Introduction to Quantitative Methods

Parina Patel

October 15, 2009

Contents

List of Tables

1 Definition of Key Terms

1. Unit of Analysis (also referred to as cases): The most elementary part of what is being studied or observed. Some examples include individuals, households, court cases, countries, states, firms, industries, etc.

2. Variables: Concepts, characteristics, or properties that can **vary, or change, from one unit of analysis to another**. Please note that all variables must vary, if there is no variation among the different cases then it is not a variable. Some examples of variables include gender, social class, education, age, level of public enforcement, type of bankruptcy, etc.

   (a) Dependent Variable–DV: Variables whose change the researcher wishes to explain

   (b) Independent Variable–IV: Variables that help explain the change in the dependent variable

3. Hypothesis: An empirical statement which seeks to test the relationship between at least two variables. For instance, As levels of public enforcement increases, levels of stock development also increases. This hypothesis has two variables: (1) public enforcement–independent variable, and (2) stock development–dependent variable.

4. Levels of Measuring Variables

   (a) Nomial: A nominal variable has qualitative categories that cannot be ranked in a meaningful way in terms of degree or magnitude. Examples of nominal variables include RACE, TYPE OF BANKRUPTCY, TYPE OF CORPORATION, NAME. All of these variables have qualitative categories that cannot be ordered in terms of magnitude or degree. This is the least powerful type of variable. [1]

   (b) Ordinal: An ordinal variable has qualitative categories that are ordered in terms of degree or magnitude. Examples of a nominal variable include CLASS or DEGREE OBTAINED. The variable DEGREE OBTAINED may include the following categories:

[1] Alphabetizing the categories does not count as ordering the variable, because the ordering has to be in terms of degree or magnitude.

None, High School Diploma, College/University Degree, Masters, Advanced Degree (JD/PHD/MD). All of these categories are qualitative and are ordered in terms of the amount of education each individual has completed.

(c) Interval/Ratio: An interval variable has quantitative values (or numbers). Some examples of interval variables include AGE (in years), NUMBER OF SHARES OUTSTANDING, and AMOUNT IN DEBT (in dollars). For all of these variables the response is going to be a number or value. This is most powerful type of variable because you can do the most with it statistically.
Note that if a variable has qualitative categories that ARE ordered and there are numerical values assigned to each category which are also ordered, we can treat this variable like an interval level variable. An example would be questionnaire that asks respondents about their feelings towards President Obama's handling of the economy on a scale of 1 to 5 where (1=very bad job, 2=bad job, 3=neither bad nor good, 4=good job, and 5=very good job). The respondents are asked to choose a category that is ordered, but since it has ordered numbers attached to the categories, we can treat it as an interval level variable with some restrictions. [2]

(d) Dichotomous/Dummy: A dichotomous variable is a variable with two (and only two) categories. These categories can be qualitative or quantitative values. [3]

2 Descriptive Statistics

Descriptive statistics are often used to describe variables. Descriptive statistics are performed by analyzing one variable at a time (univariate analysis). All researchers perform these descriptive statistics before beginning any type of data analysis.

[2] One such restriction being the dependent variable in regression analysis. In order to perform regression (see section 3.4) your dependent variable must be a proper interval variable.

[3] It is possible to convert nominal variables into numerous dichotomous/dummy variables.

2.1 Frequency Tables

Frequency tables are a detailed description of the categories/values for one variable. A frequency table most often includes all of the following: [4]

1. Absolute frequency (or just frequency): This tells you how many times a particular category in your variable occurs. This is a tally, count, or frequency of occurrence of each individual category/value in the table.

2. Relative frequency (or percent): This tells you the percentage of each category/value relative to the total number of cases.

3. Cumulative frequency: This is simply a cumulation of the relative frequency for each category/value.

   Table 1 provides an example of a frequency table for an ordinal variable (note it is ordinal because the categories are qualitative and ordered) named Socioeconomic Class. If there were numbers assigned to each category that were also ordered, we could treat this as an interval level variable.

