Applied Statistics
Desktop Reference

Phil Crewson
Preface

Approach Used in this Desktop Reference

The Applied Statistics Desktop Reference Guide provides an overview of hypothesis testing and the interpretation of output generated by AcaStat statistical software. The Guide will be most useful for those who have completed a course in statistics and need a quick refresher or for students who would like to have a reference to support course instruction. It is not intended to serve as a stand-alone statistical text.

The approach of the Guide is to connect classically taught statistics with statistical software output. The Guide presents commonly used hypothesis tests and formulas, provides an example of how to do the calculations, and relates the hypothesis test to output generated by statistical software.

For many, the most valuable portions of the Guide will be the Glossary and the annotated presentation of output from statistical software.

AcaStat Statistical Software

The software output examples provided in the text and the Glossary should be familiar to anyone who has used one of the popular statistical software packages. Most of the software output presented in the Guide is created with AcaStat statistical software. AcaStat is an easy-to-use statistical software system that analyzes raw data and summary statistics. Easily create data files, drag and drop tables from spreadsheets, or import data from delimited text files. AcaStat simplifies the creation of
crosstabulations, descriptive statistics, correlation, and common significance tests without learning programming code or complex interface commands.

The AcaStat statistical software application is available for a small fee at http://www.acastat.com or the Mac App Store. AcaStat is compatible with Windows (Vista or later) or Mac OS X (10.7 or later) operating systems.

Glossary

The Glossary contains over 200 definitions of statistical terms and concepts and includes a quick ‘Applied Stat’ review of 26 common statistical techniques with annotated statistical output examples.

About the Author

Philip E. Crewson, Ph.D. taught undergraduate and graduate statistics at American University for 15 years. He was co-editor of a monthly series in Radiology on biostatistics for radiologists and has served as a Director of Research in government and non-profit organizations and as a peer reviewer for academic journals. The content of this Guide is an outline of academic course material used in undergraduate and graduate university courses.

Copyright

Copyright © 2016 by Philip E. Crewson. All rights reserved under International and Pan-American Copyright Conventions. You have been granted the non-exclusive, non-transferable right to access and read the text of this e-book on-screen. No part of this text may be reproduced, transmitted, down-loaded, decompiled, reverse engineered, or stored in or introduced into any information storage and retrieval system, in any form or by any means, whether electronic or mechanical, now known or hereinafter invented, without the express written permission of the author.
1 Research Design and Reporting

1.1 Research Design

The scientific model guides research design to ensure findings are quantifiable (measured in some fashion), verifiable (others can substantiate our findings), replicable (others can repeat the study), and defensible (provides results that are credible to others--this does not mean others have to agree with the results). For some the scientific model may seem too complex to follow, but it is often used in everyday life and should be evident in any research report, paper, or published manuscript. The corollaries of common sense and proper paper format with the scientific model are given below.

*Scientific Model, Common Sense, and Paper Format Corollaries*

<table>
<thead>
<tr>
<th>Scientific Model</th>
<th>Common Sense</th>
<th>Paper Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research Question</td>
<td>Why</td>
<td>Introduction</td>
</tr>
<tr>
<td>Develop a theory</td>
<td>Your answer</td>
<td>Introduction</td>
</tr>
<tr>
<td>Identify variables</td>
<td>How</td>
<td>Method</td>
</tr>
<tr>
<td>Identify hypotheses</td>
<td>Expectations</td>
<td>Method</td>
</tr>
<tr>
<td>Test hypotheses</td>
<td>Analyze data</td>
<td>Results</td>
</tr>
<tr>
<td>Evaluate the results</td>
<td>What it means</td>
<td>Conclusion</td>
</tr>
<tr>
<td>Critical review</td>
<td>What it doesn’t mean</td>
<td>Conclusion</td>
</tr>
</tbody>
</table>
Overview of first four elements of the Scientific Model

The following discussion provides a very brief introduction to the first four elements of the scientific model. Although these four elements are not the primary focus of this text, they are the foundations for systematic research and data analysis and should be carefully considered before, during, and after statistical analyses. The remaining elements that pertain to testing hypotheses and evaluating results are the primary focus for the remainder of the Desktop Reference Guide.

1. Research Question

The research question should be a clear statement about what the researcher intends to investigate. It should be specified before research is conducted and openly stated in reporting the results. One conventional approach is to put the research question in writing in the introduction of a report starting with the phrase "The purpose of this study is . . . ." This approach forces the researcher to clearly identify the research objective, allow others the to benchmark how well the study design answers the primary goal of the research, and identify the key abstract concepts involved in the research

Abstract concepts: The starting point for measurement. Abstract concepts are best understood as general ideas in linguistic form that help us describe reality. They range from the simple (hot, long, heavy, fast) to the more difficult (responsive, effective, fair). Abstract concepts should be evident in the research question and/or purpose statement. An example of a research question is given below along with how it might be reflected in a purpose statement.

Research question: Is the quality of public sector and private sector employees different?

Purpose statement: The purpose of this study is to determine if the quality of public and private sector employees is different.
2. **Develop Theory**

A theory is one or more propositions that suggest why an event occurs. It is the researcher’s explanation for how the world works. These theoretical propositions provide a framework for analysis that are predisposed to determine “What is reality” not “What should reality be.” A sound theory must have logical integrity and should consider current knowledge on the event being explored. In other words, a thorough literature review before the design and implementation of a study is the hallmark of good research.

