# Descriptive and Inferential Statistics

Chanoknath Sutanapong★ & Louangrath, P.I. ★★

**About the author**

★Chanoknath Sutanapong is an independent researcher. She may be reached by email at: chanoknath.sutanapong@gmail.com

★★ Louangrath, P.I. is an Assistant Professor in Business Administration at Bangkok University, Bangkok, Thailand. He could be reached by email at: Lecturepedia@gmail.com

**ABSTRACT**

This paper introduces two basic concepts in statistics: (i) descriptive statistics and (ii) inferential statistics. Descriptive statistics is the statistical description of the data set. Common description include: mean, median, mode, variance, and standard deviation. Inferential statistics is the drawing of inferences or conclusion based on a set of observations. These observations had been described by the descriptive statistics. From these descriptive statistics, an inference is made subject to a predefined limit or error or confidence interval. The error in concluding the inference is called inferential error. There are two types of inferential errors: (i) Type I error and (ii) Type II error. Type I error occurs when the researcher accepts the alternative hypothesis despite contrary evidence. Type II evidence occurs when the researcher rejects the alternative hypothesis despite supporting evidence.

**Keywords:** Descriptive statistics, inferential statistics, sample size

**CITATION:**

Sutanapong, C. and Louangrath, P.I. (2015). "Descriptive and Inferential Statistics" *Inter. J. Res. Methodol. Soc. Sci.*, Vol., 1, No. 1: pp. 22-35. (Jan. –Mar. 2015).

## 1.0 INTRODUCTION

### 1.1 Population

Population is defined as a finite totality of the data set. Generally, population size is given by the symbol $N$. Statistics is commonly defined as a population study. There are two ways that a population may be studied: (i) the entire population may be studied in detailed, i.e. census study where every head is counted, or (ii) sampling a portion of the population and make an inference from the descriptive statistic obtained from the sample. A census study is generally not economically feasible if the population is large. Sampling is a common form of population study.

### 1.2 Sample

Sample $n$ is a portion of a population $N$ where $n \ni N : (n < N)$. A sample is taken from a population for the purpose of learning the characteristics of the population through estimation. The estimation made from a sample is called an inference. The inference is made from the basic descriptive statistic

of a sample. These descriptions include: (i) sample size, (ii) sample mean, (iii) sample variance, and (iv) sample standard deviation. Relevant to the discussion of "sample" is the idea of randomness; hence, random sampling. The issue of randomness is not insignificant and deserves due attention and independent treatment.

**1.3 Sample size and the misuse of the Yamane equation**

Sample size is the size or count of the sample elements taken from the population for the purpose of ascertaining the descriptive statistic. The descriptive statistic obtained from the sample would allow the researcher to make an inference (rationale conclusion) about the population through inferential statistics. Sample size is commonly denoted with a symbol: $n$.

There has been a misunderstanding on how to calculate sample size by using the standard error equation. The standard error equation is given by:

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \text{(A)}$$

…where $\sigma$ is the estimated population standard deviation and $n$ is the size of the test sample. A "test sample" is a sample taken for the purpose of preliminary determination of the characteristic of the population; it is not the minimum sample needed to complete the study or research. If the population is normal, i.e. $N(0,1)$ with mean zero and variance of 1, equation (A) would become:

$$SE_{\bar{x}} = \frac{1}{\sqrt{n}} \qquad \text{(B)}$$

At this point, researchers would attempt to determine the value for $n$ by assuming that $SE = 0.05$. The calculation under equation (B) follows:

$$SE_{\bar{x}} = \frac{1}{\sqrt{n}}$$

$$0.05 = \frac{1}{\sqrt{n}} =$$

$$0.05\left(\sqrt{n}\right) = 1$$

$$\sqrt{n} = \frac{1}{0.05}$$

$$n = \left(\frac{1}{0.05}\right)^2 = \frac{1}{0.0025}$$

$$n = 400$$

It is concluded: "therefore," the minimum sample size is 400. This reasoning and logic is faulty on several grounds.

*Primo*, the logic is faulty because the researcher assumes that the data is normally distributed. This assumption is not reasonable unless it has been tested and verified that the data or population was normally distributed. This distribution verification may be accomplished by the Anderson-Darling Test.

*Secundo*, the logic is faulty because the parameter (symbol) *n* used in equation (B) is not the "minimum sample size" within the understanding of sample size needed to prove the condition: $\bar{x} = \mu$ and $S = \sigma$ for sample-population inferential analysis.

