Unit III - DATA ANALYSIS AND REPORTING

Analyzing and Interpreting Information

Data analysis is the process of collecting, modeling, and analyzing data to extract insights that support decision-making. There are several methods and techniques to perform analysis depending on the industry and the aim of the analysis.

Analyzing quantitative and qualitative data is often the topic of advanced research and evaluation methods courses. However, there are certain basics which can help to make sense of reams of dta.

Always start with your research goals

When analyzing data (whether from questionnaires, interviews, focus groups, or whatever), always start from review of your research goals, i.e., the reason you undertook the research in the first place. This will help you organize your data and focus your analysis. For example, if you wanted to improve a program by identifying its strengths and weaknesses, you can organize data into program strengths, weaknesses and suggestions to improve the program. If you wanted to fully understand how your program works, you could organize data in the chronological order in which customers or clients go through your program. If you are conducting a performance improvement study, you can categorize data according to each measure associated with each overall performance result, e.g., employee learning, productivity and results.

Basic analysis of "quantitative" information

(for information other than commentary, e.g., ratings, rankings,
yes's, no's, etc.):

Make copies of your data and store the master copy away. Use the copy for making edits, cutting and pasting, etc.

Tabulate the information, i.e., add up the number of ratings, rankings, yes's, no's for each question.

For ratings and rankings, consider computing a mean, or average, for each question. For example, "For question #1, the average ranking was 2.4". This is more meaningful than indicating, e.g., how many respondents ranked 1, 2, or 3.

Consider conveying the range of answers, e.g., 20 people ranked "1", 30 ranked "2", and 20 people ranked "3".

Basic analysis of "qualitative" information

(Respondents' verbal answers in interviews, focus groups, or
written commentary on questionnaires):

Read through all the data.

Organize comments into similar categories, e.g., concerns, suggestions, strengths, weaknesses, similar experiences, program inputs, recommendations, outputs, outcome indicators, etc.

Label the categories or themes, e.g., concerns, suggestions, etc.

Attempt to identify patterns, or associations and causal relationships in the themes, e.g., all people who attended programs in the evening had similar concerns, most people came from the same geographic area, most people were in the same salary range, what processes or events respondents experience during the program, etc.

Keep all commentary for several years after completion in case needed for future reference.

Interpreting information

Attempt to put the information in perspective, e.g., compare results to what you expected, promised results; management or program staff; any common standards for your products or services; original goals (especially if you're conducting a program evaluation); indications or measures of accomplishing outcomes or results (especially if you're conducting an outcomes or performance evaluation); description of the program's experiences, strengths, weaknesses, etc. (especially if you're conducting a process evaluation).

Consider recommendations to help employees improve the program, product or service; conclusions about program operations or meeting goals, etc.

