Mid-term Course Review

Basic Quantitative Methods
Spring 2021

Statistics

We began by saying that Statistics is the study of making sense of data by collecting, summarizing, and analyzing data and reporting the results. The two major branches of statistics are descriptive statistics and inferential statistics.

Descriptive Statistics

Descriptive statistics are used to describe and summarize the data at hand and involve data reduction techniques that may be tabular, graphical or numerical. This may take the form of a frequency table (tabular), or a graph (graphical), or single number summaries that describe what is typical of the data using mean/mode/ median and the amount of variability that is present using range/inter-quartile range/index of qualitative variation/variance/ standard deviation (numerical).

Inferential Statistics

On the other hand, inferential statistics is concerned with making predictions or drawing inferences about a population from observation and analyses of a sample.

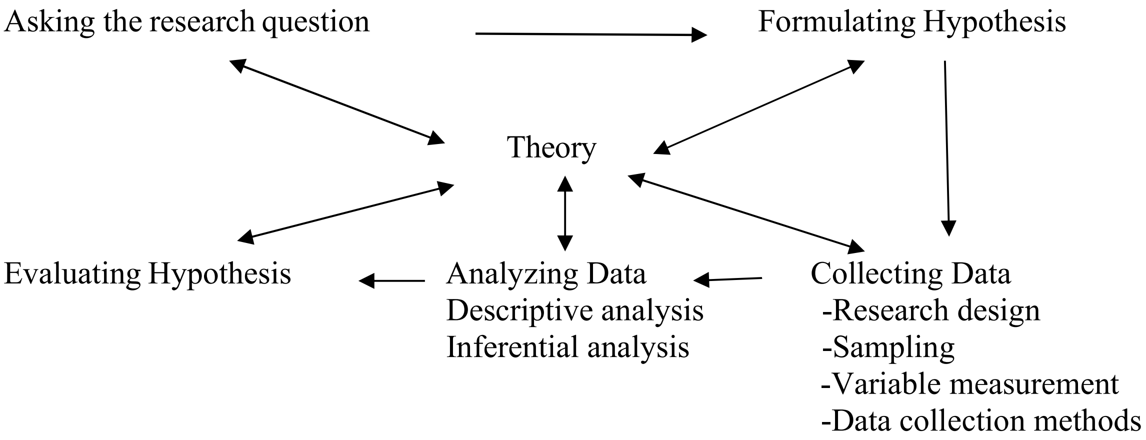
Population

A population is the total set of individuals, objects, groups, or events under consideration in a statistical study.

Sample

A sample is a relatively small subset selected from a population or that part of the population from which information is collected.

Role of Statistics in Applied Research



Asking Research Questions

Examples:

- Is there a gendered wage gap?
- Do highway tolls affect traffic patterns?
- Does urban sprawl affect health?
- Is crime more prevalent in urban areas?
- Are TODs related to transit usage?

We need to collect and analyze data before we can answer these questions—we can't answer them by relying on speculation, moral judgment or subjective preference.

Theory

A theory is a fairly elaborate explanation of the relationship between two or more variables.

Hypothesis

A hypothesis offers tentative answers to research questions.

Collecting Data

There are three basic methods of data collection:

1. Surveys - telephone, in-person, and self-administered (mail & non-mail)
2. Administrative documents & records
3. Direct observation or field research

There are advantages and disadvantages associated with each of the methods.

Variable Measurement

A variable is a property of people or objects that varies (i.e., takes on two or more values) - e.g., race, educational attainment, etc.

Our hypothesis is most often stated as relationship between two or more variables:

- The variable the researcher wants to explain is called the dependent variable.
- The variables that are expected to explain or account for the dependent variable is called the independent variable.

Cause Effect Relationships

Cause-effect relationships between variables are difficult to establish. You need to meet three criteria before causality can be properly established:

- Time order
- Correlation
- Non-spuriousness

Level of Measurement for Variables

Important to understand because the type of statistical operation will depend on this:

- Nominal level (qualitative): The nominal level is a simple categorization of observation without regard to ordering within the variables – just labeling or classifying – e.g. absence or presence of an attitude: race, gender, etc.
- Ordinal level (limited quantitative): The ordinal level categorizes a variable with an implied order. Whenever we assign numbers to rank-ordered categories ranging from low to high – e.g. social class, Likert scales, etc.
- Interval/ Ratio (quantitative): The interval/ratio level categorizes observations according to an implied order, where order is characterized by equal units or intervals between categories: number of children, temperature, age, educational attainment, etc.