3. **Identify Variables**

Variables are measurable abstract concepts that help describe relationships. This measuring of abstract concepts is referred to as operationalization. In the previous research question "Is the quality of public sector and private sector employees different?” the key abstract concepts are employee quality and employment sector. To measure "quality" a measurable representation of employee quality will need to be found. Possible quality variables may be performance on a standardized intelligence test, attendance at work, performance evaluations, etc. The variable for employment sector seems to be fairly self-evident, but a good researcher must be very clear on how they define and measure the concepts of public and private sector employment (i.e., how is non-profit employment handled?).

*Variables* represent empirical indicators of an abstract concept. This does not mean there is complete congruence between our measure and the abstract concept. Variables used in statistical analysis are unlikely to measure all aspects of an abstract concept. Put simply, all variables have an error component.

\[
\text{Abstract concept} = \text{indicator} + \text{error}
\]

Because there is always error in measurement, multiple measures/indicators of one abstract concept are felt to be better (more valid and more reliable) than one. As
shown below, one would expect that as more valid indicators of an abstract concept are used, the effect of the error term would decline:

\[ \text{Abstract concept} = \text{indicator1} + \text{indicator2} + \text{indicator3} + \text{error} \]

The appropriate design and use of multiple indicators is beyond this text, but this approach is evident in many tools commonly used today such as psychological assessments, employee satisfaction, socio-economic status, and work motivation measures.

4. Identify measurable hypotheses

A hypothesis is a formal statement that presents the expected relationship between an independent and dependent variable. A dependent variable is a variable that contains variations the researcher seeks to explain. An independent variable is a variable that is thought to affect (cause) variations in the dependent variable. Although causation is implied when statistically significant associations are identified between independent and dependent variables, causation must not rely solely on the empirical evidence. Proof (or causation) must always be an exercise in rational inference that considers empirical evidence but also includes other factors such as study design, prior research, and the basic elements required for causation (discussed later).

Levels of Data

There are four levels of variables. It is essential to be able to identify the levels of data used in a research design. The level of data used in a research design directly impacts which statistical methods are most appropriate for testing research hypotheses. The four levels of data are listed below in order of their precision.

Nominal: Classifies objects by type or characteristic (sex, race, models of vehicles, political jurisdictions).

Properties:
• categories are mutually exclusive (an object or characteristic can only be contained in one category of a variable)

• no logical order

**Ordinal:** Classifies objects by type or kind but also has some logical order (military rank, letter grades).

*Properties:*

• categories are mutually exclusive

• logical order exists

• scaled according to amount of a particular characteristic they possess

**Interval:** Classifies by type, logical order, but also requires that differences between levels of a category are equal (temperature in degrees Celsius, distance in kilometers, age in years).

*Properties:*

• categories are mutually exclusive

• logical order exists

• scaled according to amount of a particular characteristic they possess

• differences between each level are equal

• no zero starting point

**Ratio:** Classification is similar to interval but ratio data has a true zero starting point (total absence of the characteristic). For most purposes, the analyses presented in this text assume interval/ratio are the same.
Association

Statistical techniques are used to explore connections between independent and dependent variables. This connection between or among variables is often referred to as association. **Association** is also known as covariation and can be defined as measurable changes in one variable that occur concurrently with changes in another variable. A **positive association** is represented by change in the same direction (income rises with education level). **Negative association** is represented by concurrent change in opposite directions (hours spent exercising and % body fat). **Spurious associations** are associations between two variables that can be better explained by a third variable. As an example, if after taking cold medication for seven days the symptoms disappear, one might assume the medication cured the illness. Most of us, however, would probably agree that the change experienced in cold symptoms are probably better explained by the passage of time rather than pharmacological effect (i.e., the cold would resolve itself in seven days irregardless of whether the medication was taken or not).

Causation

There is a difference between determining association and causation. Causation, often referred to as a relationship, cannot be proven with statistics. Statistical techniques provide evidence that a relationship exists through the use of significance testing and strength of association metrics. However, this evidence must be bolstered by an intellectual exercise that includes the theoretical basis of the research and logical assertion. The following presents the basic elements necessary for claiming causation:

**Required elements for causation**

1. **Association:** Do the variables covary?

2. **Precedence:** Does the independent variable vary before the dependent variable?
3. **Plausibility**: Is the expected outcome consistent with theory/prior knowledge?

4. **Nonspuriousness**: Are no other variables more important?

**Reliability and Validity**

The accuracy of our measurements is affected by their reliability and validity. **Reliability** is the extent to which the repeated use of a measure obtains the same values when no change has occurred (can be evaluated empirically). **Validity** is the extent to which the operationalized variable accurately represents the abstract concept it intends to measure (cannot be confirmed empirically—it will always be in question). Reliability negatively impacts all studies but is very much a part of any methodology/operationalization of concepts. As an example, reliability can depend on who performs the measurement and when, where, and how data are collected (from whom, written, verbal, time of day, season, current public events).

There are several different conceptualizations of validity. **Predictive validity** refers to the ability of an indicator to correctly predict (or correlate with) an outcome (e.g., GRE and performance in graduate school). **Content validity** is the extent to which the indicator reflects the full domain of interest (e.g., past grades only reflect one aspect of student quality). **Construct validity** (correlational validity) is the degree to which one measure correlates with other measures of the same abstract concept (e.g., days late or absent from work may correlate with performance ratings). **Face validity** evaluates whether the indicator appears to measure the abstract concept (e.g., a person's religious preference is unlikely to be a valid indicator of employee quality).

