Data Classification and Distribution

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

The objective of this paper is to explain the four main types of data. The classification of data by type is important for statistical analysis. In particular, data classification is useful for quantitative research in social science. Data are defined as a quantitative measurement of qualitative fact. Data are classified into three types: quantitative, ordinal and nominal. Quantitative data are those that may be subject to mathematical operations: addition, subtraction, multiplication and division. Ordinal data are those that rank the values in a data set in an ascending order (from low to high) or from descending order (from high to low). Nominal data are those numbers or designation of value that is used for the purpose of identification. Nominal data cannot be subjected to mathematical operations. In addition to types, the data may also be classified according to their probability nature: (i) discrete data for discrete probability and (ii) continuous data for continuous probability.

Keywords: data types, quantitative data, nominal data, ordinal data

CITATION:

Sutanapong, C. and Louangrath, P.I. (2015). "Data Classification and Distribution" *Inter. J. Res. Methodol. Soc. Sci.*, Vol., 1, No. 2: pp. 36-47. (Apr. – Jun. 2015).

1.0 INTRODUCTION

1.1 Data classification

Data can be classified into two main types: (i) qualitative and (ii) quantitative. Qualitative data are those that are used for identification. No mathematical operations, such as multiplication and division may be allowed. Quantitative data are those that have non-arbitrary zero point and the data may be subjected to mathematical operations, such as addition, subtraction, multiplication and

division. In statistical analysis, these two types of data are further divided into subcategories. For qualitative data, there are two subcategories: (1) nominal data, and (2) ordinal data. For quantitative data, there are also two further categories: (1) interval scale, and (2) ratio scale.

1.2 Nominal data and its central tendency

Nominal data is qualitative data; it is used to differentiate between items or subjects based on names or meta-categories, such as gender, nationality, ethnicity and language. There are three measures for central tendency: mean, median, and mode.

The *mean* is a common measure used for the measurement of the central tendency. The mean is the expected value of the set. A mathematical mean is the sum of all items divided by the number of items. In a given set of $X_i : (x_1, x_2, x_3, x_4, x_5)$, the means is $\bar{X} = (x_1 + x_2 + x_3 + x_4 + x_5) \div 5$. It is a rough estimate of the central value. There are some values in the set which are located above and below the mean. The measure of that dispersion is called variance and, it is standardized in a standard unit of measurement as the standard deviation.

The median is simply the midpoint of the set such that the probability of the value falling above or below that point is equal.

The *mode* is used as the measurement for central tendency. The mode is defined as the value that appears most in frequency in a data set. In discrete probability, the mod is the X at which the probability mass function is maximized.

1.3 Ordinal data and its central tendency

Ordinal is a second type of qualitative data. This is a ranked order: 1^{st}, 2^{nd}, 3^{rd},… The data can be sorted on ascending order or descending order. The ranking may be dichotomous in form, such as (Yes | No), or non-dichotomous, such as: *completely agree*, *most agree*, *most disagree*, and *completely disagree*

The *median* is used for the measurement of central tendency. The median is the middle-ranked. The mean $\bar{x}$ is not allowed because the ranked order: 1^{st}, 2^{nd}, 3^{rd},… data may not be added or divided. Therefore, $\bar{x}$ is not the measure for central tendency. However, in addition to the median, the mode may also be used as a measure for central tendency. IQ test, for instance, is ordinal data; it is ranked data. There is no measurement to quantify intelligence (Mussen, 1973).

1.4 Interval data and its central tendency

Interval data is quantitative. Each item may be different; however, no ratios among items are allowed. This means that no division may be performed. For example, Celsius scale is an interval scale. However, a ratio of Celsius is not allowed. One cannot say that 20 degree Celsius is "twice" as hot as 10 degree Celsius because zero degree Celsius is an arbitrary number, i.e. 0 Celsius is equal to -273.15 kelvin. The interval variable is sometimes referred to as "scaled variable." A mathematical term for scaled variable is *affine line*. In affine space, there is no point of origin.