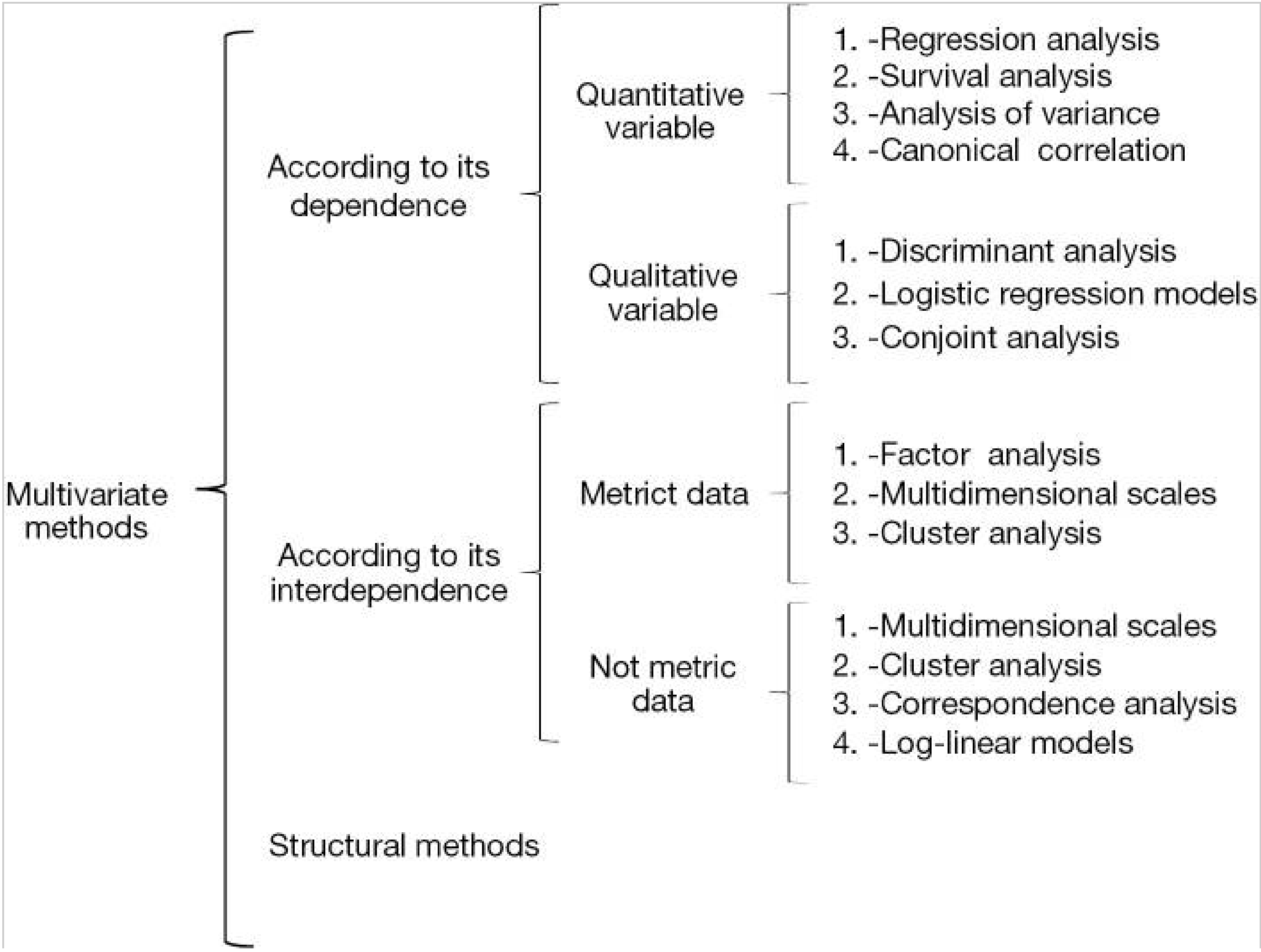
Record conclusions and recommendations in a report, and associate interpretations to justify your conclusions or recommendations.

Reporting Results

1. The level and scope of content depends on to whom the report is intended, e.g., to funders / bankers, employees, clients, customers, the public, etc.
2. Be sure employees have a chance to carefully review and discuss the report. Translate recommendations to action plans, including who is going to do what about the research results and by when.
3. Funders / bankers will likely require a report that includes an executive summary (this is a summary of conclusions and recommendations, not a listing of what sections of information are in the report — that's a table of contents); description of the organization and the program, product, service, etc., under evaluation; explanation of the research goals, methods, and analysis procedures; listing of conclusions and recommendations; and any relevant attachments, e.g., inclusion of research questionnaires, interview guides, etc. The funder may want the report to be delivered as a presentation, accompanied by an overview of the report. Or, the funder may want to review the report alone.
4. Be sure to record the research plans and activities in a research plan which can be referenced when a similar research effort is needed in the future.

Overview of Multivariate analysis

Multivariate means involving multiple dependent variables resulting in one outcome. This explains that the majority of the problems in the real world are Multivariate. For example, we cannot predict the weather of any year based on the season. There are multiple factors like pollution, humidity, precipitation, etc. Here, we will introduce you to multivariate analysis, its history, and its application in different fields.



Multivariate analysis (MVA) is a Statistical procedure for analysis of data involving more than one type of measurement or observation. It may also mean solving problems where more than one dependent variable is analyzed simultaneously with other variables.

Types of Multivariate Analyses To Be Taught

- Multiple linear regression: A linear regression method where the dependent variable Y is described by a set of X independent variables. An example would be to determine the factors that predict the selling price or value of an apartment.

- Multiple linear correlation: Allows for the determination of the strength of the strength of the linear relationship between Y and a set of X variables.

- Multivariate nonlinear regression: A form of regression analysis in which the dependent variable Y is described by a nonlinear combination of the independent variables X.