**Study Design**

There are two types of study designs; experimental and quasi-experimental.

**Experimental**: The experimental design uses a control group and applies treatment to a second group. It provides the strongest evidence of causation through extensive
controls and random assignment to remove other differences between groups. Using the evaluation of a job training program as an example, one could carefully select and randomly assign two groups of unemployed welfare recipients. One group would be provided job training and the other would not. If the two groups are similar in all other relevant characteristics, one could assume any differences between the groups employment one year later was caused by job training.

Whenever an experimental design is used, both internal and external validity can become very important factors.

**Internal validity:** The extent to which accurate and unbiased association between the IV and DVs were obtained in the study group.

**External validity:** The extent to which the association between the IV and DV is accurate and unbiased in populations outside the study group.

**Quasi-experimental:** The quasi-experimental design does not have the level of controls employed in an experimental design (most social science research). Although internal validity is lower than can be obtained with an experimental design, external validity is generally better (assuming a good random sample) and a well designed study should allow for the use of statistical controls to compensate for extraneous variables.

**Types of quasi-experimental design**

**Cross-sectional study:** Data obtained at one point in time (most surveys).

**Case study:** An in-depth analysis of one entity, object, or event.

**Panel study:** Repeated cross-sectional studies over time with the same participants (cohort study).

**Trend study:** Tracking indicator variables over a period of time (unemployment, crime, dropout rates).
1.2 Reporting Results

The following provides an outline for presenting the results of systematic research. Regardless of the specific format used, the key points to consider in reporting the results of research are:

1. Clearly state the research question up front.
2. Completely explain study assumptions and method of inquiry so that others may duplicate the study.
3. Objectively and accurately report the results of the analysis.
4. Present data in tables that are
   4.1. Accurate
   4.2. Complete, and
   4.3. Titled and documented so that they could stand on their own without a report.
5. Correctly reference sources of information and related research.
6. Openly discuss weaknesses/biases of the research.
7. Develop a defensible conclusion based on the analysis, not personal opinion.

Paper Format

The following outline may be a useful guide in formatting a research report. It incorporates elements of the research design and steps for hypothesis testing (in italics). It may be helpful to refer back to this outline after reviewing the sections on hypothesis testing. The report should clearly identify each section so the reader does not get lost or confused.
I. **Introduction:** The introduction section should include a definition of the central research question (purpose of the paper), why it is of interest, and a review of the literature related to the subject and how it relates to the hypotheses.

*Elements:*

- Purpose statement
- Theory guiding study design
- Identify abstract concepts

II. **Method:** Describe the source of the data, sample characteristics, statistical technique(s) applied, level of significance necessary to reject the null hypotheses, and how the abstract concepts were operationalized.

*Elements:*

- Independent variable(s) and level of measurement
- Dependent variable(s) and level of measurement
- **Assumptions**
  1. Random sampling
  2. Independent subgroups
  3. Population normally distributed
- **Hypotheses**
  1. Identify statistical technique(s)
  2. State null hypothesis or alternative hypothesis
- **Rejection criteria**
1. Indicate alpha (amount of error the researcher is willing to except)

2. Specify one or two-tailed tests

III. **Results:** Describe the results of the data analysis and the implications for the hypotheses. The results section should include such elements as univariate, bivariate, and multivariate analyses; significance test statistics, the probability of error and related status of the hypotheses tests (were they rejected?).

   * **Elements:**
     - Describe sample statistics
     - Compute test statistics
     - Decide results

IV. **Conclusion:** Summarize and evaluate the results. Put in plain words what the research found concerning the central research question. Identify alternative variables, discuss implications for further study, and identify the weaknesses of the research and findings.

   * **Elements:**
     - Interpretation (What do the results mean?)
     - Weaknesses (What do the results not mean?)

V. **References:** Only include literature cited in the report.

VI. **Appendix:** Additional tables and information not included in the body of the report.
Table Format

In the professional world, presentation is almost everything. Most consumers of research do not want to review or wade through the output produced by statistical software. As a result, most researchers develop the ability to create one-page summaries of the research findings that contain one or more tables presenting summary data. An example of survey responses to three questions by customer sex and age is given below:

Survey of Customers

<table>
<thead>
<tr>
<th>Customer Characteristics</th>
<th>Sex</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total (1)</td>
<td>Female</td>
</tr>
<tr>
<td>Sample Size -&gt;</td>
<td>3686</td>
<td>1466</td>
</tr>
<tr>
<td>% of Total (1) -&gt;</td>
<td>100%</td>
<td>41%</td>
</tr>
<tr>
<td>Margin of Error -&gt;</td>
<td>1.6%</td>
<td>2.6%</td>
</tr>
</tbody>
</table>

Staff Professional?
- Yes: 89.4%  89.6%  89.5%  89.5%  89.4%  89.5%  89.4%  89.6%  90.0%
- No: 10.6%  10.4%  10.5%  10.5%  10.6%  10.5%  10.4%  10.0%

Treated Fairly?
- Yes: 83.1%  82.8%  83.8%  82.6%  83.0%  83.5%  84.3%  87.9%
- No: 16.9%  17.2%  16.2%  17.4%  17.0%  16.5%  15.7%  12.1%

Served Quickly?
- Yes: 71.7%  69.3%  74.1%  73.3%  69.3%  72.1%  74.2%  78.0%
- No: 28.3%  30.7%  25.9%  26.7%  30.7%  27.9%  25.8%  22.0%

(1) Total number of cases is based on responses to the question concerning being served quickly. Non-responses to survey items cause the sample sizes to vary.
Critical Review Checklist

The following checklist may help when critically evaluating research prepared by others. Is there anything else that should be added to the list?