Attributes of Interval/ratio Variables

Interval/ratio variables can be discrete or continuous

- Discrete variables arise from a counting process – e.g. number of children, number of visits to the health clinic
- Continuous variables arise from a measurement process – e.g. weight, height, etc.

Research Design

The research design is a set of activities that help structure a research study.

- Experimental studies are characterized by a) random assignment of individuals/subjects to experimental conditions and b) active role for the researcher in manipulating the treatment. Examples of experimental design are: completely randomized design, randomized block design, Latin square design, and factorial design.
- In observational (or non/quasi/experimental) studies, the researcher has a passive role and there is no randomization. Examples include time series comparisons, pre-post comparisons, survey research, etc.

Sampling

Only if we use probability-sampling methods can we use probability theory to make our inferences. Our purpose is two folds:

1. Representativeness
2. Assess the errors in our decisions.

Types of Sampling

- Simple random sampling
- Systematic random sampling
- Stratified random sampling
- Cluster sampling

Summarizing Data – 3 ways to do it – Tabular, Graphical, and Numerical

Depends on how data are measured.

- **Tabular** Frequency tables can be used with any level of measurement – but for variables measured at interval/ratio, the categories of the variable need to be collapsed to form class intervals. These class intervals need to be wide enough so we don’t end up with too many categories, but not too wide so as to mask meaningful differences among categories.
- **Graphical** Pie charts, bar graphs are ideal for nominal/ordinal levels of measurement. Histograms and time series graphs are ideal for interval/ratio levels of measurement.

Percentages	**Proportions**	**Rates**
Proportion x100	rel. freq. = # in each category Total #	# of actual occurrences # of possible occurrences

Data or distribution of a single variable can be summarized with respect to three characteristics – central tendency, variability and shape.

Measures of Central Tendency - Numerical summary

- Mode (most frequently occurring data value) or the category w/the most frequency
- Median – need at least ordinal level data. Locate the median by using (n+1)/2 If decimals, take the average of the two middle cases.
- Mean

Population Mean $$\mu = \frac{\sum x_i}{N}$$

Population Mean for Grouped data:

$$\mu = \frac{\sum f_i x_i}{N}$$ Where f_i is the frequency in category i.

Measures of Variability

- Range: maximum – minimum
- IQR: Q3 – Q1

  Steps:

  1) order data from low to high
  2) find median location:

$$\frac{(n+1)}{2}$$

3) find Q_1 & Q_3 using:

$$\frac{(\text{truncated median location} + 1)}{2}$$

4) find Q_1 & Q_3
5) IQR = $Q_3 - Q_1$

- IQV – Index of Qualitative Variation – only meant for variables measured at nominal level.

  $$IQV = \frac{k(n^2 - \sum f^2)}{n^2(k-1)}$$

  Where n is sample size, k is number of categories of variable and f is the frequency in each category.

- Standard Deviation: Sample standard deviation s= $\sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

population std.dev = σ. For the population standard deviation, σ , replace $n-1$ with N.

The variance is simply the standard deviation squared. Sample variance = s^2 ; population variance = σ

Shape of Distribution

Symmetrical	Right Skewed	Left Skewed
Mean = mode = median	mean > median	mean < media

Probability

- Probability is a quantitative measure of our belief that an event will occur.
- The relative frequency (the number of times a particular event occurred divided by total observations) can be interpreted as the long run average or probability of the particular event occurring.
- A probability distribution lists all possible values of a variable along with their associated probabilities.
- All random variables have probability distributions
- Probability distributions can be discrete or continuous depending on whether the variables being modeled arise from counting or measuring processes.

The normal distribution is an important distribution in statistics. Almost all of the other distributions can be approximated using the normal distribution. The normal distribution is completely defined by parameters μ and σ . Since there are many normal curves corresponding to each pair of μ and σ , we standardize our X to take on the scale of Z (which gives the number of standard deviation units X is away from the mean μ) so we can refer to one normal curve (standard normal) table to compute the probabilities with which X takes on a particular value.