*Tertio*, the logic is faulty by assuming that $SE = 0.05$. The parameter $SE$ is known as standard error, hence the abbreviation SE. This abbreviation and its attendant meaning has been misinterpreted to mean "sampling error" and that the sampling error is mistakenly limited to an arbitrary attachment of the value 0.05 by the extension of further faulty reasoning that 0.05 comes from the precision level or random error level used in normal distribution curve. Further points of criticism follow this third faulty of logic:

(i) The parameter SE in equation standard for standard error. The word standard refers to the standard score. The standard score is determined by the measurement unit that is counted in "standard deviation" or amount of distances counted in standard deviation unit placed away from the mean. To the right of the mean, it is called +S and the left of the mean, it is called –S; where S standards for standard deviation. However, in the equation, instead of S, the Greek symbol sigma ($\sigma$) is used. It means further that this sigma comes from the assumption of normal distribution where the sample and population are assume to have equivalence statistical information, i.e. t = Z and $\sigma = [(\overline{x} - \mu) / Z]\sqrt{n}$. For that reason, the expression of SE is followed by a subscript of the sample mean, thus: $SE_{\overline{x}}$. When this reasoning and definition of SE are explained, then it becomes clear why the attempt of re-defining of SE to mean "sampling error" is truly erroneous; and

(ii) the use of 0.05 is also faulty. The precision level of random chance error of 0.05 is used in statistical significance test. In order to reach any conclusion of significance test, there must be a test statistic equation from which the result is used as a yard-stick to read the critical value from the significance test table, i.e. t-Table, Z-Table, chi-squared Table, F-table, etc. The value 0.05 in the erroneously interpreted equation (B) comes from nowhere. It is arbitrary picked and equated to the precision level that is 'commonly used" in statistics. This type of approach to statistics is spurious.

It is concluded that equations (A) and (B) are not formulae used to determine minimum sample size. Minimum sample size is given by the following formulae:

$$n_Y = \frac{N}{1 - N(\alpha^2)} \qquad \text{(C)}$$

where  is the population size and α is the error level which is set at 0.05 for 0.95confidence interval.

This equation is known as the Yamane equation. It may be used only when the population size is known. This is called finite population formula. However, if the population size is not known, then the Yamane equation is useless. In real life, we are faced with non-finite or unknown population size.

If the size of the population is not known, the following formula is used;

$$n = \frac{Z^2\sigma^2}{E^2} \qquad \text{(D)}$$

…where $Z$ is the standard score determine by: $Z = \frac{\overline{x} - \mu}{S / \sqrt{n}}$ for the data set and $Z = \frac{x_i - \overline{x}}{S}$ for an item within a set. A set is defined as $x_i : (x_1, x_2, ..., x_n)$. The parameter $E = SE$ in equation (A). We will revisit the issue of minimum sample size in Sect. 5, *infra*.

## 2.0 DESCRIPTION STATISTICS

Descriptive statistics are the properties of a data set; it describes the data. Descriptive statistics are used before formal inferences are made (Evans *et al.*, 2004). The data set comes from a sample. A sample comes from the population. Assume that the following data about student height is taken from a class room: $(195,170,165,165,160)$. The data may be described below. The sample size $n=5$ since there are 5 counts of data elements in the set $(195,170,165,165,160)$. The entire string of data $(195,170,165,165,160)$ is called a data set.

### 2.1 Sample mean

The mean is the mathematical average of the data set. We have been given a data set: $(195,170,165,165,160)$. The mean is determined by the following formula:

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \qquad (1)$$

where $i=$ item; $x_i=$ each item observed; $n=$ totals count of item of $x_i$, i.e. $x_i:(x_1,x_2,...,x_n)$; and $\sum_{i=1}^{n}$ = sum of items from 1st to nth term. Therefore, the mean of the data set $(195,170,165,165,160)$ may be calculated as:

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{(195+170+165+165+160)}{5}$$
$$=\frac{855}{5}$$
$$=171$$

It is said, the mean is of the data set $(195,170,165,165,160)$ is 171. The *mean* of 171 is an estimate of all 5 elements in the set. However, this estimate does not give an exact number. Some of the items in the set may be located above 171 and some may be found below 171. This difference is called *dispersion*. This dispersion illustrated by the mean difference. The mean difference is given by:

$$\delta=x_i-\overline{x} \qquad (2)$$

For the data set $(195,170,165,165,160)$ with the mean of 171, the mean difference may be calculated as:

**Table 1.** The mean of a data set is an approximation of the central value

| $i$ | $x_i$ | $\overline{x}$ | $(x_i-\overline{x})$ | Locate $x_i$ |
|---|---|---|---|---|
| 1 | 195 | 171 | 24 | Above mean |
| 2 | 170 | 171 | -1 | Below mean |
| 3 | 165 | 171 | -6 | Below mean |
| 4 | 165 | 171 | -6 | Below mean |
| 5 | 160 | 171 | -11 | Below mean |

This estimation is not accurate. This inaccuracy is shown through the mean difference $(x_i - \bar{x})$. Since some data point is located above and some data points are located below the mean, in order to get total possible dispersion per observation, the mean difference is calculated. The total dispersion among all data points from the mean is determined by the sum of the individual mean difference square. This sum squared mean difference is illustrated in the table below.

**Table 2.** Mean difference

| $i$ | $x_i$ | $\bar{x}$ | $(x_i - \bar{x})$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|---|
| 1 | 195 | 171 | 24 | 576 |
| 2 | 170 | 171 | -1 | 1 |
| 3 | 165 | 171 | -6 | 36 |
| 4 | 165 | 171 | -6 | 36 |
| 5 | 160 | 171 | -11 | 121 |
| | | | | $\sum(x_i - \bar{x})^2 = 770$ |

The difference each data point from the mean, i.e. missed estimation, may be located above the mean, at the mean or below the mean. This dispersion is made uniformed by squaring each mean difference: $(x_i - \bar{x})^2$.

The total sum of the mean difference is 770. This is the measurement of the total dispersion of all data points from the mean. This total dispersion is not helpful because it does not given any information for the individual dispersion in the data set $(195,170,165,165,160)$; therefore, it is necessary to distribute the total dispersion to each data point in the set $(195,170,165,165,160)$ by dividing the total dispersion by $n = 5$. This calculation follows the following formula:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (3)$$

Equation (1.6) is called *sample variance*. The calculation for the sample variance of the data set $(195,170,165,165,160)$ follows:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{770}{5}$$

$$S^2 = 154$$

The variance or average dispersion per data point in the set $(195,170,165,165,160)$ with mean 171 is $S^2 = 154$. The variance represents the error of the estimate. Recall that the estimate was the mean. The mean value was 171. Comparing 154 variance to the mean of 171, the error of the estimate appears large. The variance appears to be a large error because the variance is a square of the dispersion to accommodate for "above" and "below" the estimated value: $\bar{x} = 171$.

In order to minimize the error of the estimate, it is necessary to standardize the error into a standard score called standard deviation. Sample standard deviation is given by:

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (4)$$

Equation (1.4) is the square root of the variance. Standard deviation is defined as the square root of the variance. If the variance represents "error" of the estimate, standard deviation represents the 'minimization" of the error. The calculation for the standard deviation of the data set $(195,170,165,165,160)$ follows:

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} = \sqrt{154}$$

$$S = 12.41$$

The standardized error of the estimate is 12.41; it has no unit of measurement because it is a standard score. The standard deviation represents the common parlance of "given and take" or "plus or minus" jargons when a person gives a certain value. For example, the average height of students in this class is 171 plus or minus 12.41.

To make the calculation easier, we generally construct a table to calculate all descriptive statistics of a sample. This table is produced below.

**Table 3:** The uniform dispersion measured by the variance

| **Item** | **Data** | **Mean** | **Mean Difference** | **Mean Difference Squared & Variance** | **Standard Deviation** |
|---|---|---|---|---|---|
| $i$ | $x_i$ | $\overline{x}$ | $\left(x_i - \overline{x}\right)$ | $\left(x_i - \overline{x}\right)^2$ | $S = \sqrt{S^2}$ |
| 1 | 195 | 171 | 24 | 576 | $S = \sqrt{\frac{1}{n}\sum\left(x_i - \overline{x}\right)^2}$<br>$S = \sqrt{154}$<br>$S = 12.41$ |
| 2 | 170 | 171 | -1 | 1 | |
| 3 | 165 | 171 | -6 | 36 | |
| 4 | 165 | 171 | -6 | 36 | |
| 5 | 160 | 171 | -11 | 121<br>$\sum\left(x_i - \overline{x}\right)^2 = 770$<br>$S^2 \frac{1}{n}\sum\left(x_i - \overline{x}\right)^2 = \frac{770}{5}$<br>$S^2 = 154$<br>Variance = 154 | |

The variance gives uniformity of the dispersion, but it does not standardize the dispersion about (around) the mean (estimate); therefore, to standardize the measurement of the dispersion, the square root of the variance is taken. This is called *standard deviation*.