1. Research question(s) clearly identified
2. Variables clearly identified
3. Operationalization of variables is valid and reliable
4. Hypotheses evident
5. Statistical techniques identified and appropriate
6. Significance level (alpha) stated a priori
7. Assumptions for statistical tests are met
8. Tables are clearly labeled and understandable
9. Text accurately describes the data in the tables
10. Statistical significance properly interpreted
11. Conclusions fit analytical results and original research question(s)
12. Inclusion of all relevant variables in the study design
13. Study design weaknesses are identified and addressed by researcher
14. Other?
15. __________________________________________
16. __________________________________________
1.3 AcaStat Statistical Software

AcaStat
Data analysis simplified.

Most of the examples presented in the following chapters were created with AcaStat statistical software. AcaStat is an inexpensive and easy-to-use data analysis tool. Easily create data files, drag and drop or paste data from spreadsheet tables, or import delimited text files. Run crosstabulations, descriptive statistics, correlation, and common significance tests without learning programming code or complex interface commands.

$19.99 - Visit www.acastat.com or the Mac App Store for more information.
Compatible with Windows Vista or later — Mac OS X 10.7 or later.
Features

- Intuitive data analysis interface
- Drag and drop variable selection
- Running record of analyses
- Data file capacity ~200,000 observations
- 19 statistical procedures
- 18 summary statistic procedures
- Format, compute and recode variables
- Create mean centered variables
- Trim lengthy string variables
- Split data file module for reducing number of variables
- Create, save, export delimited or AcaStat data files
- Paste or drag and drop spreadsheet data
- Import delimited text files
- Random sampling module
- Weight procedure (frequencies and descriptives)
- Charts module
- Filter (exclude) observations
- Merge data files by records or variables
- Print and save output
- Convert output for pasting into spreadsheets
- Drag and drop data and output files into AcaStat
- Example data files with variable formatting
- Statistics glossary module
- Quick Start and User Manuals
- On-line How To Guide
**Statistical Procedures**

- **Frequencies tables** - chi-square goodness of fit
- **List observations**
- **Descriptives** - count, sum, min, max, mean, median, range, population/sample variance and deviation, skewness, standard error, coefficient of variation, 95% confidence interval, one-sample t-test
- **Explore descriptive statistics** by subgroups
- **Crosstabulation** - cell count, row %, col %, total %, chi-square, Yates's correction, Fisher's, odds ratio, relative risk, Cramer's V, Pearson C, lambda, kappa, McNemar’s test
- **Independent and Paired T-test of Means** - homogeneity of variance
- **Wilcoxon Signed-Rank and Mann-Whitney U** – Nonparametric test
- **One-way ANOVA** - Bonferroni post hoc, eta, eta squared, Levene
- **Kruskal-Wallis** – Nonparametric test
- **Bivariate Correlation** – Pearson & Spearman rho and matrices
- **Point-Biserial Correlation** – summary statistics, Brown-Forsythe test
- **Multiple regression (OLS & Logistic)** - summary statistics, output residuals and predictive values
- **Diagnostic accuracy** - area under the curve, sensitivity, specificity, positive and negative predictive values, likelihood ratios, kappa
- **Appraisal Analysis** – ratios, price-related differential (PRD), coefficient of dispersion (COD), coefficient of variation (CV)
- **Repeated Random Samples** – simulate a data file of means
- **Summary Statistics Module** – weighted mean, z-scores, confidence intervals, chi-square, one-way anova, z-test of proportions, t-test of means, diagnostic accuracy, Epi Curve, probability distributions
Data File Basics

There are two general sources of data; primary and secondary data. Primary data is designed and collected specifically to answer one or more research questions. Secondary data was collected by others for purposes that may or may not match the research questions being explored. An example of a primary data source would be an employee survey the researcher designs and implements for an organization to evaluate job satisfaction. An example of a secondary data source would be the use of census data or other publicly available data such as the General Social Survey to explore research questions that may not have been specifically envisioned when the data design was created.

Designing data files

The best way to envision a data file is to use the analogy of the common spreadsheet software. In spreadsheets, there are columns and rows. For many data files, a spreadsheet provides an easy means of organizing and entering data. In a rectangular data file, columns represent variables and rows represent observations. Variables are commonly formatted as either numerical or string. A numerical variable is used whenever the research plans to manipulate the data mathematically. Examples would be age, income, temperature, and job satisfaction rating. A string variable is used whenever the research plans to treat the data entries like words. Examples would be names, cities, case identifiers, and race. Most variables that could be considered string are coded as a numeric variable. As an example, data for the variable "sex" might be coded 1 for male and 2 for female instead of using a string variable that would require letters (e.g., "Male" and "Female"). This has two benefits. First, numeri-
cal entries are easier and quicker to enter. Second, manipulation of numerical data with statistical software is generally much easier than manipulating string variables.

Data file format

There are several common data file formats. As a general rule, there are data files that are considered system files and data files that are text files. System files are created by and for specific software applications. Examples would be Microsoft Excel, SAS, STATA, and SPSS. Text files contain data in ASCII format and are almost universal in that they can be imported into most statistical programs.