The central tendency of an interval data is measured by the mode, median and arithmetic mean. The measurement of the dispersion include range and standard deviation. In interval scaled data, multiplication and division are not allowed. Since division is not allowed, studentized range and coefficient of variation may not be calculated. The point of origin is arbitrary defined; thus, the central moment may be determined. Coefficient of variation may not be determined since the mean is a moment about the origin.

1.5 Ratio data and its central tendency

Ratio scaled data is quantitative. The measurement is the estimation of the ratio between a magnitude of a continuous and a unit magnitude of the same kind (Michell, 1997). Ratio scale has unique and meaningful zero. Mass, length, duration, plane angle and energy are measured by ratio scale. Quantitative data that is obtained through a measurement is this type of data. Since zero is not

arbitrary value; therefore, ratios are allowed. Multiplication and division are allowable mathematical operations.

The central tendency is measured by the mode, median, arithmetic mean are the basic measurements. In addition, geometric mean and harmonic mean may also be used. Studentized range and the coefficient of variation are used to measure dispersion.

Central tendency is the measure of the central value of a probability distribution (Weisberg, 1992). It is the average or the center point of a distribution. Common measures of the central tendency include the arithmetic mean, media, and mode. These three measurements are referred to as central tendency (Dodge, 2003).

Based on the law of large number (LLN), as the number of observation gets larger, there is a tendency for the estimated value to gravitate towards the mean of the group. This LLN also is another tool to understand the concept of central tendency theory.

In statistics, central tendency refers to the probability distribution of continuous data whose divider at 50/50 separating the lower range and upper range in equal area to be the mean. This mathematical mean is taken to define the central limit of the data distribution for estimating the value of the distribution, i.e. the expected value of the observation.

2.0 Probability distribution

A distribution is the fraction of individual events in relations to the whole number of observation. Adding all the individual events, the sum of the distribution is 1.0, i.e. each event is a proportion to the whole where the whole (of whatever is being observed) is 100% or simply 1.0.

A probability distribution is the probability of a subset of the possible outcome in relations to the entire observation. There are two types of probability based on the nature of the data: (i) discrete data produces discrete probability, and (ii) continuous data produces continuous probability function. Graphically, discrete probability produces a picture of a histogram where each data point stands alone and not connected to any other data point. Graphically, a continuous data set produces a continuous line or curve. The distribution is cumulative.

2.1 Discrete probability distribution

A data that is categorical, such as (Yes | No) is called discrete data. It is discrete because the observer must select one over the other. By selecting one choice, i.e. Yes, the other choice (No) is precluded. Two things are required: (i) event of interest which is generally defined as “Yes” or “Success.” This event of interest is assigned a score; that score is generally a number 1.0. The other event which is not an event of interest is “No”; it is assigned a score of zero (0); (ii) the second requirement is the total known observation. The total number of observation must be known; if this number is not known, probability cannot be calculated.

For discrete probability, some times called binomial distribution, the probability of a specified score may be determined if the number of past success and the number of total event are known. The binomial distribution is given by:

$$P(X) = \frac{n!}{(n-X)!\,X!}\, p^X q^{n-X} \qquad (1)$$

where $p = \dfrac{s+1}{n+2}$ and $q = 1 - p$.

The test statistic is given by:

$$Z_{bin} = \frac{\frac{X}{n} - p}{\sqrt{\frac{pq}{n}}} \qquad (2)$$

The critical value depends on the level of confidence. Generally, the level of confidence is set at 0.95; therefore, the critical value for Z is 1.65.

2.2 Continuous probability distribution

Continuous probability distribution is produced by continuous data. Continuous data is the type of data in which all data points are connected. The measurement of the probability of this type of data is in a form of probability density function. The data is cumulative and continuous. Graphically, it is represented as a continuous line or curve. A common form of representation is the Gaussian function which represents normal distribution:

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right] \qquad (3)$$

Where …

x = actual sample event or observation;
μ = estimated mean of the assumed ideal population; and
σ = estimated population standard deviation.