Response Surface Regression: A form of multivariate non-linear regression where the influences of several independent or “response” variables on a dependent variable are determined. The goal of response surface regression is to optimize a response.

- Discriminant analysis: In an original survey of males for possible factors that can be used to predict heart disease, the researcher wishes to determine a linear function of the many putative causal factors that would be useful in predicting those individuals that would be likely to have a heart attack within a 10-year period.

- Principal component analysis (PCA): Is used to simplify the description of a set of interrelated variables. PCA considers all variables equally; they are not divided into dependent and independent variables. In PCA, the interrelated variables are in essence transformed into new, uncorrelated values. Using the data from the lung function example, the data for each individual are highly interrelated since they were all recorded on one breath. Because the data are interrelated, you need to use a method that develops a new set of measurements that are uncorrelated with each other. PCA allows development of new uncorrelated measurements called principal components. It is hoped that the first 2-3 of the principal components can be used to explain the original variation in lung function. Use of PCA may allow you to use fewer principal components than the number of variables in the original data set and help to simply the interpretation and explanation of the results.

- Factor analysis: Is similar to PCA in that it allows one to determine the interrelationships among a set of variables. Like PCA, factor analysis does not have a dependent variable that is described by a set of independent variables. Using our political survey example, factor analysis will allow you to group each of the questions into subgroups that are uncorrelated with each other.

- Cluster analysis: Is a method for grouping individuals or objects into unknown groups. This method differs from discriminant analysis in that the number and the characteristics of the groups are unknown prior to the analysis.

Types of Variables

1. Qualitative variable:
    a. • One in which numerical measurement is not possible.
    b. • An observation is made when an individual is assigned to one of several mutually exclusive categories (i.e. cannot be assigned to more than one category).
    c. • Non-numerical data.
    d. • Observations can be neither meaningfully ordered nor measured (e.g. hair color, resistance vs. susceptibility to a pathogen, etc.).
2. Quantitative variable:
    a. • One in which observations can be measured.
    b. • Observations have a natural order of ranking.
    c. • Observations have a numerical value (e.g. yield, height, enzyme activity, etc.)

• Quantitative variables can be subdivided into two classes:

1. Continuous: One in which all values in a range are possible (e.g. yield, height, weight, etc.).

2. Discrete: One in which all values in a range are not possible, often counting data (number of insects, lesions, etc.).

Steven's Classification of Variables

• Stevens (1966)1 developed a commonly accepted method of classifying variables.

1. Nominal variable:
   a. Each observation belongs to one of several distinct categories.
   b. • The categories don't have to be numerical.
   c. • Examples are sex, hair color, race, etc.
2. Ordinal variable:
   a. • Observations can be placed into categories that can be ranked.
   b. • An example would be rating for disease resistance using a 1-10 scale, where 1=very resistant and 10=very susceptible.
   c. • The interval between each value in the scale is not certain.
3. Interval variables:
   a. • Differences between successive values are always the same.
   b. • Examples would be temperature and date.
4. Ratio variables:
   a. • A type of interval variable where there is a natural zero point or origin of measurement.
   b. • Examples would be height and weight.
   c. • The difference between two interval variables is a ratio variable.

THE OBJECTIVE OF MULTIVARIATE ANALYSIS

MVA or Multivariate Analysis considers multiple factors. The objectives of MVA are listed below.

- Reduction in data or simplification of the structure

  MVA helps to simplify the data as much as possible without losing out on the critical information. This aids in drawing interpretation later.

- Grouping and Sorting the data

  MVA has multiple variables. The variables are grouped based on their unique features.

- Data is verified based on the variables

  Understanding the variables and collected data is verified. Concluding, the state of the variables is critical. The variables can be independent or dependent on the other variables.

- Establishing a connection between the variables

  The relationship between the variables is vital to understand the behavior of the variables based on observations and other variables present.

- Testing and construction of hypothesis

  Creating a statistical hypothesis based on the parameters of the multivariate data is tested. This testing is done to understand if the assumptions are correct or not.

Advantages and Disadvantages of Multivariate Analysis

Advantages

- The main advantage of multivariate analysis is that since it considers more than one factor of independent variables that influence the variability of dependent variables, the conclusion drawn is more accurate.
- The conclusions are more realistic and nearer to the real-life situation.
- MVA considers multiple variables. These variables can be independent or dependent on each other. The analysis considers the factors and draws an accurate conclusion.
- The analysis is tested and conclusions are drawn. The drawn conclusions are close to real-life situations.