Text files

Text files can be either fixed or free formatted.

Fixed: In fixed formatted data, each variable will have a specific column location. When importing fixed formatted data into a statistical package, these column locations must be identified for the software program. The following is an example:

```
++++|++++|++++|++++|++++|
10123HoustonTX12Female1
```

Reading from left to right, the variables and their location in the data file are:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Column location</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>1-3</td>
<td>101</td>
</tr>
<tr>
<td>Age</td>
<td>4-5</td>
<td>23</td>
</tr>
<tr>
<td>City</td>
<td>6-12</td>
<td>Houston</td>
</tr>
<tr>
<td>State</td>
<td>13-14</td>
<td>TX</td>
</tr>
<tr>
<td>Education</td>
<td>15-16</td>
<td>12</td>
</tr>
<tr>
<td>Sex</td>
<td>17-22</td>
<td>Female</td>
</tr>
<tr>
<td>Marital status</td>
<td>23</td>
<td>1</td>
</tr>
</tbody>
</table>

1=single 2=married
**Free**: In free formatted data, either a space or special value separates (delimits) each variable. Common delimiters are tabs or commas. When importing free formatted data into a statistical package, the software assumes that when a delimiter value is found that it is the end of the previous variable and the next character will begin another variable.

The following is an example of a comma separated value data file (know as a csv file):

101,23,Houston,TX,12,Female,1,

Reading from left to right, the variables are:

Case, Age, City, State, Education, Sex, Marital status

When reading either fixed or free data, statistical software counts the number of variables and assumes when it reaches the last variable that the next variable will be the beginning of another observation (case).

**Data dictionary**

A data dictionary defines the variables contained in a data file (and sometimes the format of the data file). To properly define and document a data file, the researcher should record information on the coding of each variable. The following is an example of a description of variable coding for a three-question survey.

| Employee Survey | Response #:
|-----------------|-----------------------|
| Q1: How satisfied are you with the current pay system? | a. Very satisfied  
| | b. Somewhat satisfied  
| | c. Satisfied  
| | d. Somewhat dissatisfied  
| | e. Very dissatisfied  
| Q2: How many years have you been employed here? |  
| Q3: Please fill in your department name: |  

To properly define and document a data file, record the following information:

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable name</td>
<td>An abbreviated name used by statistical software to represent the variable (generally 8 characters or less)</td>
</tr>
<tr>
<td>Variable type</td>
<td>String, numerical, date</td>
</tr>
<tr>
<td>Variable location</td>
<td>If a fixed data set, the column location and possibly row in data file</td>
</tr>
<tr>
<td>Variable label</td>
<td>Longer description of the variable</td>
</tr>
<tr>
<td>Value label</td>
<td>If a numerical variable is coded to represent categories, the categories represented by the values must be identified</td>
</tr>
</tbody>
</table>

For the employee survey, the data dictionary for a comma separated data file will look like the following:

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable name</td>
<td>CASEID</td>
</tr>
<tr>
<td>Variable type</td>
<td>String</td>
</tr>
<tr>
<td>Variable location</td>
<td>First variable in csv file</td>
</tr>
<tr>
<td>Variable label</td>
<td>Response tracking number</td>
</tr>
<tr>
<td>Value labels</td>
<td>(not normally used for string variables)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable name</td>
<td>Q1</td>
</tr>
<tr>
<td>Variable type</td>
<td>Numerical (categorical)</td>
</tr>
<tr>
<td>Variable location</td>
<td>Second variable</td>
</tr>
<tr>
<td>Variable label</td>
<td>Satisfaction with pay system</td>
</tr>
<tr>
<td>Value labels</td>
<td>1=Very satisfied</td>
</tr>
<tr>
<td></td>
<td>2=Somewhat satisfied</td>
</tr>
<tr>
<td></td>
<td>3=Satisfied</td>
</tr>
<tr>
<td></td>
<td>4=Somewhat dissatisfied</td>
</tr>
<tr>
<td></td>
<td>5=Very dissatisfied</td>
</tr>
</tbody>
</table>
Variable name: Q2  
Variable type: Numerical  
Variable location: Third variable  
Variable label: Years employed  
Value labels: None

Variable name: Q3  
Variable type: String  
Variable location: Fourth variable  
Variable label: Department name  
Value labels: None

If the data for four completed surveys are entered into a spreadsheet, it will look like the following:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CASEID</td>
<td>Q1</td>
<td>Q2</td>
</tr>
<tr>
<td>2</td>
<td>1001</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>1002</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>1003</td>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>1004</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

The data will look like the following if saved in a text file as comma separated (note: the total number of commas for each record equals the total number of variables):

CASEID, Q1, Q2, Q3,
1001, 2, 5, Admin,
1002, 5, 10, MIS,
1003, 1, 23, Accounting,
1004, 4, 3, Legal,
Some statistical software will not read the first row in the data file as variable names. In that case, the first row must be deleted before saving to avoid computation errors. The data file without variable names will look like the following:

1001, 2, 5, Admin,
1002, 5, 10, MIS,
1003, 1, 23, Accounting,
1004, 4, 3, Legal,

This is a very efficient way to store and analyze large amounts of data, but it should be apparent at this point that a data dictionary would be necessary to understand what the aggregated data represent. Documenting data files is very important. Although this was a simple example, many research data files have hundreds of variables and thousands of observations that will be uninterpretable without careful documentation.
Descriptive statistics are used to describe a population or sample. Population descriptive statistics are calculated using all possible observations in a population. Sample descriptive statistics are calculated using a subset of observations in a population. In the case of samples, the descriptive statistic only represents the sample, not the underlying population they originated from.