The variables μ and σ are considered *inferential statistics* which may be determined through the t-equation and Z-equation:

$t = \frac{\overline{x} - \mu}{S/\sqrt{n}}$, solve for population mean $\mu = t\left(\frac{S}{\sqrt{n}} \right) - \overline{x}$

and …

$Z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$, solve for population standard deviation: $\sigma = \left(\frac{\overline{x} - \mu}{Z} \right)\sqrt{n}$.

2.2.1 Normal distribution

Normal distribution is a function that explains the continuous probability of any given data point in the data set would fall between "two numbers." These two numbers are the upper bound and lower bound with the mean as the reference point, i.e. $\overline{x} \pm S$. The function that produces a normal curve is called the Gaussian function:

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right] \qquad (4)$$

The Gaussian function depends on three factors: the observation (x), the mean (μ) and the standard deviation (σ).

Probability density function (PDF) is defined as the density of a continuous variable. It is a function that gives the relative likelihood that a particular vale would take for a univariate, i.e. X. The function for the density is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{5}$$

where X is given. If more and more X are given then we will start to see a curve produced by f. The expected value for X is given by:

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx \tag{6}$$

Note that only continuous data has probability density function. Discrete data does not have probability density function. The requisite property of a probability density function is the cumulative distribution function (CDF). If there is no CDF, there cannot be PDF because this is a property that differentiates continuous random variable from discrete random variable.

Cumulative density function (CDF) is the area under the probability density function (PDF) from negative infinity up to point X. Recall that point X is a given value where we ask "what is the probability of X occurring? CDF exists only when the data variable is continuous and non-discrete. It cannot exists where the variable is discrete because in discrete data distribution, each data point is independent and does not connect to any other data points. Since the cumulative density is comprised of all area or region under the curve, cumulative density can only exists in continuous data form. The cumulative probability distribution function is given by:

$$F_X(x) = \int_{-\infty}^{x} f_X f(t)dt \tag{7}$$

where X = the entire data set of continuous data set $X_i : (-\infty, ..., \infty)$; x = a given data point that we want to determine its cumulative probability density up to that point; t = generally equated to time since each time we ask about the CDF of x we do so at a distinct time. We can answer about the cumulative probability of x one data point at a time since CDF is the cumulative of all probability up to that point x.

2.2.2 Standard normal distribution

Standard normal distribution is the Gaussian function for the normal distribution that has a mean of zero ($\mu = 0$) and standard deviation of one ($\sigma = 1$). This is a perfect bell shape curve. When we speak of normal distribution, we generally refer to the ideal form or standard normal distribution where the mean is equal to zero and the variance is one. The probability density function (PDF) of the standard normal distribution is given by:

$$\phi(x) = \frac{\exp\left(-\frac{1}{2}x^2\right)}{\sqrt{2\pi}} \tag{8}$$

Note that the factor $1/\sqrt{2\pi}$ is used to ensure that the entire area under the curve is equal to one, i.e. a unit distribution. The value of ½ in the exponent ensures a unit value or 1.0 variance and 1.00 standard deviation. The PDF is symmetric about zero; it means that it is a perfect mirror image of itself at zero mean ($\mu = 0$). The curve shifts direction or shows inflection points at -1 and +1.