Formula to go from X to z:

$$z = \frac{x - \mu}{\sigma}$$

When dealing with a sample, replace μ and σ with $\bar{x}$ and s.

Uses of Z-Scores

1. Find various areas under the normal curve
2. Comparison of scores with different means and standard deviations
3. Computing percentile rankings

Sampling Distribution of X-bar and P$_s$

All estimates or statistics are considered random variables because their value varies with the particular sample drawn. Therefore, these statistics ($\bar{x}$, s , P_s) have their own probability distribution called the sampling distribution, which tell us the probability with which these statistics assume various values, which gives us a way of assessing the accuracy of these estimates.

Central Limit theorem tells us that as long as sample sizes are large (>=30), most of these statistics (x , $\bar{P}_s$), particularly means and sums of variables, will be normally distributed, so we can use the normal distribution to calculate the various probabilities associated with the estimates.

Sampling distribution of the sample mean x-bar

$$\bar{x} \sim N\left[\mu_{\bar{x}} = \mu, \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\right]$$

The z transformation here takes on the form:

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

Sampling distribution of sample proportion Ps:

According to the Central Limit Theorem →

$$P_s \sim N\left[\mu_{P_s} = P_u, \quad \sigma_{P_s} = \sqrt{\frac{P_u(1-P_u)}{n}}\right] \qquad z = \frac{P_s - P_u}{\sqrt{\frac{P_u(1-P_u)}{n}}}$$

The particular sampling distributions that we examined in this course are: The z distribution (or the standard normal distribution)

The t distribution
The F distribution
The χ^2 distribution

Estimation

We can compute both point estimates and interval estimates.

Point estimates ($\bar{x}$, s , P_s)

Point estimates need to be:

1. Unbiased
2. Efficient
3. Consistent

Interval Estimates

You can assess the accuracy of your sample estimates by providing an interval of values that you believe with a certain level of confidence, contains the true population parameter.

Confidence Interval for mean μ

$\bar{x} \pm z_{\alpha/2} \cdot \sigma_{\bar{x}}$ where $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ (population parameter σ is known)

$\bar{x} \pm z_{\alpha/2} \cdot s_{\bar{x}}$ where $s_{\bar{x}} = \frac{s}{\sqrt{n}}$ (where population parameter σ is unknown but sample is large)

$$\bar{x} \pm t_{(\alpha/2, df)} \cdot s_{\bar{x}}$$ (Population parameter σ is unknown and sample size is small)

$\mathbf{z_{critical}} = z_{\alpha/2}$; $\alpha = 1$ – confidence level; $\mathbf{t_{crtical}} = t_{(\alpha/2, df)}$; degrees of freedom = n-1

Confidence Interval for population proportion Pu

$$p_s \pm z_{\alpha/2} \cdot \sigma_{p_s} \quad \text{where} \quad \sigma_{ps} = \sqrt{\frac{p_u(1-p_u)}{n}}$$

Sample size formula

Given tolerable error E, confidence level (1 - α) and σ, then sample size *n* is given by the formula

$$n = \left[\frac{z_{\alpha/2} \cdot \sigma}{E}\right]^2 = \frac{z_{\alpha/2}^{\;2} \cdot \sigma^2}{E^2}$$

Similarly, working with proportions for tolerable error E, confidence level (1-α), and population proportion P_u, sample size *n* is given by

$n = [\{z_{\alpha/2}\}^2 . \{P_u(1\text{-}P_u)\}] \,/ E^{\,2}$

Hypothesis Testing

Five Step Model:

1. State assumptions
2. State H_o and H_a
3. Define sample distribution, critical region(s)
4. Compute test statistic → general form is:

   $$\frac{\text{Point estimate - Hypothesized value}}{\text{Standard Error of Point Estimate}}$$
5. Make a decision and state conclusion

Hypothesis Testing for a single mean μ

1. Assumptions: normal distribution, random sample, interval/ratio level of measurement.

2. H_o: μ = some value as sample mean
   H_1: $\mu \neq$ some value (two-tailed) or
   H_1: $\mu >$ some value (one-tailed, right) or
   H_1: $\mu <$ some value (one-tailed, left)