When a descriptive statistic is presented to represent one variable, it is commonly referred to as a univariate statistic. Examples would be mean income of residents in a city, percent of city residents with a high school education, etc. When the relationship between two variables is presented, this is commonly referred to as a bivariate statistic. Separately calculating mean income and percent with a high school education for males and females would be an example of bivariate statistics.

3.1 Univariate Statistics – Counts and Percents

This section presents techniques for using counts and proportions to create univariate descriptive statistics. Use the techniques presented in this section if the analysis variable meets the following characteristics.
Univariate descriptive statistics are used to describe one variable through the creation of a summary statistic. Common univariate statistics for nominal/ordinal data include ratios, rates, proportions, and percentages.

Ratios

A ratio measures the extent to which one category of a variable outnumbers another category in the same variable.

**Ratio Example:** The community has 1370 Protestants and 930 Catholics.

\[
Ratio = \frac{\text{LargerFreq}}{\text{SmallerFreq}}
\]

or

\[
Ratio = \frac{1370}{930}
\]

For every one Catholic there are 1.47 Protestants in the given population.
Rates

A rate measures the number of actual occurrences out of all possible occurrences per an established period of time.

Rate Example: Create the death rate per 1,000 given there are 100 deaths a year in a population of 10,000.

\[
Rate = \frac{\text{# Per Year}}{\text{Total Population}} \times \text{Rate Metric} \\
\text{or}
\]

\[
Rate = \frac{100}{10,000} \times 1,000
\]

The population death rate is 10 deaths per 1000 people.

Proportions

Proportions are a common method used to describe frequencies as they compare to a total (relative frequencies). A proportion weights the frequency of occurrence against the total possible. It is often reported as a percentage.

\[
Proportion = \frac{\text{OccurrenceFreq}}{\text{TotalPossible}}
\]

\[
Percent = \frac{\text{OccurrenceFreq}}{\text{TotalPossible}} \times 100
\]
Software Output: An example using survey data.

Frequencies

<table>
<thead>
<tr>
<th>Value</th>
<th>Count</th>
<th>Percent</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>08</td>
<td>1</td>
<td>2.00</td>
<td>2.00</td>
</tr>
<tr>
<td>09</td>
<td>2</td>
<td>4.00</td>
<td>6.00</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>4.00</td>
<td>10.00</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>4.00</td>
<td>14.00</td>
</tr>
<tr>
<td>12</td>
<td>21</td>
<td>42.00</td>
<td>56.00</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>4.00</td>
<td>60.00</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>8.00</td>
<td>68.00</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>6.00</td>
<td>74.00</td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>12.00</td>
<td>86.00</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>6.00</td>
<td>92.00</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>2.00</td>
<td>94.00</td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>4.00</td>
<td>98.00</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>2.00</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Interpretation

[a] 2 out of 50 have 9 years of education

[b] 4.00% have 9 years of education (2/50 = 0.04; 0.04 * 100 = 4.00%)

[c] 6.00% have 9 years or less education (3/50 = 0.06; 0.06*100 = 6.00%)
3.2 Univariate Statistics – Mean, Median, and Mode

This section presents techniques for using interval/ratio data to create descriptive statistics. For the methods reviewed in this section, the analysis variable must be interval or ratio level data.

<table>
<thead>
<tr>
<th>Interval Data</th>
<th>Ratio Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifies objects by type or kind but also has some logical order in consistent intervals.</td>
<td>Classifies objects by type or kind but also has some logical order in consistent intervals.</td>
</tr>
<tr>
<td>★ Categories are mutually exclusive</td>
<td>★ Same as Interval but also has a zero starting point</td>
</tr>
<tr>
<td>★ Logical order exists</td>
<td></td>
</tr>
<tr>
<td>★ Scaled according to amount of a particular characteristic they possess</td>
<td></td>
</tr>
<tr>
<td>★ Differences between each level are equal</td>
<td></td>
</tr>
<tr>
<td>★ No zero starting point</td>
<td></td>
</tr>
</tbody>
</table>

Examples of Interval/Ratio

<table>
<thead>
<tr>
<th>Miles to work</th>
<th>Income</th>
<th>IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAT score</td>
<td>Education in years</td>
<td>Temperature</td>
</tr>
<tr>
<td>Age</td>
<td>Weight</td>
<td>Height</td>
</tr>
</tbody>
</table>

The following reviews common univariate descriptive statistics used to represent central tendency and univariate statistics that describe variation in the data.

Measures of Central Tendency

**Mode**: The mode is the most frequently occurring score. A distribution of scores can be unimodal (one score occurred most frequently), bimodal (two scores tied for most frequently occurring), or multimodal. In the table that follows the mode is 32. If there were also two scores with the value of 60, the distribution would be bimodal (32 and 60).
**Median:** The median is the point on a rank ordered list of scores below which 50% of the scores fall. It is especially useful as a measure of central tendency when there are very extreme scores in the distribution, such as would be the case if we had someone in the age distribution provided below who was 120. If the number of scores is odd, the median is the score located in the position represented by \((n+1)/2\). In the table below the median is located in the 4\(^{th}\) position \((7+1)/2\) and would be reported as a median of 42. If the number of scores are even, the median is the average of the two middle scores.