2.2.3 General normal distribution

Every normal distribution is a version of standard normal distribution, meaning that it shares properties with the standard normal distribution. The mean (μ) is the reference point and the deviation from the mean is measured in units of standard deviation (σ). The probability density of the general normal distribution is given by:

$$f(x,\mu,\sigma) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) \tag{9}$$

The factor of $1/\sigma$ is used as a scale to ensure that the probability density is equal to 1.0. If X is a general normal distribution, then:

$$Z = \frac{(x-\mu)}{\sigma} \tag{10}$$

From the above statement, it also means that if Z is a standard normal distribution, then:

$$X = Z\sigma + \mu \tag{11}$$

where the distribution of Z has the expected value equal to μ and standard deviation σ. Every normal distribution is an exponent of a *quadratic* function:

$$f(x) = e^{ax^2+bx+c} \tag{12}$$

The property of the exponent may be summarized as follows:

$a < 0$	The *a* is negative.
$c = -\ln(-4\pi)/2$	The constant term
$\mu = -\frac{b}{a}$	The mean is a negative ratio of b/a
$\sigma^2 = -\frac{1}{2}a$	The variance is negative ½ of *a*.

For standard normal distribution, the properties are:

$a = -\frac{1}{2}$,
$b = 0$
$c = -\ln(2\pi)/2$

The *T-distribution* is used to analyze normal distribution of a sample. The sample is assumed to have normal distribution. Normal distribution is denoted as $N(0,1)$, i.e. identical, independent, distribution (IID) with mean of zero and variance of 1.0. Recall that a sample is a portion of a population taken with sample size of n from population where in the size of the population may be known or unknown or non-finite. This is the first scenario of normal distribution. The critical value of the t-distribution is given by the t-equation:

$t = \frac{\bar{x}-\mu}{S/\sqrt{n}}$ (the value of *t* is given in the T-Table).

The *Z-distribution* is a normal distribution of a population. When the population (not the sample) is the unit of analysis, use the Z-table. All properties and assumptions of normal distribution used in the t-distribution scenario are applicable. The population distribution is explained by the standard score equation for the normal distribution of the population. The Z-equation is given by:

$$Z = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}}$$ (the value of Z is given in the Z-Table).

3.0 NON-NORMAL DISTRIBUTIONS

One approach of differentiate data characteristic is to verify whether the data is normally distributed. A normal curve manifests at least five distinct characteristics. First, *Symmetry* around the mean where the mean, median and mode are equal. Second, the distribution curve is *unimodal*. Third, the area under the curve is *unity*. Fourth, there is an *inflection* point at +/- 1 standard unit about the mean. Lastly, the density of the curve is *log-concave*. If the distribution curve of the data breaks these characteristics, it is considered not normally distributed. One short hand indicator for testing the data's normality are skewness and kurtosis. A normal distribution, skewness is 0 and kurtosis is less than 3. Skewness and kurtosis are calculated by:

$$SKEW = \frac{n}{(n-1)(n-2)} \sum \left(\frac{X_i - \bar{X}}{S} \right)^3 \quad (13)$$

$$KURT = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{X_i - \bar{X}}{S} \right)^4 \quad (14)$$

If the kurtosis exceeds 3.0, the data does not have normal distribution. The expected value for skewness is zero; however, there is a problem to conclude when skewness and kurtosis leads to a different conclusion. For stance, kurtosis is less than 3, but skewness is non-zero. In such a case, we have a conflicting conclusion. The solution to this apparent conflict is to convert skewness to a Z score and reads the percentage probability whether it exceeds a certain threshold, i.e. 5% error. This reconciliation may be accomplished by D'Agostino's K square statistic (D'Agostino, 1970; D'agostino *et al.*, 1990).

The K square statistic or Omnibus test is used to test whether the non-zero skew conforms to normal distribution (D'Agostino, Belanger, and D'Agostino, 1990). With known skewness and kurtosis, determine $Z_1(g_1)$ and $Z_2(g_2)$, then calculate K squared. If K sq. ≤ 2.0, it means the data is normally distributed. If K sq. > 2.0, it means the data is not normally distributed.