Standard deviation is the standard score, i.e. uniform of dispersion of data, about the mean of the data set. The standard deviation is used as a correcting value for the estimated mean. Therefore, the standard deviation is given as plus or minus $\pm S$ about the mean. When the mean is given, it must be given with plus or minus standard deviation in a form $\overline{x} \pm S$ because $\overline{x}$ is an estimated value and this estimate is not 100% accurate.

## 3.0 STANDARD SCORE

The standard score of measurement of the error or missed forecast about the mean is measure by a Z-score. The Z-score is a standard score telling us how far is an individual data point away from the mean. This standard score measurement is given by the t-formula. The t-formula is given by;

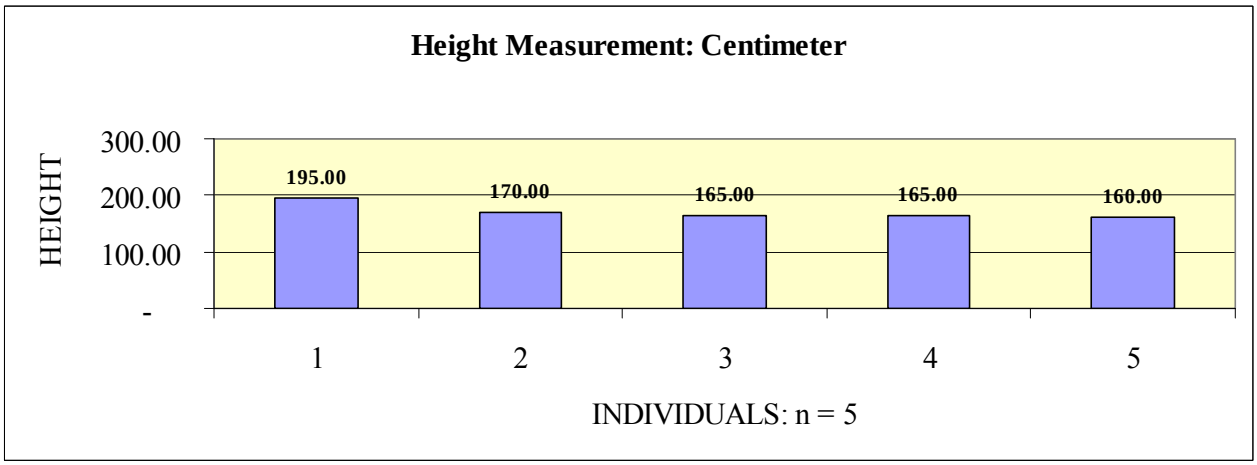
$$t = \frac{\bar{x} - \mu}{S / \sqrt{n}} \qquad (5)$$

where …

| | |
|---|---|
| $t$ | = sample standardized score, i.e. how far away from the mean; |
| $\bar{x}$ | = sample mean; |
| $\mu$ | = *ideal* mean or assumed population mean; |
| $S$ | = standard deviation of the sample data; and |
| $n$ | = sample size. |

The meaning of the t-formula may be described as the probability distribution of the data points in the set. For example, we are working with the data set $(195,170,165,165,160)$ which may be illustrated by the histogram below.

**Fig. 1.** Height measurement




The estimated height for the group is 171; this is the mean for the group. However, data set $(195,170,165,165,160)$ shows that each data each point does not equal to 171. The t-equation is a tool to provide the distance between each data point to the mean in a standard score form.

The assumption for the standard score measurement assumes that if there was an ideal data distribution, it would have been *normally distributed* in a perfect bell shaped curved call a normal curve. This curve is illustrated below.

For the data set $(195,170,165,165,160)$, the standard score under the t-equation may be calculated. Using equation (1.5), the standard score is calculated thus:

$$t = \frac{\bar{x} - \mu}{S / \sqrt{n}}$$
$$= \frac{171 - \mu}{12.41 / \sqrt{5}}$$

The value for $\mu$ is missing. The value of $\mu$ is the ideal height which may be estimated via the t-equation if the value of $t$ is known. The t-value is called the critical value. The critical value is given by the t-table. In order to read the t-table, two pieces of information are required: (i) degree of freedom of the data set and (ii) the level of confidence.

The *degree of freedom* is defined as the range of the data points from its first data point to its last data point. The degree of freedom is formally defined as: $df = n - 1$. In the present case, $n = 5$; therefore, the degree of freedom is $df = n - 1 = 5 - 1 = 4$. This degree of freedom may be read on the t-table at the first column.