As an example, if the last score (65) is dropped from the following table, the median would be represented by the average of the 3\(^{rd}\) (6/2) and 4\(^{th}\) score, or \((32+42)/2=37\). Always remember to order the scores from low to high before determining the median.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Age</th>
<th>Also known as X</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Mode</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>Median</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Number of scores or cases (n)</td>
<td></td>
</tr>
<tr>
<td>310</td>
<td>Sum of scores</td>
<td></td>
</tr>
<tr>
<td>44.29</td>
<td>Mean</td>
<td></td>
</tr>
</tbody>
</table>
Mean:

\[ \bar{X} = \frac{\sum X_i}{n} \]

The mean is the sum of the scores (\( \sum X_i \)) divided by the number of scores (n) to compute an arithmetic average of the scores in the distribution. The mean is the most often used measure of central tendency. It has two properties: 1) the sum of the deviations of the individual scores (Xi) from the mean is zero, 2) the sum of squared deviations from the mean is smaller than what can be obtained from any other value created to represent the central tendency of the distribution. In the above table the mean age is \((310/7)\) or 44.29.

Weighted Mean:

\[ \bar{X}_w = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2 + n_3 \bar{X}_3}{n_1 + n_2 + n_3} \]

When two or more means are combined to develop an aggregate or grand mean, the influence of each mean must be weighted by the number of cases in its subgroup.

Example

\[ \bar{X}_1 = 12, n = 10 \]

\[ \bar{X}_2 = 14, n = 15 \]

\[ \bar{X}_3 = 18, n = 40 \]

Wrong Method: \[ \frac{(12 + 14 + 18)}{3} = 14.7 \]
Correct Method:
\[
\frac{10(12) + 15(14) + 40(18)}{10 + 15 + 40} = 16.2
\]

Measures of Variation

Range: The range is the difference between the highest and lowest score (high-low). It describes the span of scores but cannot be compared to distributions with a different number of observations. In the table below, the range is (65-24) or 41.

Variance: The variance represents the average of the squared deviations between the individual scores and the mean. The larger the variance the more variability there is among the scores in a given distribution. When comparing two samples with the same unit of measurement (e.g., age), the variances are comparable even though the sample sizes may be different. Generally, smaller samples have greater variability among the scores than larger samples. The formula is almost the same for estimating population variance.

Population Variance:
\[
\sigma^2 = \frac{\Sigma (X_i - \mu)^2}{N}
\]

Sample Variance:
\[
S^2 = \frac{\Sigma (X_i - \bar{X})^2}{n - 1}
\]
Standard deviation: The standard deviation represents the square root of variance. It provides a representation of the variation among scores that is directly comparable to the raw scores.

\[ \sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} \]

Population Standard Deviation:

\[ S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n - 1}} \]

Sample Standard Deviation:

Example

<table>
<thead>
<tr>
<th>Variable (Xi)</th>
<th>Age</th>
<th>(Xi-mean)</th>
<th>(Xi-mean)* (Xi-mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>-20.29</td>
<td></td>
<td>411.68</td>
</tr>
<tr>
<td>32</td>
<td>-12.29</td>
<td></td>
<td>151.04</td>
</tr>
<tr>
<td>32</td>
<td>-12.29</td>
<td></td>
<td>151.04</td>
</tr>
<tr>
<td>42</td>
<td>-2.29</td>
<td></td>
<td>5.24</td>
</tr>
<tr>
<td>55</td>
<td>10.71</td>
<td></td>
<td>114.70</td>
</tr>
<tr>
<td>60</td>
<td>15.71</td>
<td></td>
<td>246.80</td>
</tr>
<tr>
<td>65</td>
<td>20.71</td>
<td></td>
<td>428.90</td>
</tr>
</tbody>
</table>

\[ n = 7 \]

\[ \text{mean} = 44.29 \]

\[ \text{sum of squares} = 1509.43 \]

\[ \text{sample variance} = 251.57 \]

\[ \text{sample standard deviation} = 15.86 \]
**Software Example**: Uses data from previous example.

### Descriptive Statistics

**Variable**: Age

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>7</td>
</tr>
<tr>
<td>Sum</td>
<td>310.0000</td>
</tr>
<tr>
<td>Mean</td>
<td>44.2857</td>
</tr>
<tr>
<td>Median</td>
<td>42.0000</td>
</tr>
<tr>
<td>Min</td>
<td>24.0000</td>
</tr>
<tr>
<td>Max</td>
<td>65.0000</td>
</tr>
<tr>
<td>Range</td>
<td>41.0000</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.0871</td>
</tr>
</tbody>
</table>

**Pop Var** | 215.6327 [b] |
**Sam Var** | 251.5714 |
**Pop Std** | 14.6844 |
**Sam Std** | 15.8610 |
**Std Error** | 5.9949 |
**CV%** | 35.8152 [c] |
**95% CI (+/-)** | 14.6690 [d] |
**One sample t-test (mu=0)|p<|** | 0.0003 [e] |

**Missing Cases** | 0

### Interpretation

[a] Skewness provides an indication of the how asymmetric the distribution is for a given sample. A negative value indicates a negative skew. Values greater than 1 or less than -1 indicate a non-normal distribution.