$$K^2 = Z_1(g_1)^2 + Z_2(g_2)^2 \quad (15)$$

where $Z_1(g_1) = \delta a \sinh\left(\frac{g_1}{\alpha \sqrt{\mu_2}} \right)$

$$W^2 = \left(\sqrt{2\gamma_2 + 4}\right) - 1$$

$$\delta = \frac{1}{\sqrt{\ln W}}$$

$$\alpha^2 = \frac{2}{W^2 - 1}$$

$$\mu_{1(g1)} = 0$$

$$\mu_{2(g1)} = \frac{6(n-2)}{(n+1)(n+3)}$$

$$Z_2(g_2) = \sqrt{\frac{9A}{2}}\left\{1 - \frac{2}{9A} - \left(\frac{1 - \frac{2}{A}}{1 + \left(\frac{g_2 - \mu_1}{\sqrt{\mu_2}}\right)\sqrt{\frac{2}{(A-4)}}}\right)^{1/3}\right\}$$

$$A = 6 + \frac{8}{\gamma_1}\left(\frac{2}{\gamma_1} + \sqrt{1 + \frac{4}{\gamma_1^2}}\right)$$

$$\mu_{1(g2)} = -\frac{6}{n+1}$$

$$\mu_{2(g2)} = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

Table 1. K square statistic

Sample size	Expected value	Standard deviation	95% quantile
20	1.971	2.339	6.373
50	2.017	2.308	6.339
100	2.026	2.267	6.271
250	2.009	2.174	6.129
500	2.012	2.113	6.063
1000	2.009	2.062	6.038
$\chi^2(df = 2)$	2.000	2.000	5.991

3.1 Chi square distribution

When the sample size is small, the distribution of the data will not be normally distributed. In order to determine the goodness-of-fit, we need to compare our small sample to an ideal normal distribution. What does that mean? It means that with the small sample that we have, the data is not normally distributed. However, assume that the data is large enough then it would have been normally distributed. This "normal distribution," through assumption, is then compared to the assumed unit normal distribution (standard distribution) and compare our presumed "normal had we had large enough data" to the standard normal distribution and see how does our distribution fit to the standard one. If it is closely fit, we say it would meet the requirement of goodness-of-fit.

The goodness-of-fit under chi-square test is given by:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \qquad (16)$$

Recall that s^2 is the sample variance which is given by:

$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ and σ^2 population variance comes from $\sigma = \sqrt{\sigma^2}$ where:

$\sigma = \left(\frac{\bar{x} - \mu}{Z}\right)\sqrt{n}$ from the Z-equation: $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$. The chi-square test statistic for goodness-of-fit is read from a chi-square table.

The chi-square equation is also given as:

$$\chi^2 = \sum_{i=1}^{k}\frac{(O_i - E_i)^2}{E_i} \qquad (17)$$

where ..

O_i = observed frequency;

E_i = expected frequency.

The critical value for the chi-square shown in the table depends on the reading of the degrees of freedom. The degree of freedom is determined by: $df = (K - 1 - 2)$ where K is the number of classes. If the theoretical distribution contains m parameters as in $K : \{1, 2, ..., m\}$, then the degree of freedom may be defined as $df = (K - 1 - m)$. This degree of freedom is used to read off the critical value in the chi-square table. The degree of freedom is located in the first column of the chi-square table.

If the critical value for the chi-square is less than the critical value given by the chi-square table, it is considered good. It is said that the data or model provides a good fit. "Good fit" of what? It is a goodness-of-fit that fits into the mantel of the standard normal distribution. The purpose of the chi-square is to use the small sample to fit the normal distribution "had the sample been the size that would have been adequate to produce a normal distribution." If the observed critical value is larger than the expected value then it is said that the data or model does not fit the normal distribution. It can be said that the data is "truly non-normal" or significantly different from the normal distribution.

3.2 Poisson distribution

Recall that when the sample size is small and produces a non-normal distribution, the chi-square distribution is used to compare the assume distribution had the sample been increase to the magnitude that "would have produced a normal distribution." The chi-square situation involves one sample. In case where there are two small samples and both of which are non-normally distribute and are taken at different time intervals, chi-square study would have be able to accommodate our analysis. The Poisson distribution analysis accommodates this second scenario: two chi-square distribution with time interval.