Disadvantages

- The main disadvantage of MVA includes that it requires rather complex computations to arrive at a satisfactory conclusion.
- Many observations for a large number of variables need to be collected and tabulated; it is a rather time-consuming process.
- MVA is laborious and as it includes complex computations.
- The analysis requires a huge amount of observations for multiple variables that are collected and tabulated. This observation process is time-consuming.

Hypotheses testing

Hypothesis tests are statistical tools widely used for assessing whether or not there is an association between two or more variables. These tests provide a probability of the type 1 error (p-value), which is used to accept or reject the null study hypothesis.

Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis.

Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process. The word "population" will be used for both of these cases in the following descriptions.

- Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data.
- The test provides evidence concerning the plausibility of the hypothesis, given the data.
- Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed.

Hypothesis Testing Works

In hypothesis testing, an analyst tests a statistical sample, with the goal of providing evidence on the plausibility of the null hypothesis.

Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed. All analysts use a random population sample to test two different hypotheses: the null hypothesis and the alternative hypothesis.

The null hypothesis is usually a hypothesis of equality between population parameters; e.g., a null hypothesis may state that the population mean return is equal to zero. The alternative hypothesis is effectively the opposite of a null hypothesis (e.g., the population mean return is not equal to zero). Thus, they are mutually exclusive, and only one can be true. However, one of the two hypotheses will always be true.

Hypothesis testing is a formal procedure for investigating our ideas about the world using statistics. It is most often used by scientists to test specific predictions, called hypotheses, that arise from theories.

There are 5 main steps in hypothesis testing:

1. State your research hypothesis as a null hypothesis (Ho) and alternate hypothesis (Ha or H1).
2. Collect data in a way designed to test the hypothesis.
3. Perform an appropriate statistical test.
4. Decide whether to reject or fail to reject your null hypothesis.

Present the findings in your results and discussion section.

Step 1: State your null and alternate hypothesis

After developing your initial research hypothesis (the prediction that you want to investigate), it is important to restate it as a null (Ho) and alternate (Ha) hypothesis so that you can test it mathematically.

The alternate hypothesis is usually your initial hypothesis that predicts a relationship between variables. The null hypothesis is a prediction of no relationship between the variables you are interested in.

You want to test whether there is a relationship between gender and height. Based on your knowledge of human physiology, you formulate a hypothesis that men are, on average, taller than women. To test this hypothesis, you restate it as:

Ho: Men are, on average, not taller than women.
Ha: Men are, on average, taller than women.

Step 2: Collect data

For a statistical test to be valid, it is important to perform sampling and collect data in a way that is designed to test your hypothesis. If your data are not representative, then you cannot make statistical inferences about the population you are interested in.

To test differences in average height between men and women, your sample should have an equal proportion of men and women, and cover a variety of socio-economic classes and any other control variables that might influence average height.

You should also consider your scope (Worldwide? For one country?) A potential data source in this case might be census data, since it includes data from a variety of regions and social classes and is available for many countries around the world.

Step 3: Perform a statistical test

There are a variety of statistical tests available, but they are all based on the comparison of within-group variance (how spread out the data is within a category) versus between-group variance (how different the categories are from one another).

If the between-group variance is large enough that there is little or no overlap between groups, then your statistical test will reflect that by showing a low p-valrue. This means it is unlikely that the differences between these groups came about by chance.

Alternatively, if there is high within-group variance and low between-group variance, then your statistical test will reflect that with a high p-value. This means it is likely that any difference you measure between groups is due to chance.

Your choice of statistical test will be based on the type of data you collected.

Based on the type of data you collected, you perform a one-tailed t-test to test whether men are in fact taller than women. This test gives you:

an estimate of the difference in average height between the two groups.

a p-value showing how likely you are to see this difference if the null hypothesis of no difference is true.

Your t-test shows an average height of 175.4 cm for men and an average height of 161.7 cm for women, with an estimate of the true difference ranging from 10.2cm to infinity. The p-value is 0.002.

Step 4: Decide whether to reject or fail to reject your null hypothesis

Based on the outcome of your statistical test, you will have to decide whether to reject or fail to reject your null hypothesis.

In most cases you will use the p-value generated by your statistical test to guide your decision. And in most cases, your predetermined level of significance for rejecting the null hypothesis will be 0.05 – that is, when there is a less than 5% chance that you would see these results if the null hypothesis were true.