Table 1: Frequency Table–Socioeconomic Class

Socioeconomic Class	Frequency	Percent	Cumm. Percent
Upper	50	7.14%	7.14%
Upper Middle	150	21.43%	28.57%
Middle	300	42.86%	71.43%
Lower Middle	150	21.43%	92.86%
Lower	50	7.14%	100%
Total	700	100%	

4. Crosstabulations: This is also referred to as a grouped frequency table for two variables. A crosstab simply presents the absolute frequency broken down by categories of two or more variables. It is also possible to find percentages in these types of tables. For instance, using the

[4] The stata command for frequency is **fre** or **tab**. Before you use the **fre** command you need to install it onto your computer, so you need to type the following command: "**ssc install fre**," which will install the **fre** command onto your computer. For a frequency table of a variable named "class," type either "**fre class**" or "**tab class**"

example below, we can find the percentage of young people that listen to music. [5]

Table 2: Crosstab of Music Preference and Age

	AGE		
Preference	*Young*	*Middle Age*	*Old*
Music	14	10	3
News-talk	4	15	11
Sports	7	9	5

2.2 Measures of Central Tendencies

Measures of central tendencies provide the most occurring or middle value/category for each variable. There are three measures of central tendencies–mode, median, and mean. See Table 3 for a summary of measures of central tendencies.

2.3 Measures of Variability

Measures of variability is defined as the dispersion (or deviation) away from the mean for each variable. Measures of variability only exist for interval level variables. There are three measures of variability–range, standard deviation, and variance. A discussion of each can be found below followed by a summary table (Table 3).

1. Range: The range is found by taking the highest value of a variable minus the lowest value of that variable.

2. Standard deviation: The standard deviation exists for all interval variables. It is the average distance of each value away from the sample mean. The larger the standard deviation, the farther away the values are from the mean; the smaller the standard deviation the closer, the values are to the mean. Suppose you passed out a questionnaire

[5] The stata command for a crosstab is either **tab** or **tab2**. For a crosstab of two variables named "age" and "preference," type the following command into stata: "**tab age preference**"

asking randomly selected individuals to rate President Obama's job performance on a scale from 1 to 10. You find that on average these individuals give the President a rating of 5.8, and this variable has a standard deviation of 1.2. This means that on average, each rating of the President is approximately 1.2 points away from 5.8 (the sample mean).

3. Variance: The variance is always going to the the standard deviation squared. The variance cannot be interpreted as meaning anything other than the standard deviation squared. [6]

2.4 Summary of Central Tendencies and Variability

Table 3: Summary of Univariate Statistics

Univariate Statistic	**Variables**	**Description**
Mode	Nominal, Ordinal, and Interval	most frequent category/value
Median	Ordinal and Interval	category/value that lies in the middle
Mean	Interval	value that represents the average
Range	Interval	highest value minus lowest value
Standard Deviation	Interval	on average how much each individual value is dispersed around the mean
Variance	Interval	standard deviation squared