The Poisson distribution is a distribution of a discrete data. A discrete random variable X is said to have a Poisson distribution with parameter lambda: λ where $\lambda > 0$ if $k = 1, 2, 3, ...$ The probability mass function (PMF) is given by:

$$f(k; \lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad (18)$$

where $e = 2.71828$ and $k!$ is the factorial of k. The value of lambda is $\lambda = E(X) = Var(X)$.

Although Poisson distribution is used with discrete random variable, the Poisson itself can be counted---that is, two Poisson distribution can be counted and compared. Recall that the simple Poisson distribution involves two binomial distribution comparing counts from two different time frames.

The count of each Poisson is represented as N_1 and N_2 taken from two time frames t_1 and t_2. The average of the two frequencies are:

$$R_1 = \frac{N_1}{t_1} \qquad \text{and} \qquad R_2 = \frac{N_2}{t_2} \tag{19}$$

It is assumed that the assumed frequencies are equal. The test statistic for the equality is given by:

$$Z = \frac{R_1 - R_2}{\sqrt{\frac{R_1}{t_1} + \frac{R_2}{t_2}}} \tag{20}$$

The hypothesis statement follows: $H_0 : R_1 = R_2$. The null hypothesis argues that both frequencies have the same average, i.e. there is no significant difference. $H_A : R_1 \neq R_2$. The alternative hypothesis argues that the two average frequencies are not the same, i.e. they are statistically significantly different.

Generally, this type of test is used to prove changes after a stimulus is introduced into the system, i.e. new procedures or new policy, and a measurement is taken to prove whether the stimulus causes any changes or make the system respond by comparing the data from two time period. This type of comparison study is useful in empirical research. The phrase "comparison study" is not a grammatical error. The phrase is used when two data sets are compared. It is incorrect to use the term "comparative study." Comparison focuses on a certain characteristic of property of the data. Comparative study involves the entire set or all characteristics of the set.

3.3 F Distribution

Recall that a population study is tested by the Z-equation and the critical value for the distribution is given by the Z-table. Where the study involves two populations, the F-table is used. The two populations or groups may have different sizes. Different sizes imply different degrees of freedom. Recall that the degree of freedom is defined as $df = n - 1$. Thus, the F-table is read by using the degrees of freedom from each group.

F-distribution is a continuous probability distribution. It is used in the analysis of variance. In the t-distribution and Z-distribution, the means are used; therefore, those two tests are called means analysis. The variance is the shape of the distribution curve. Therefore, F-distribution is the comparison study of the two curves via their variances or the shape of the curves. The test statistic used to verify whether the two populations are significantly different is given by:

$$F = \frac{\left(S_1^2 / \sigma_1^2\right)}{\left(S_2^2 / \sigma_2^2\right)} \tag{21}$$

where $S_1^2 = \frac{\sum \left(x_i - \bar{x}\right)^2}{n_1 - 1}$, and $S_2^2 = \frac{\sum \left(y_i - \bar{y}\right)^2}{n_2 - 1}$;

Assuming that there are two groups of data call $X_i : (x_1, x_2, ..., x_n)$ and $Y_i : (y_1, y_2, ..., y_n)$, the mean for each group is given by:

$$\overline{x} = \sum_{i=1}^{n} x_i \qquad \text{and } \overline{y} = \sum_{i=1}^{n} y_i \, .$$

If the distribution is normal, the variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 1$. Therefore, F becomes:

$$F = \frac{S_1^2}{S_2^2} \tag{22}$$

The degree of freedom is defined as $df(n_1 - 1, n_2 - 1)$ where $n_1 - 1$ is the numerator degree of freedom and $n_2 - 1$ is the denominator degree of freedom. The hypothesis statement follows: $H_0 : S_1^2 = S_2^2$. The null hypothesis argues that both variances are equal, i.e there is no significant difference. The alternative hypothesis is $H_A : S_1^2 \neq S_2^2$. The alternative hypothesis argues that the two variances are not the same, i.e. they are statistically significantly different.