In some cases, researchers choose a more conservative level of significance, such as 0.01 (1%). This minimizes the risk of incorrectly rejecting the null hypothesis (Type I error).

In your analysis of the difference in average height between men and women, you find that the p-value of 0.002 is below your cutoff of 0.05, so you decide to reject your null hypothesis of no difference.

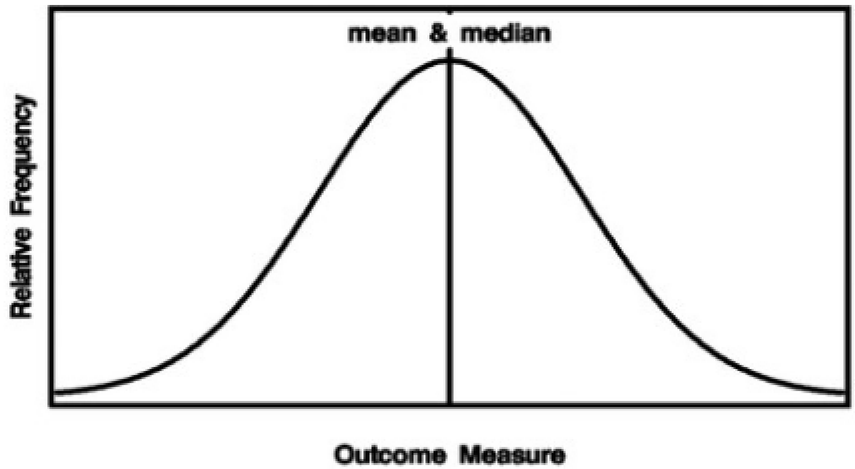
Step 5: Present your findings

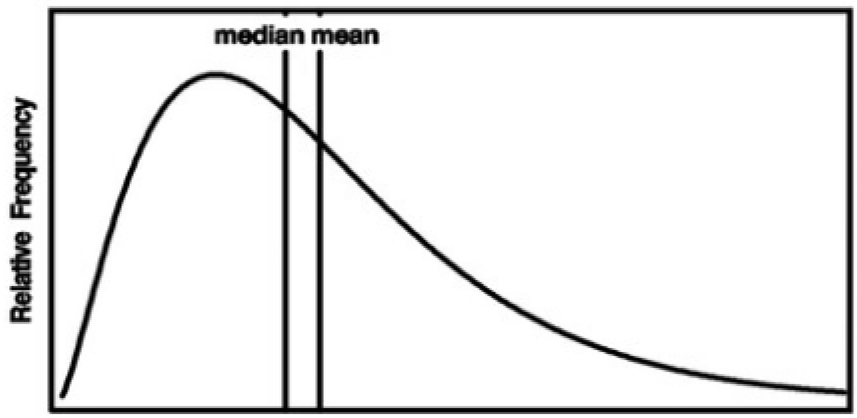
The results of hypothesis testing will be presented in the results and discussion sections of your research paper.

In the results section you should give a brief summary of the data and a summary of the results of your statistical test (for example, the estimated difference between group means and associated p-value). In the discussion, you can discuss whether your initial hypothesis was supported by your results or not.

In the formal language of hypothesis testing, we talk about rejecting or failing to reject the null hypothesis. You will probably be asked to do this in your statistics assignments.

In our comparison of mean height between men and women we found an average difference of 13.7 cm and a p-value of 0.002; therefore, we can reject the null hypothesis that men are not taller than women and conclude that there is likely a difference in height between men and women.





Real-World Example of Hypothesis Testing

If, for example, a person wants to test that a penny has exactly a 50% chance of landing on heads, the null hypothesis would be that 50% is correct, and the alternative hypothesis would be that 50% is not correct.

Mathematically, the null hypothesis would be represented as Ho: P = 0.5. The alternative hypothesis would be denoted as "Ha" and be identical to the null hypothesis, except with the equal sign struck-through, meaning that it does not equal 50%.

A random sample of 100 coin flips is taken, and the null hypothesis is then tested. If it is found that the 100 coin flips were distributed as 40 heads and 60 tails, the analyst would assume that a penny does not have a 50% chance of landing on heads and would reject the null hypothesis and accept the alternative hypothesis.