3. Mostly use normal distribution, but for small samples we use the t-distribution with degrees of freedom: **(df) = n-1**
4. Compute test statistic:

$$z = \frac{\bar{x}-\mu}{\sigma_{\bar{y}}} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \quad \textbf{or} \quad \frac{\bar{x}-\mu}{s/\sqrt{n}}$$

For small samples:

$$t = \frac{\bar{x}-\mu}{s_{\bar{y}}} = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$

5. Decide whether to reject the null hypothesis or not and state your conclusion in terms of original problem.

Types of error

Type I error (alpha error). The probability of rejecting a null hypothesis that is, in fact, true.

Type II error (beta error). The probability of failing to reject a null hypothesis that is, in fact, false.

Decision-making and the five-step model

	Decision	
The H_o is Actually	Reject	Fail to Reject
True	Type I, or alpha error	OK
False	OK	Type II, or, beta error

Hypothesis Testing for a single proportion Pu

1. Assumptions: normal distribution, random samples, nominal level of measurement.

2. H_o: $p = p_o$, some value
   H_1: $p \neq p_o$, (two-tailed) or
   H_1: $p > p_o$, (one-tailed, right) or
   H_1: $p < p_o$, (one-tailed, left)

3. Normal distribution
4. Compute test statistic:

$$z = \frac{p_s - p_u}{\sqrt{p_u(1-p_u)/n}}$$

5. Decide whether to reject the null hypothesis and state your conclusion.

Hypothesis Testing for two independent means μ_1 - μ_2

1. Assumptions: normal distribution, independent random samples, interval/ratio level of measurement

2. H_o: $\mu_1 = \mu_2$, or $\mu_1 - \mu_2 = 0$
   H_1: $\mu_1 \neq \mu_2$
   H_1: $\mu_1 > \mu_2$
   H_1: $\mu_1 < \mu_2$

3. Use normal distribution, or the t-distribution for small samples, with degrees of freedom: **(df) =$n_1 + n_2 - 2$**

4.Test Statistics:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

or

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

For small sample sizes n_1 & n_2:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Step 5. Decide- If test statistics value of z or t > z (critical) or t (critical) then reject null hypothesis.

Hypothesis Testing for two independent proportions (Pu1 – Pu2)

1. Independent random samples, nominal level of measurement, normal sampling distribution.

2. Null Hypothesis: $p_{u_1} = p_{u2}$

   Alternate Hypothesis: p_{u_1} not equal to p_{u2} (two tailed)

   $p_{u_1} > p_{u2}$ (one tailed – right)

   $p_{u_1} < p_{u2}$ (one tailed – left)

3. Sampling distribution = Z distribution

   Alpha = (choose)

   Z(critical) = (find)

4. Test Statistic Z (observed) = $p_{s_1} - p_{s_2} \big/ \sigma_{p_1 - p_2}$

   where we use $s_{p_{s1}} - s_{p_{s2}}$ as our estimate of $\sigma_{p_1 - p_2}$ most of the time, and calculate it as follows:

   $$Z_{\text{obs}} = \frac{p_{s_1} - p_{s_2}}{s_{p_{s_1} - p_{s_2}}}$$

   Where $s_{p_{s1} - p_{s2}} = \sqrt{p_u(1-p_u)} \cdot \sqrt{(n_1 + n_2) \big/ n_1 \cdot n_2}$

   and,

   $$p_u = \frac{(n_1 \cdot p_{s_1}) + (n_2 \cdot p_{s_2})}{n_1 + n_2}$$

5. Decide. If Z(observed) falls in the critical (rejection) region, we reject the null hypothesis.

Hypothesis Testing for equality of more than two independent means

ANOVA (Analysis of Variance)

Use when the independent variable has three or more categories, and the dependent variable is Interval Ratio.

- *Significance Test Using Anova – The F Test*

The F-test is an overall test of the null hypothesis that group means on the dependent variable do not differ.