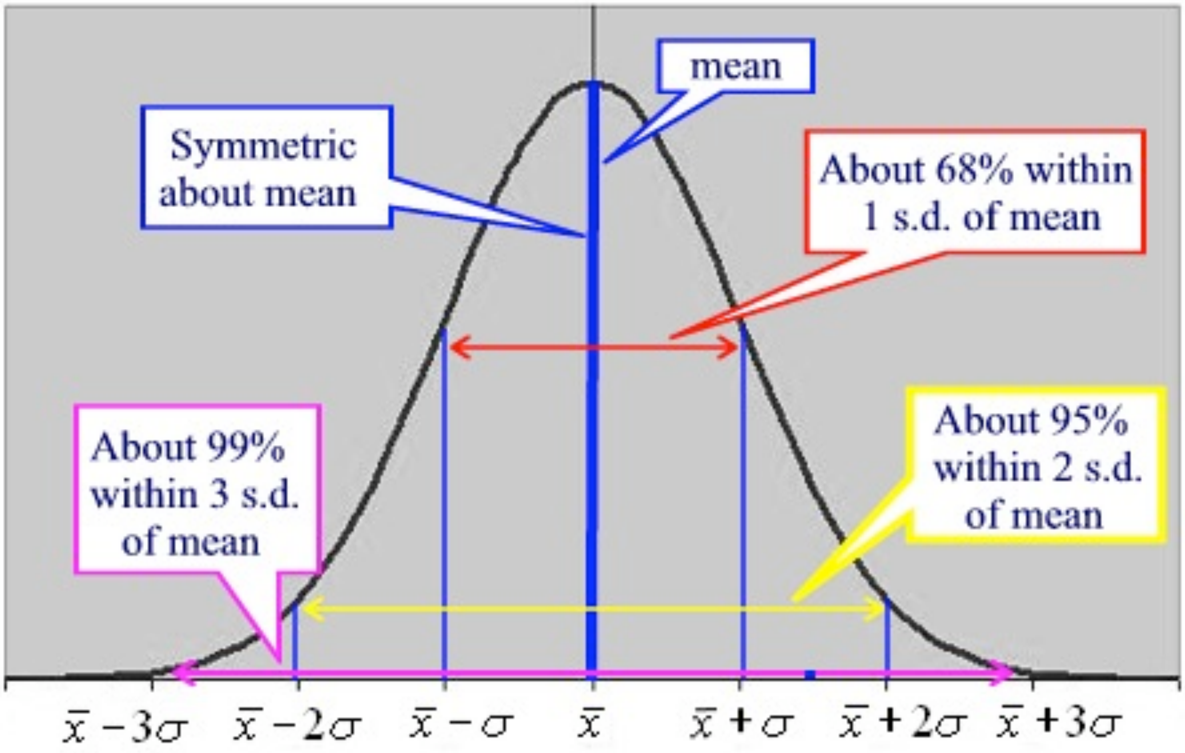
3 Inferential Statistics

3.1 More Definitions and Terms

1. Normal Curve
An interval variable is said to be normally distributed if it has all of the following characteristics:

   (a) A bell shape curve.

[6] In stata, the easiest way to find the mode is by looking at a frequency table and finding the value/category that occurs most frequently. The median, mean, standard deviation, and variance can be found by using the following command: **sum var1 var2 ..., detail**



(b) It is perfectly symmetrical.

(c) All measures of central tendencies (mode, median, and mean) lie in the middle middle of the curve. These measures of central tendencies divide the curve in half (where 50% of the values lie to the left of the mean, and 50% lie to the right).

(d) Approximately 95% of the values are found two standard deviations away from the mean (in both directions).

Variables that are determined by nature are normally distributed (graphically they have a normal curve) such as age, weight, height, etc. It is important to understand what a normal curve looks like and its characteristics because almost all methods described below assume normality. If this assumption is violated (i.e. a variable is not normally distributed) it can have an effect on the statistical results (resulting in significance when in reality it is not significant, or not resulting in statistical significance when it is significant). If variables are not normally distributed, it is easy to make transformations, such as logging or taking the square root, in order to achieve normality. [7]

2. Confidence Intervals: Confidence Intervals are used to estimate a range of the population based on some sample of any interval level variable. Confidence intervals are two numbers which represent the higher and lower limits of a statistic, coefficient, or paramater. Confidence intervals assume the interval level variable has a normal distribution, and uses the sample in order to find a range of the entire population. When dealing with confidence intervals, a confidence or an α level (see discussion below on statistical significance for explanation of these terms) must be specified.

3. Standard Error: Standard error is the estimated standard deviation, and the standard error squared is the estimated variance. Standard error plays a large role in testing for significance, and can drastically affect the outcome. For instance, large standard errors will cause variables to be insignificant, which may indicate an incorrect use of a statistical method or analysis.

4. Statistical Significance: Statistical significance represents the results of some statistical test that is being performed. The statistical test varies depending on the levels of measurement of the variables, and the objective of the research or hypothesis. There are numerous different tests but they all have some similarities and include all of the following:

   (a) One Null Hypothesis: The null hypothesis usually states there is no relationship between the variables being tested. The null hypothesis is already determined and based on the method being used. Most null hypotheses state that one statistic or number is equal to another statistic or number. This is usually displayed as: $H_0 : a = b$

   (b) One Alternative Hypothesis: [8] The alternative hypothesis usually states that the two or more variables are somehow related.

[7] The best way to see if an interval variable is normal is with a histogram. A histogram is a graph which places the values of the interval variable on the X axis, and the frequency or density on the Y axis. In Stata, the command for a histogram is **histogram var1, freq normal**

[8] This is also referred to as a research hypothesis. I refer to this as a research hypothesis or an alternative hypothesis

Like the null hypothesis, the alternative hypothesis is also already determined based on the method being used. The alternative hypothesis is the opposite of the null hypothesis and usually states that one statistic or number is not equal to another statistic or number. Alternative hypotheses can be displayed using one of the following forms:

i. $H_a : a \neq b$

ii. $H_a : a > b$

iii. $H_a : a < b$

(c) Outcome of Statistical Test: All tests will always have one of two outcomes:

i. Reject H_0 and Accept H_a, which means that the null is wrong and the alternative hypothesis is right. When this occurs, there is statistical significance, and a confidence or an α level is always specified (discussed below).

ii. Accept H_0 and Reject H_a, which means that the null is right and the alternative hypothesis is wrong. When this occurs, there is no statistical significance, and therefore no need to specify a confidence or an α level.

