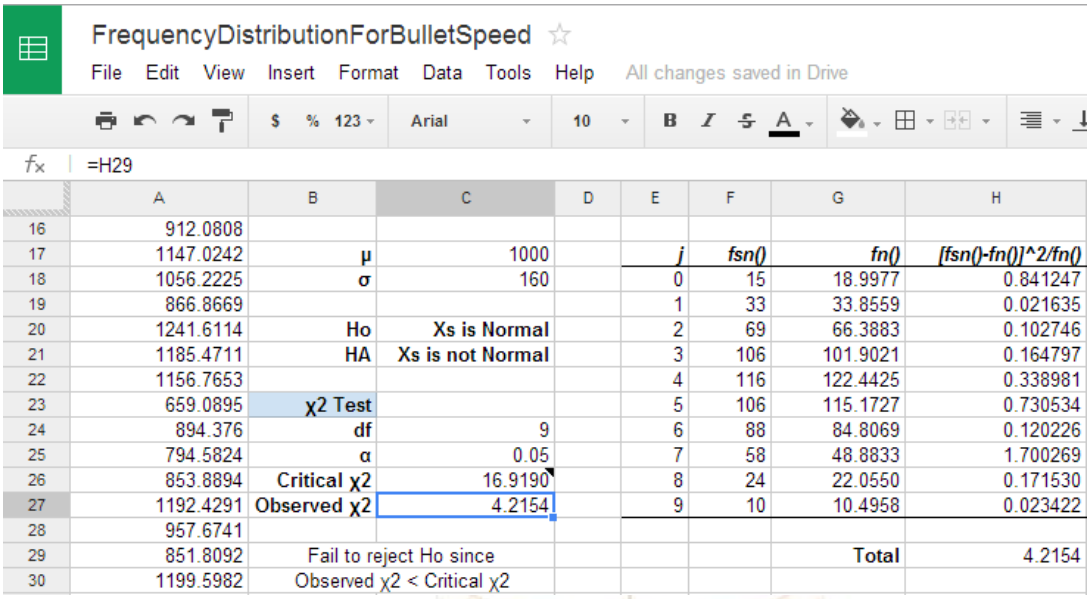
FrequencyDistributionForBulletSpeed

File Edit View Insert Format Data Tools Help All changes saved in Drive

=Countif($A$2:$A$626,"<="&F4)

	A	B	C	D	E	F	G	H	I	J	K
1	Xs							*fsn()*	*fs()*	*Fsn()*	*Fs()*
2	812.4736	n	625					*absolute*		*absolute cumulative*	*cumulative*
3	1356.4018				*j*	*bin*	*interval*	*frequency*	*frequency*	*frequency*	*frequency*
4	865.6049	log2(n)	9.29		0	700	(-inf, 700]	15	0.024	15	0.024
5	1055.224	m	10		1	780	(700,780]	33	0.0528	48	0.0768
6	647.2554				2	860	(780,860]	69	0.1104	117	0.1872
7	1107.915	min	630.4197		3	940	(860,940]	106	0.1696	223	0.3568
8	845.3313	max	1,393.6150		4	1020	(940,1020]	116	0.1856	339	0.5424
9	868.1111	range	763.1953		5	1100	(1020,1100]	106	0.1696	445	0.712
10	1099.8349	range/m	76.3195		6	1180	(1100,1180]	88	0.1408	533	0.8528
11	1115.6755	w	80		7	1260	(1180,1260]	58	0.0928	591	0.9456
12	1167.2926				8	1340	(1260,1340]	24	0.0384	615	0.984
13	718.2729	lo	700		9		(1340,+inf)	10	0.0160	625	1

Figure 4 Computing a frequency distribution in a Google Spreadsheet, using function *countIf*()



FrequencyDistributionForBulletSpeed

File Edit View Insert Format Data Tools Help All changes saved in Drive

=H29

	A	B	C	D	E	F	G	H
16	912.0808							
17	1147.0242	μ	1000		*j*	*fsn()*	*fn()*	*[fsn()-fn()]^2/fn()*
18	1056.2225	σ	160		0	15	18.9977	0.841247
19	866.8669				1	33	33.8559	0.021635
20	1241.6114	Ho	Xs is Normal		2	69	66.3883	0.102746
21	1185.4711	HA	Xs is not Normal		3	106	101.9021	0.164797
22	1156.7653				4	116	122.4425	0.338981
23	659.0895	χ2 Test			5	106	115.1727	0.730534
24	894.376	df	9		6	88	84.8069	0.120226
25	794.5824	α	0.05		7	58	48.8833	1.700269
26	853.8894	Critical χ2	16.9190		8	24	22.0550	0.171530
27	1192.4291	Observed χ2	4.2154		9	10	10.4958	0.023422
28	957.6741							
29	851.8092	Fail to reject Ho since					Total	4.2154
30	1199.5982	Observed χ2 < Critical χ2						

Figure 5 Implementation of the Goodness-of-fit test for Setup 2.

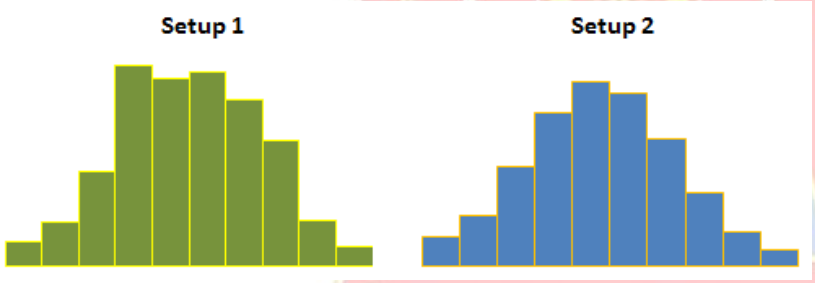


Figure 6 Comparing the histograms derived from the frequency distributions resulting from different interval settings.

H5 {=FREQUENCY(A2:A21,F5:F10)}

	A	B	C	D	E	F	G	H
1	Audit Time					Using the *frequency* () Function		
2	12		Using Pivot Table					Frequency
3	13		Row Labels	Count of Audit Time		Bin	Interval	*fsn* ()
4	14		10-14	4		-∞		
5	14		15-19	8		10	(-∞,10]	0
6	15		20-24	5		15	(10,15]	6
7	15		25-29	2		20	(15,20]	7
8	16		30-34	1		25	(20,25]	4
9	17		Grand Total	20		30	(25,30]	2
10	18					35	(30,35]	1
11	18					+∞	(35, +∞]	0
12	18							
13	19							

Figure 7 Pivot Table vs. function *Frequency*().