- *Hypothesis Testing*

1. Step:1 – Normal sampling distribution, Independent Random Sample, Interval/Ratio Measurement, Equal Variance
2. Step:2 – Ho: All population means are equal.
   Ha: At least one of the population means is different.
3. Step:3 - sampling distribution –F
   Alpha = (choose),
   df (between) = k-1,
   df (within) = n-k,
   F(critical) = (find)
   Where k is the number of categories of the independent variable and n is sample size.
4. Step:4 – Calculate test statistic F and the anova summary table
   - Compute group means
   - Compute grand mean
   - Calculate Within Sum of Squares
   - Calculate Between Sum of Squares
   - Calculate Mean Square Estimate
   - Calculate F = Mean Sq Estimate (between)/Mean Sq Est (within)
5. Step:5 – Decision – If F(obs) > F(critical) then reject null hypothesis.

Summarizing data from more than one variable – Contingency Tables or CrossTabs

Contingency tables can be used to determine and describe relationships between two or more variables.

- Identify Independent Variable and Dependent Variable (variables such as age, race, and sex almost always independent)
- Rules for computing % so you can understand the relationship
  1) calculate % within each category of the independent variable.
  2) interpret table by comparing % for different categories of the independent variable.

Common practice is to assign the independent variable to head the columns and the dependent variable to head the rows.

Hypothesis Testing when both independent and dependent variables are measured at nominal or ordinal level.

Chi-Square

Use when the dependent variable is nominal or ordinal
Chi-square Statistic - a measure of discrepancy existing between observed and expected frequencies
If $\chi^2 = 0$, observed and expected frequencies match exactly
Expected frequencies are calculated based on a null hypothesis.

Hypothesis Testing

- Step:1 – Independent Random Samples, Nominal Measurement
- Step:2 – Ho: The variables are independent of each other
  Ha: The variables are not independent of each other

- Step:3 - sampling distribution – χ^2
  Alpha = ____,
  df = (r-1)*(c-1),
  χ^2(critical) = ____
- Step:4 – Calculate test Statistic: $\chi^2 = \sum \frac{(ObservedFreq - ExpectedFreq)^2}{ExpectedFreq}$
- Step:5 – Decision

Coefficient of Determination, r^2: The coefficient of determination is the square of the Pearson correlation coefficient. It represents the percent of the variance in the dependent variable explained by the independent variable, or the shared variance between the two variables.

Correlation is a necessary but not sufficient condition for causality.

Regression Analysis

1. Can be used for prediction or explanation. Prediction involves predicting the value of the dependent variable from knowledge of the independent variable(s). Explanation involves explaining the dependent variable with the help of theoretically relevant independent variable(s). Explanatory uses of regression are very theory driven.
2. Given a set of data, the logic of regression is to estimate the strength, direction and nature of the relationship between the dependent and the independent variable(s), assuming that the relationship is functionally linear (that is, it can be described by a line in two dimensions or a plane or hyperplane in higher dimensions), and by using the principle of least squares, we fit the best fitting line/plane/hyperplane. The principle of least squares (also called Ordinary Least Squares or OLS) is to minimize the sum of squared distances between the data points and the line (or plane/hyperplane).
3. This process provides us with an equation describing the nature of the relationship between dependent and independent variable(s).

Simple regression – one independent (predictor) variable

Population Regression Equation: $Y_i = \alpha + \beta X_i + \varepsilon_i$

Sample Regression Equation: $Y_i = a + bX_i + e$

R-squared is a measure we can use to assess how good the estimated model fits the data. It also tells us the proportionate reduction in prediction error of the dependent variable given one or more independent variables. In this context, you compute one set of errors without the knowledge of the independent variable(s) and another set with the knowledge of independent variable(s). The amount of difference between these two types of errors as a proportion of the first type of error, then gives you R-squared.

Where a = intercept or where the regression line crosses the Y axis or the value of Y when X = 0; b = slope coefficient, or the change in Y associated with a unit change in X; and ε = error or all that is not captured by the model.

Population Prediction Equation: $E(Y|X_i) = \alpha + \beta X_i$

Sample Prediction Equation: $\hat{Y} = a + bX_i$

Least squares solution for *a* and *b*:

$$a = \overline{Y} - b\overline{X}$$

$$b = \frac{\sum (X_i - \overline{X}) \cdot (Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$ = Covariance of X and Y/ Variance of X

a and *b* are sample estimates of population parameters α and β.