(d) Confidence Level: With every test that is being performed, the researcher must specify a confidence level if and only if the H_0 is rejected and the H_a is accepted. The confidence level literally translates into the percentage or probability of correctly rejecting the null hypothesis (or the level in which the outcome is correct). In social science and statistics, the most common confidence levels are AT LEAST 90%. This means that with each test, researchers want to be right at least 90% of the time. Usually researchers report confidence levels at 90%, 95%, 99%, or 99.9%, the higher the confidence level the better the results. The confidence level can also be displayed as a probability, where 90% confidence level translates into being right with a probability of 0.90.

(e) α level or $p-value$: [9] In addition to specifying a confidence level when rejecting the H_0, an α level must also be specified. Similar to

[9] These are just different names for the same thing, and can be seen in articles as statistical significance, α, p-value, or even Type I error. I use these terms interchangeably, but they all mean the same thing.

the confidence level, the α level can be displayed as a percentage or a probability. The α level represents the probability or percentage of rejecting the null when it should not have been rejected. In other words, the α level is the probability or percentage of making a mistake, and the lower the α level the better the results. A statistical significance (or α level) of 1% is better than a statistical significance of 5%.
α (probability) = 1− probability confidence level
α (percentage) = 100 – percentage confidence level

(f) What does a statistical test do: When performing any type of inferential statistics and any type of statistical testing, a value is generated based on the data (either a T, F, Z, or χ^2), and this value is being compared to some corresponding critical value [10] (T, F, Z, or χ^2) in order to determine statistical significance.

3.2 Comparing Two or More Groups

The following methods are used if a researcher is interested in comparing differences in two or more group means, and determining whether the difference is statistically significant or a result of sampling error. This portion will provide a table of some of the differences between the methods, and a brief discussion of important details and clarification of each method.

Table 4: Comparing Group Means

Method	**Grouping Variable**	**Mean Variable**	**Stata Command**
Two sample T	Dichotomous	Interval	**ttest var1, by(var2)**[11]
Paired T	Before and After	Interval	**ttest var1==var2**
One way ANOVA	Nominal	Interval	**oneway var1 var2**[12]

1. Two sample T Test
   Two sample t test compares means across TWO (and only two) groups.

[10] These critical values can be found by looking at tables in the back of any statistics or research methods textbook.

[11] In two sample t test var1 is the interval variable, and var2 is the grouping variable. Note the grouping variable MUST be a dummy with 2 categories

[12] In ANOVA var1 is the interval variable, and var2 is the grouping variable

For instance, a researcher wishes to know if there is a difference in the amount of debt (in dollars) in Chapter 7 and Chapter 13 bankruptcies. The researcher is interested in two variables, (1) the amount of debt in dollars (interval), and (2) Chapter 7 or Chapter 13 (dummy). The null hypothesis for this example is that there is no difference in the amount of debt between Chapter 7 and Chapter 13 bankruptcies. In other words, the average amount of debt for Chapter 7 equals the average amount of debt for Chapter 13. The alternative hypothesis is that the average amount of debt is different for these two types of bankruptcies. This test assumes that the interval variable is normally distributed across both groups, and that the variances are equal across both groups.

2. Paired T Test

   Paired t (also referred to as matched t) test compares means across the SAME VARIABLE and the SAME CASES at two different times. Suppose you were interested in looking to see if neighborhood watch was effective. You had data on the number of calls to 911 in neighborhoods one week before the watch was implemented and one week after the watch started. The null hypothesis is that the neighborhood watch was ineffective (or that the number of 911 calls before the watch remained unchanged after the watch). Matched t test assumes that the interval variable is normally distributed at both times.

3. ANOVA

   ANOVA (Analysis of Variance) is similar to the two sample t test, but it compares means across MORE THAN TWO groups. A researcher would use ANOVA if s/he was interested in comparing the difference in the amount of debt (in dollars) between Chapters 7, 12, and 13. Now the researcher is interested in comparing three types of bankruptcies, and wants to test the null that there is no difference in the amount of debt for these three types of bankruptcies, or that these three types of bankruptcies has equal group means. The research hypothesis is that at least one type of bankruptcy is different from another in terms of the average amount in debt. Conducting ANOVA alone will not tell you which groups are different from each other, but only if there is a difference between groups. In order to see which specific groups are different you need to conduct a post-hoc test (this is usually done if and only if you find there is a difference between the groups). ANOVA assumes that the interval variable is normally distributed.