In the F-table, the top row (from left to right marked as d1) is the numerator degree of freedom. The first column (top to bottom marked as d2) is the denominator degree of freedom. For example, if d1 = 6 and d2 = 5, the F-critical value is 4.95. *What does it mean?* Recall that the F-test measures the significance of the variance between two groups. If the critical value from the observed data is less than 4.95, it means that the difference between the two groups or populations is not significant. However, if the calculation for $F = S_1^2 / S_2^2$ is greater than 4.95, it means that the difference among these two populations is significant, i.e. they are real difference.

For a second example, d1 = 6 and d2 = 10, the critical value for F is 3.22. The null hypothesis argues that the difference in variance among the two populations is not statistically significantly different, i.e. $H_0 : F_{obs} < 3.22$ and the alternative argument (your argument) states that $H_A : F_{obs} > 3.22$. If you have the data of the two populations, calculate the variance comparison according to the formula $F = S_1^2 / S_2^2$ and reach a conclusion according to the decision rule of H_0 and H_A then reject or accept H_0 accordingly.

4.0 CONCLUSION

This paper is a foundational materials on statistics needed for quantitative research in social science. In this part 2, we explore three main types of data: quantitative, ordinal, and nominal, and their central tendency. In addition, we traced four common types of test statistics, namely Student T, Z, chi square and F tests. The Student T test is used for sample analysis. The Z test is used for population or expected value analysis. Both T and Z test requires the data distribution to be normal. The chi square test is a test of fitness. The fitness attempt is to fit the distribution of the empirical data to the assumed normal distribution. If the sample is not normally distributed, the chi square test is used. For two samples, which are not normally distributed, the F test is used. The F test verifies the fitness via ratio analysis of two samples in order to differentiate one from the other. If there is a signif9cant difference, it means that the two samples came from a different source or process. Conversely, if the difference between the two samples is not significant, it is concluded that both sample may have come from the same source or process.

REFERENCES

D'Agostino, Ralph B. (1970). "Transformation to normality of the null distribution of 1." *Biometrika*, **57** (3): 679–681. doi:10.1093/biomet/57.3.679. JSTOR 2334794.

D'Agostino, Ralph B.; Albert Belanger; Ralph B. D'Agostino, Jr (1990). "A suggestion for using powerful and informative tests of normality." *The American Statistician*. 44 (4): 316–321. https://web.archive.org/web/20120325140006/http://www.cee.mtu.edu/~vgriffis/CE%205620%20materials/CE5620%20Reading/DAgostino%20et%20al%20-%20normaility%20tests.pdf

Dodge, Y. (2003) The Oxford Dictionary of Statistical Terms, OUP for International Statistical Institute.ISBN 0-19-920613-9 (entry for "central tendency"); and Upton, G.; Cook, I. (2008) Oxford Dictionary of Statistics, OUP ISBN 978-0-19-954145-4 (entry for "central tendency").

Henk, Tijms (2004). Understanding Probability: Chance Rules in Everyday Life. Cambridge: Cambridge University Press. p. 169. ISBN 0-521-54036-4.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. British Journal of Psychology, 88, 355–383.

Mussen, Paul Henry (1973). Psychology: An Introduction. Lexington (MA): Heath. p. 363. ISBN 0-669-61382-7. "The I.Q. is essentially a rank; there are no true 'units' of intellectual ability."

Pólya, George (1920). "Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem" [On the central limit theorem of probability calculation and the problem of moments]. *Mathematische Zeitschrift* (in German). 8 (3–4): 171–181. doi:10.1007/BF01206525. https://gdz.sub.uni-goettingen.de/id/PPN266833020_0008

Weisberg H.F. (1992) Central Tendency and Variability, Sage University Paper Series on Quantitative Applications in the Social Sciences, ISBN 0-8039-4007-6. p.2.