If, on the other hand, there were 48 heads and 52 tails, then it is plausible that the coin could be fair and still produce such a result. In cases such as this where the null hypothesis is "accepted," the analyst states that the difference between the expected results (50 heads and 50 tails) and the observed results (48 heads and 52 tails) is "explainable by chance alone."

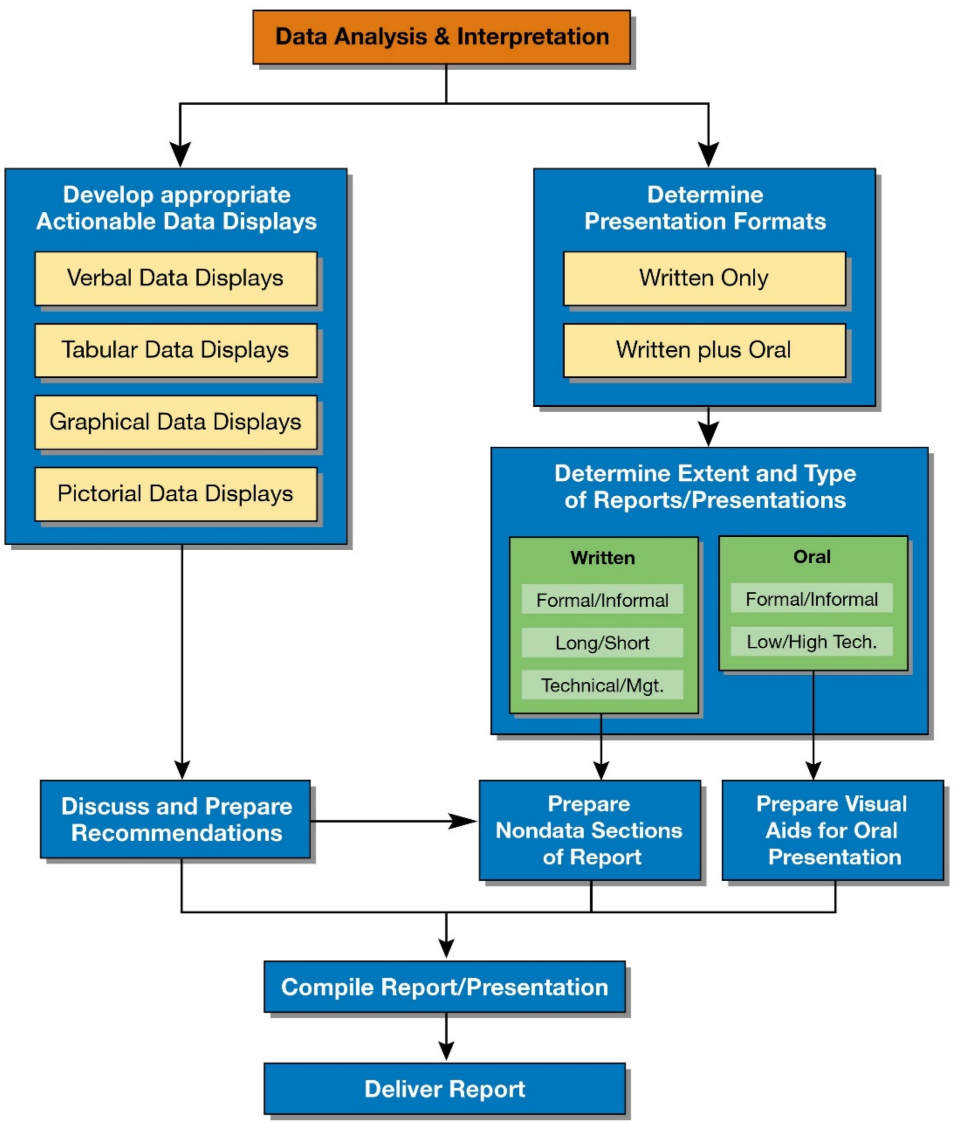
Measure of Association

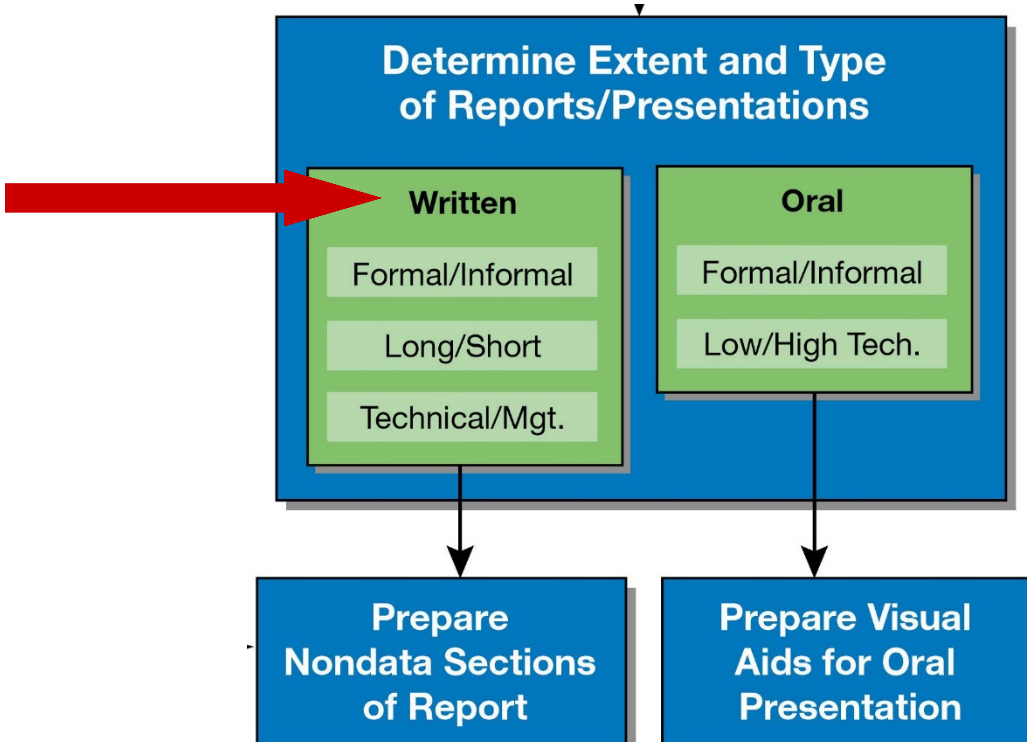
Association

An association is present if probability of

Presenting Insights and Findings: Written and Oral Reports

- That a quality presentation of research findings can have an inordinate effect on a reader's or a listener's perceptions of a study's quality.
- The contents, types, lengths, and technical specifications of research reports.
- That the writer of a research report should be guided by questions of purpose, readership, circumstances/ limitations, and use.
- That while some statistical data may be incorporated into the text, most statistics should be placed in tables, charts, or graphs.
- That oral presentations of research findings should be developed with concern for organization, visual aids, and delivery in unique communication settings





Guidelines for Short Reports

Guidelines for Short Reports

Remind reader of request

Write in an expository style

Write report and hold for review

Attach detailed materials in appendix

Components:
Short Report: Memo or Letter-Style