3.3 Association and Correlation

Tests of association only tell you if two or more variables are associated (if two variables tend to occur together). Correlations tell you not only if the variables are associated but also the direction and strength of the relationship. Correlations only range from -1 to 1. A correlation of 0 means that the variables are not related. A positive correlation indicates a positive relationship (an increase in one variable leads to an increase in another variable), while a negative correlation indicates a negative relationship (an increase in one variable leads to a decrease in another variable). The closer a correlation is to -1 or 1 the stronger the relationship between the variables. For instance a correlation of 0.01 to 0.3 indicates a weak positive relationship, while a correlation of -0.01 to -.3 indicates a weak negative relationship. A correlation of 0.31 to 0.69 indicates a moderate positive relationship while a correlation of -0.31 to -0.69 indicates a moderate negative relationship. A correlation above 0.7 indicates a strong positive, and a correlation below -0.7 indicates a strong negative relationship.

Table 5: Association and Correlation

Method	**Type**	**Level of Variables**	**Stata Command**
Chi Square	Association	Nominal	**tab var1 var2, chi2**
Fishers Exact	Association	Dichotomous	**tab var1 var2, exact**
Pearson	Correlation	Interval	**corr var1 var2**
Spearman	Correlation	Ordinal	**spearman var1 var2**

Statistical significance does not tell you the “strength” of the relationship but only how often the relationship between the two variables holds. For instance, association only tells you if the two variables are statistically significant, or how often these two variables are not independent (a statistical significance of 5% indicates that 95 out of 100 times these two variables tend to occur together). Correlation, on the other hand, tells you how often these two variables are related as well as how strong the relationship is between the two variables. A correlation of 0.8 that has a statistical significance of 5% indicates that 95 out of 100 times these two variables strongly and positively occur together. The null for all of the methods described in Table 5 is that the two variables are independent (or that they are not associated/correlated).

3.4 Explaining a Dependent Variable

If you are looking to explain a variable using one or more variables then you want to use any of methods described below depending on the level of measurement of the dependent variable. Note that for all of the models described below, you can never have a nominal variable (with ore than two categories) as in independent variable. If you want to use a nominal level variable as an IV, then you must recode it into one or more dummy variables. Suppose you wanted the model to control for race. Since race is a nominal variable you cannot include it in your analysis as an IV, however, you could create dichotomous variables for all the categories of race and include these newly formed dichotomous variables.

Table 6: Explaining a Dependent Variable

Method	**Dependent Variable**	**Independent[13] Variable**	**Stata Command**
OLS Regression	Interval[14]	Dummy/Interval	**regress dv iv1 iv2 iv3...**[15]
Logit/Probit	Dummy	Dummy/Interval	**logit dv iv1 iv2 iv3...**
Cumulative Logit	Ordinal	Dummy/Interval	**ologit dv iv1 iv2 iv3...**
Multinominal Logit	Nominal	Dummy/Interval	**mlogit dv iv1 iv2 iv3...**
Poisson	Count[16]	Dummy/Interval	**poisson dv iv1 iv2 iv3...**

The null hypothesis for all of these methods is that the independent variable does not have an effect on the dependent variable. This null hypothesis is performed for each independent variable. [17]

[13] You can also use ordinal level variables as independent level variables as long as the categories are ordered in terms of degree or magnitude and the numbers corresponding to the categories are also ordered.

[14] The DV in regression analysis MUST be an interval variable. It cannot be an ordinal variable that is being treated as an interval variable, or a dichotomous variable being treated as an interval variable.

[15] dv=your dependent variable, iv=your independent variable(s)

[16] The dependent variable must not have any negative values and can be non-normal.

[17] For all of these methods, the null hypothesis is that the coefficient (β) is equal to 0, and the alternative hypothesis is that the coefficient is a value no equal to zero. This literally means different things for the different methods, but can be interpreted the same way. For regression a coefficient equal to zero literally means that there is no linear change in the dependent variable as the independent variable changes, or that the slope is equal to 0. For the logit and poisson models, the null hypothesis literally means that the logit or the log of the odds is zero.

3.4.1 Assumptions

All of these models have certain assumptions. If these assumptions are not met, it affects the way the results are interpreted and can lead to serious errors in the statistical tests. All of these models have some similar assumptions about the observations, independent, and dependent variables. All models assume that the observations are independent and are randomly sampled from the entire population of interest. [18] The models also assume that no important independent variables are omitted from the model, and that all variables are measured without error. All of these models are stating that there is some functional relationship between the independent and dependent variables. The exact functional relationship is will depend on the method you use to fit your data. For instance, if you are using OLS Regression, this will assume the functional relationship between the independent and dependent variables is linear or that it takes the form of a straight line.

[18] Instead of handpicking cases that will ensure statistical significance.