The *confidence level* is the percentage distribution of the data within the probability distribution curve (*see* Figure 2.0) within which we accept as "normal." If the observation falls within this range of confidence, it is said that there is no significance because the data value is the value that is classified as a normal occurrence. If the data value falls outside of the confidence range, it is said to be significant because it is not within the range of normal expectation. Generally, by common practice a range of $\overline{x} \pm 2S$ or plus or minus two units of standard deviation about the mean is used. This range of $\overline{x} \pm 2S$ encompasses 0.95 or 95% of the data under the curved shown in Figure 2.0.

With 4 degrees of freedom at 95% confidence interval level, the critical value of t-score is 2.13. The reading of the t-table is illustrated below.

**Fig. 2.** Reading the Student T table

| | Percentile Point | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **df** | **70** | **75** | **80** | **87.5** | **90** | **95** | **97.5** | **99** | **99.5** |
| 1 | 0.73 | 1.00 | 1.38 | 2.41 | 3.08 | 6.31 | 12.71 | 31.82 | 63.66 |
| 2 | 0.62 | 0.81 | 1.06 | 1.60 | 1.89 | 2.92 | 4.30 | 6.96 | 9.92 |
| 3 | 0.58 | 0.79 | 0.98 | 1.42 | 1.64 | 2.35 | 3.18 | 4.54 | 5.84 |
| 4 | 0.57 | 0.77 | 0.94 | 1.34 | 1.53 | **2.13** | 2.78 | 3.75 | 4.60 |
| 5 | 0.56 | 0.75 | 0.92 | 1.30 | 1.48 | 2.02 | 2.57 | 3.36 | 4.03 |
| 6 | 0.55 | 0.74 | 0.91 | 1.27 | 1.44 | 1.94 | 2.45 | 3.14 | 3.71 |
| 7 | 0.55 | 0.73 | 0.90 | 1.25 | 1.42 | 1.89 | 2.36 | 3.00 | 3.50 |
| 8 | 0.55 | 0.72 | 0.89 | 1.24 | 1.40 | 1.86 | 2.31 | 2.90 | 3.36 |
| 9 | 0.54 | 0.71 | 0.88 | 1.23 | 1.38 | 1.83 | 2.26 | 2.82 | 3.25 |
| 10 | 0.54 | 0.70 | 0.88 | 1.22 | 1.37 | 1.81 | 2.23 | 2.76 | 3.17 |

When the sample size is 5, the degree of freedom is 4 because $df = n - 1 = 5 - 1 = 4$; this is used as a referenced ROW. This degree of freedom is located on the first column. The confidence interval selected is 0.95; this is used as referenced COLUMN. Where the row and column intersects, a critical value for *t* is found. In this case, the critical value is 2.13.

With known standard score or critical value $t = 2.13$, the value of the assumed population mean $\mu$ may be calculated using the t-equation. The calculation follows:

$$t = \frac{\overline{x} - \mu}{S / \sqrt{n}}$$

$$|\mu| = t\left(\frac{S}{\sqrt{n}}\right) - \overline{x}$$

$$= 2.13\left(\frac{12.41}{\sqrt{5}}\right) - 171$$

$$= 2.13\left(\frac{12.41}{2.24}\right) - 171$$

$$= 2.13(5.55) - 171$$

$$= 11.82 - 171$$

$$\mu = 159.18$$

The value of $\mu = 159.18$; it means that the "expected" average height of the population from which the sample data set $(195,170,165,165,160)$ is 159.18. However, this number is an estimate. Like the estimated sample mean, the estimated population is also not 100% accurate. The standard used in this estimation is 0.95 or 955 confidence interval. Therefore, it is necessary to give the estimated population height of 159.18 is an interval form. In order to construct an interval, it is necessary to determine the *population standard deviation*.

The t-equation gives us the population mean; however, it does not have a population mean. We need to look for a population standard deviation elsewhere. We have mentioned the term "assumed population" which we construct as an ideal population. This ideal population must be also fitted to the 0.95 confidence interval in order to give us a standard score for the population. The standard score for the population is given by the Z-equation. The Z-equation is given by:

$$Z = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}} \qquad (6)$$

where $Z$ = population standardized score, i.e. how far away from the mean; $\overline{x}$ = sample mean; $\mu$ = *ideal* mean or assumed population mean; $\sigma$ = standard deviation of the ideal population; and $n$ = sample size.

Similar to the exercise we did above in finding the critical value for *t*, under equation (6) with the facts given, we need to find the critical value for Z. he Z-table gives the critical value for Z by a confidence level. We have been using 0.95 confidence level in our calculation in the t-equation. Using 0.95 as the confidence interval, the Z-critical value is 1.645 or 1.65. The reading of this value is illustrating below.