Multiple regression is an extension of regression when we include more than one independent (predictor) variable.

Population Regression Equation: $Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots. + \varepsilon_i$

Sample Regression Equation: $Yi = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots\ldots + e_i$

Where a = intercept; b1 = slope coefficient associated with X1 or the effect of X1 on Y controlling for all the other Xs, b2 = slope coefficient associated with X2, controlling for all the other Xs, and so on., and e = error

Population Prediction Equation: $E(Y|X) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots. + \beta_k X_k$

Sample Prediction Equation: $\hat{Y} = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots\ldots + b_k Xk$

Hypothesis Testing: (1) Perform an overall test to see if any of the independent variables has a significant relationship with the dependent variable. We do this using the F statistic given as part of the output, testing the Ho: $\beta_1 = \beta_2 = \ldots = \beta k = 0$, against the alternative hypothesis that at least one of these coefficients is non-zero. If the significance (or p-value) associated with this F is less than 0.05, we can reject the null and conclude that there is at least one useful independent variable. This hypothesis test is also equivalent to testing whether the $R^2 = 0$, i.e., the independent variables, as a set, do not explain or predict any part of the dependent variable's variability.

An alternate interpretation revolves around explaining the variability in the dependent variables. R-squared in this case tells us the proportion of variance in the dependent variable that is explained or accounted for by the independent variable(s). If the total variability in the dependent variable can be standardized to be 1 or 100 percent, then R-squared is the proportion or percent of the total that is explained by the regression and so R-squared ranges from 0 to 1 or 0 to 100 percent, with 1 or 100 percent indicating perfect explanation. The total variability in the dependent variable or the total sum of squares $\sum(Y - \overline{Y})^2$ can be decomposed or partitioned into two other sums of squares, one, the sum of squares due to regression (that part that gives us the R-squared: $\sum(\hat{Y} - \overline{Y})^2$ and two, the sum of squares of errors or residuals:- $\sum Y - \hat{Y})^2$. So R-squared can also be calculated as the ratio of $\frac{\sum(\hat{Y} - \overline{Y})^2}{\sum(Y - \overline{Y})^2}$

(2) We can perform hypothesis tests of the slope coefficients, with the null hypothesis being that the population slope for the particular independent variable is equal to 0, that is, there is no relationship between the particular independent variable and the dependent variable. To perform the hypothesis test, we calculate the test statistic t by dividing the slope coefficient estimate by its standard error – and if this t value is greater than or equal to 2, we conclude that the variable associated with the slope coefficient has a statistically significant association with the dependent variable.

Having assessed model fit through (1) and the significance of the effect of each independent variable through (2), we can now proceed to interpret each coefficient, i.e., discuss the magnitude and direction of the effect of each independent variable.

Multiple Regression

Least square multiple regression equation with two independent variables

$Y = a + b_1X_1 + b_2X_2$

Where b_1 = the partial slope of the linear relationship between the first independent variable and Y;
b 2 = the partial slope of the linear relationship between the second independent variable and Y

Partial Slopes calculation

A major difference between the multiple and bivariate regression equations concerns the slopes (b's). In the case of multiple regression, the b's are called partial slopes, and they show the amount of change in Y for a unit change in the independent variable while controlling for the effects of the other independent variables in the equation. The partial slopes are thus analogous to partial correlation coefficients and represent the direct effect of the associated independent variable on Y.

The partial slopes for the independent variables are determined by formulas below

$b_1 = (s_y/s_1)\ [(r_{y1} – r_{y2}\ r_{12})/(1-r_{12}^2)]$

$b_2 = (s_y/s_2)\ [(r_{y2} – r_{y1}\ r_{12})/(1-r_{12}^2)]$

where
b_1 =the partial slope of X_1 on Y
b_2 =the partial slope of X_2 on Y
s_y = the standard deviation of Y
s_1 = the standard deviation of the first independent variable X_1
s_2 = the standard deviation of the second independent variable X_2
r_{y1} = the bivariate correlation between Y and X_1
r_{y2} = the bivariate correlation between Y and X_2
r_{12} = the bivariate correlation between X_1 and X_2