1. Introduction
   - Problem statement
   - Research objectives
   - Background
2. Conclusions
   - Summary and conclusions
   - Recommendations

Components:

Short Report: Technical

1. Prefatory Information (all)
2. Introduction (all, plus brief methodology and limitations)
3. Findings
4. Conclusions
5. Appendices

Components:
Long, Report: Management

1. Prefatory Information (all)
2. Introduction (all, plus brief methods and limitations)
3. Conclusions and Recommendations
4. Findings
5. Appendices

Components:
Long Report: Technical

1. Prefatory Information
2. Introduction
3. Methodology (full, detailed)
4. Findings
5. Conclusions
6. Appendices
7. Bibliography

Findings:

1. In this city, *commercial banks are not the preferred savings medium.* Banks are in a weak third place behind money market accounts.
2. Customers of the Central City Bank have a *somewhat more favorable attitude toward bank savings* and less of a preference for government bonds.

Question: Suppose that you have just received an extra $1,000 and have decided to save it. Which of the savings methods listed would be your preferred way to save it?

- ❒ Government bonds
- ❒ Savings and loan
- ❒ Bank savings
- ❒ Credit union
- ❒ Stock
- ❒ Other

Savings Method	Total Replies	Central City Bank Customers	Other Bank Customers
Government bonds	24%	20%	29%
Savings and loan	43	45	42
Bank	13	18	8
Credit union	9	7	11
Stock	7	8	5
Other	4	2	5
Total	100%	100%	100%
	n = 216	n = 105	n = 111