**Fig. 3.** Reading the Z table

| $Z_{1-\alpha}$ | $1-\alpha$ | $Z_{1-\alpha}$ | $1-\alpha$ | $Z_{1-\alpha}$ | $1-\alpha$ |
|---|---|---|---|---|---|
| 1.40 | **0.919** | 1.75 | **0.960** | 2.10 | **0.982** |
| 1.41 | **0.921** | 1.76 | **0.961** | 2.11 | **0.983** |
| 1.42 | **0.922** | 1.77 | **0.962** | 2.12 | **0.983** |
| 1.43 | **0.924** | 1.78 | **0.962** | 2.13 | **0.983** |
| 1.44 | **0.925** | 1.79 | **0.963** | 2.14 | **0.984** |
| 1.45 | **0.926** | 1.80 | **0.964** | 2.15 | **0.984** |
| 1.46 | **0.928** | 1.81 | **0.965** | 2.16 | **0.985** |
| 1.47 | **0.929** | 1.82 | **0.966** | 2.17 | **0.985** |
| 1.71 | **0.956** | 2.06 | **0.980** | 2.41 | **0.992** |
| 1.72 | **0.957** | 2.07 | **0.981** | 2.42 | **0.992** |
| 1.73 | **0.958** | 2.08 | **0.981** | 2.43 | **0.992** |
| 1.74 | **0.959** | 2.09 | **0.982** | 2.44 | **0.993** |
| | | | | | |
| 1.645 | **0.950** | 2.326 | **0.990** | 3.090 | **0.999** |
| 1.960 | **0.975** | 2.576 | **0.995** | 3.291 | **0.9999** |

At 0.95 confidence interval, the critical value for Z is 1.645. This value may be rounded to 1.65. Throughout this Tutorial Note, the value 1.65 is used as a standard critical value for Z.

Using the Z-equation, the estimated population standard deviation may be calculated. The calculation for the population standard deviation follows:

$$Z = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}}$$

$$\sigma = \left( \frac{\overline{x} - \mu}{Z} \right) \sqrt{n}$$

$$= \left( \frac{171 - 159.18}{1.65} \right) \sqrt{5}$$

$$= \left( \frac{11.82}{1.65} \right) 2.24$$

$$= 7.16(2.24)$$

$$\sigma = 16.05$$

The estimated population standard deviation is $\sigma = 16.05$. Using 0.95 confidence interval, where 0.95 of the data falls within 2 units of standard deviation about the mean, the range of 0.95 confidence interval for the estimated population may be calculated thus: $\mu \pm 2\sigma$. The value $\mu + 2\sigma$ may be called the upper range and $\mu - 2\sigma$ may be called the lower range.

The upper range is: $\mu + 2\sigma = 159.18 + 2(16.05) = 159.18 + 32.10 = 191.28$ and the lower range is $\mu + 2\sigma = 159.18 - 2(16.05) = 159.18 - 32.10 = 127.08$. The estimated population mean of 159.18 now is more meaningful between it has a range between 127.08 and 191.28.

Recall that the sample data set was $(195,170,165,165,160)$, the sample mean was 171 and the sample standard deviation is 12.41. To construct a range for 0.95 confidence interval, we simply write $\overline{x} + 2S$ for the upper range and $\overline{x} - 2S$ for the lower range. The value of these two end points of the interval may be calculated thus: $\overline{x} + 2S = 171 + 2(12.41) = 171 + 24.82 = 195.82$ for the upper end of the range and $\overline{x} + 2S = 171 - 2(12.41) = 171 - 24.82 = 146.18$ for the lower end of the range of the sample.

## 4.0 INFERENTIAL STATISTICS

Inferential statistics is defined as using the sample descriptive statistics to make an inference (estimation) of the population. The sample is the observation; the estimated population is the inferred value without observation. In our example, we took a sample of 5 people with the height recorded as $(195,170,165,165,160)$. It was found that the estimated height was 171 which is the mathematical mean of the sample. The assumption here is that the data obeys the central limit theorem (Brewer, 2002, p. 6)

The estimated central value of 171 is not an accurate estimation. In order to speak with any degree of confidence about the estimated value, we need to select a standard. This standard is the confidence interval. The confidence interval selected is 0.95 or 95% confidence interval. This 0.95 confidence interval is equal to 2 units of standard deviation about the mean. The $\pm 2S$ helps us define the range of the 0.95 confidence interval. The upper bound of the sample height is 195.82 and the lower bound for the sample height is 127.08.

With the sample data in hand, we wanted to make an inference about the population. In order to accomplish this task, we need the sample and the population estimate to be uniform or expressed in common value. The commonality of the sample and the assumed population is accomplished through the standardized scoring system. The standard score measurement for the sample is given by the t-critical value. The standard score measurement for the assumed population is given by the Z-equation. By reading both the t-equation and Z-equation at a common confidence interval of 0.95, the sample is equated to the assume population. This equating t-equation to Z-equation may be illustrated.

Given that $t = \frac{\overline{x} - \mu}{S/\sqrt{n}}$ and $Z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$, set the two equations equal to come another thus $t = Z$, the simplification of the terms follows:

$\frac{\overline{x} - \mu}{S/\sqrt{n}} = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$ , multiply both sides by one of the denominator:

$\left(\sigma/\sqrt{n}\right)\frac{\overline{x} - \mu}{S/\sqrt{n}} = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}\left(\sigma/\sqrt{n}\right)$ the term $\sigma/\sqrt{n}$ cancels out on the right side of the equation. Now divided both sides of the equation by $\overline{x} - \mu$, thus:

$\left(\sigma/\sqrt{n}\right)\frac{\overline{x} - \mu}{S/\sqrt{n}}\left(\frac{1}{\overline{x} - \mu}\right) = \overline{x} - \mu\left(\frac{1}{\overline{x} - \mu}\right)$ the term $\overline{x} - \mu$ on both sides is reduce to 1. The only term left is:

$\frac{\sigma/\sqrt{n}}{S/\sqrt{n}} = 1$ , the term $\sqrt{n}$ may be removed by multiplying both sides by $\sqrt{n}$, thus leaving: $\sigma = S$. The population standard population is equal to the sample standard deviation when both sample and population is expressed in 0.95 confidence interval or at any level of confidence interval where both are of equal size.

Recall from equation (4) that the standard deviation is given by: $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$ and note further that standard deviation is obtained through taking the square root of the variance: $S = \sqrt{S^2}$. Therefore, the comparative term of $\sigma = S$ actually is a comparison study of the population variance and the sample variance. The sample variance is given by equation (3) as: $S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$. It is noted that the term $x_i$ is the individual data point of observation which may varied within a set of $x_i : [x_1, x_2, ..., x_n]$. The only fixed term is the estimated value or the mean: $\overline{x}$. Therefore, the comparison of $\sigma = S$ also yield another property about normal distribution: $\overline{x} = \mu$; the sample mean is equal to the population mean *within 0.95 confidence interval*. A question arises: what sample size would yield this condition: $\overline{x} = \mu$?

## 5.0 MINIMUM SAMPLE SIZE DETERMINATION

Sample size determination in qualitative research is less structured. There is no concrete or fix approach to determine sample size. Samples are taken as the research work progresses (Sandelowski, 1995; p. 179-183). For example, participants or respondents are taken into the research until a saturation point is reached (Glaser, 1965, pp. 436-445). There is no reliable guidance for sample size determination in qualitative research (Onwuegbuzie and Leech, 2007; pp. 105-121). However, this is not the case in quantitative research. This paper focuses on how minimum sample size is determined in quantitative research.

Minimum sample size is defined as a sample count or size that is not too large so as to cause inefficiency, i.e. waste of resources, and not too small so as to cause bias in sample-population inferential estimation. In another word, a minimum sample size is the size of sample from a population that would satisfy the condition: $\overline{x} = \mu$. Recall that a population size is defined as $N$ and a sample size is defined as $n$. Minimum sample size may be determined in two scenarios: (i) the population size $N$ is known. This is called finite population, and (ii) the population size is unknown; this is called non-finite population.

**5.1 Minimum sample size for finite population**

There is a general "rule of thumb" for minimum sample size. Sudman suggests minimum sample size to be at least 100 (Sudman, 1976). Kish suggests minimum sample size to be 30 to 100 elements (Kish, 1965). However, it is better not to do a guessing work if the sample is to be a good representative of the population. Where the population size is known, the means to calculate the sample size is through the use of population proportion technique. The Yamane equation is a common tool for sample size calculation in this scenario (Yamane, 1967). The Yamane's population proportion technique is given by:

$$n_Y = \frac{N}{1+N(\alpha^2)} \tag{7}$$

where $N$ = known population size and $\alpha$ is the precision level or error level. Conventionally, the value of $\alpha$ is set at $\alpha = 0.05$ where the confidence interval is defined as 0.95.

The example that we are working with a data set of $(195,170,165,165,160)$; the population size was not given. Assume that the population size is 2000 people. From the Yamane equation the minimum sample size may be determined. The calculation under equation (7) follows:

$$n_Y = \frac{N}{1+N(\alpha^2)} = \frac{2000}{1+2000(0.05^2)} = \frac{2000}{1+2000(0.0025)}$$
$$= \frac{2000}{1+5} = \frac{2000}{6}$$
$$= 333.33$$
$$n_Y = 333$$

The minimum sample size under this method is 333 people.

**5.2 Minimum sample size for non-finite population**

A second scenario involves unknown population size. This case calls for a different formula for minimum sample size calculation. The formula for minimum sample size in a non-finite population is given by:

$$n = \frac{Z^2\sigma^2}{E^2} \tag{8}$$

where $Z$ = standard score where $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$; $\sigma$ = population standard deviation; and $E^2$ = standard error where $E = \frac{\sigma}{\sqrt{n}}$.

Recall the t-equation: $t = \frac{\bar{x}-\mu}{S/\sqrt{n}}$, solve for $\mu$; thus: $\mu = t\left(\frac{S}{\sqrt{n}}\right) - \bar{x}$. From our prior calculation for the data set: $(195,170,165,165,160)$ we have:

$t = 2.13$
$S = 12.41$ solve for $\mu = t\left(\frac{S}{\sqrt{n}}\right) - \bar{x}$:
$\bar{x} = 171.00$

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

$$|\mu| = t\left(\frac{S}{\sqrt{n}}\right) - \bar{x}$$

$$= 2.13\left(\frac{12.41}{\sqrt{5}}\right) - 171$$

$$= 2.13\left(\frac{12.41}{2.24}\right) - 171$$

$$= 2.13(5.55) - 171$$

$$= 11.82 - 171$$

$$\mu = 159.18$$

Now solve for the population standard deviation using the Z-equation;

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$\sigma = \left(\frac{\bar{x} - \mu}{Z}\right)\sqrt{n}$$

$$= \left(\frac{171 - 159.18}{1.65}\right)\sqrt{5}$$

$$= \left(\frac{11.82}{1.65}\right)2.24$$

$$= 7.16(2.24)$$

$$\sigma = 16.05$$

Recall that $E = \frac{\sigma}{\sqrt{n}}$; therefore, $E = \frac{16.05}{\sqrt{5}} = \frac{16.05}{2.24} = 7.17$.

We are now ready to determine the minimum sample size using equation (8):

$$n = \frac{Z^2\sigma^2}{E^2} = \frac{1.65^2\left(16.05^2\right)}{7.17^2} = \frac{2.72(257.60)}{51.34} = \frac{701.32}{51.34}$$

$$n = 13.66$$

The minimum sample size is about 14 compared to 333 under the Yamane equation. It is clear from this example that the Yamane equation is not an efficient means for minimum sample size calculation.

## 6.0 CONCLUSION

In this introduction to descriptive and inferential statistics, we provide series of illustrations on how descriptive and inferential statistics are calculated. In addition, the minimum sample size required for a non-biased representation of the population by the sample was also explained. We also pointed out that the use of 400 counts as the sample size under the Yamane method is a misuse and misunderstanding of how to calculate sample size. The Yamane method is reserved for a finite population scenario. In non-finite case, the Yamane method is not appropriate.

## REFERENCES

Brewer, Ken (2002). Combined Survey Sampling Inference: Weighing of Basu's Elephants. Hodder Arnold. p. 6. ISBN 978-0340692295.

Glaser, B. (1965). "The constant comparative method of qualitative analysis." *Social Problems*, **12**, 436–445.

Michael J. Evans, Jeffrey S. Rosenthal W. H. Freeman (2004). Probability and Statistics: The Science of Uncertainty. Freeman and Company. p. 267. ISBN 9780716747420.

Kish, Leslie. 1965. Survey Sampling. New York: John Wiley and Sons, Inc.; p. 17.

Onwuegbuzie, A. J., & Leech, N. L. (2007). "A call for qualitative power analyses." *Quality & Quantity*, **41**, 105–121. doi:10.1007/s11135-005-1098-1

Sandelowski, M. (1995). "Sample size in qualitative research." *Research in Nursing & Health*, **18**, 179–183.

Sudman, Seymour. 1976. Applied Sampling. New York: Academic Press.

Yamane, Taro (1967). *Statistics, An Introductory Analysis*, 2nd Ed., New York: Harper and Row.