[b] Population variance (Pop Var) and population standard deviation (Pop Std) will always be less than sample variance and sample standard deviation, since the sum of squares is divided by n instead of n-1.

[c] Coefficient of Variation (CV) is the ratio of the sample standard deviation to the sample mean: (sample standard deviation/sample mean)*100 to calculate CV%. It is used as a measure of relative variability, CV is not affected by the units of a variable.

[d] Add and subtract this value to the mean to create a 95% confidence interval.

[e] This represents the results of a one-sample t-test that compares the sample mean to a hypothesized population value of 0 years of age. In this example, the sample mean age of 44 is statistically significantly different from zero.
3.3 Standardized Z-Scores

A standardized z-score represents the relative position of an individual score in a distribution as compared to the mean and the variation of scores in the distribution. A negative z-score indicates the score is below the distribution mean. A positive z-score indicates the score is above the distribution mean. Z-scores will form a distribution identical to the distribution of raw scores; the mean of z-scores will equal zero and the variance and standard deviation of a z-distribution will always equal one.

To obtain a standardized score, subtract the mean from the individual score and divide by the standard deviation.

\[ Z = \frac{X_i - \bar{X}}{S} \]

**Example:** mean=44.29 and standard deviation=15.86

<table>
<thead>
<tr>
<th>Variable (Xi)</th>
<th>Age</th>
<th>(Xi-mean)</th>
<th>(Xi-mean)/s</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doug</td>
<td>24</td>
<td>-20.29</td>
<td>-20.29/15.86</td>
<td>-1.28</td>
</tr>
<tr>
<td>Mary</td>
<td>32</td>
<td>-12.29</td>
<td>-12.29/15.86</td>
<td>-0.77</td>
</tr>
<tr>
<td>Jenny</td>
<td>32</td>
<td>-12.29</td>
<td>-12.29/15.86</td>
<td>-0.77</td>
</tr>
<tr>
<td>Frank</td>
<td>42</td>
<td>-2.29</td>
<td>-2.29/15.86</td>
<td>-0.14</td>
</tr>
<tr>
<td>John</td>
<td>55</td>
<td>10.71</td>
<td>10.71/15.86</td>
<td>0.68</td>
</tr>
<tr>
<td>Beth</td>
<td>60</td>
<td>15.71</td>
<td>15.71/15.86</td>
<td>0.99</td>
</tr>
<tr>
<td>Ed</td>
<td>65</td>
<td>20.71</td>
<td>20.71/15.86</td>
<td><strong>1.31</strong> [a]</td>
</tr>
</tbody>
</table>

**Interpretation**

[a] Ed is 1.31 standard deviations above the mean age for those represented in the sample.
Software Output: Data from the previous example.

The following displays z-scores created by statistical software when conducting a descriptive analysis. The statistical software calculates the mean and standard deviation for the sample data and then calculates a z-score for each observation and outputs the result to the data file thereby creating a new variable (in this case Z-Age).

<table>
<thead>
<tr>
<th>Obs</th>
<th>Age</th>
<th>Z-Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24</td>
<td>-1.27897</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
<td>-0.77459</td>
</tr>
<tr>
<td>3</td>
<td>32</td>
<td>-0.77459</td>
</tr>
<tr>
<td>4</td>
<td>42</td>
<td>-0.14411</td>
</tr>
<tr>
<td>5</td>
<td>55</td>
<td>0.67551</td>
</tr>
<tr>
<td>6</td>
<td>60</td>
<td>0.99075</td>
</tr>
<tr>
<td>7</td>
<td>65</td>
<td>1.30599</td>
</tr>
</tbody>
</table>

The following displays the descriptive statistics for the Z-Age variable. Notice that the mean is zero and sample variance and sample standard deviation are one.

Descriptive Statistics
Variable: Z-Age

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>7</td>
<td>Pop Var</td>
<td>0.8571</td>
</tr>
<tr>
<td>Sum</td>
<td>0.0000</td>
<td>Sam Var</td>
<td>1.0000</td>
</tr>
<tr>
<td>Mean</td>
<td>0.0000</td>
<td>Pop Std</td>
<td>0.9258</td>
</tr>
<tr>
<td>Median</td>
<td>-0.1441</td>
<td>Sam Std</td>
<td>1.0000</td>
</tr>
<tr>
<td>Min</td>
<td>-1.2790</td>
<td>Std Error</td>
<td>0.3780</td>
</tr>
<tr>
<td>Max</td>
<td>1.3060</td>
<td>CV%</td>
<td>ERR</td>
</tr>
<tr>
<td>Range</td>
<td>2.5850</td>
<td>95% CI (+/-)</td>
<td>0.9248</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.0871</td>
<td>One sample t-test (mu=0)</td>
<td>p&lt;</td>
</tr>
</tbody>
</table>

Missing Cases 0
Z-scores may also be used to compare similar individuals in different groups. As an example, to compare students with similar abilities taking two different classes with the same level of course content but different instructors, the exam scores from the two courses could be “normalized.” To more properly compare student A’s performance in class A with Student B’s performance in class B, the numerical test scores could be adjusted by the variation (standard deviation) of test scores in each class and the distance of each student’s test score from the average (mean) for the class. The student with the greater z-score performed relatively better, as measured by the test score, than the student with the lower z-score.

**Transformed Z-Scores**

Z-scores have two major disadvantages for a lay audience. First, negative scores have negative connotations. Second, z-scores and their associated decimal point presentation can be difficult for others to interpret. These disadvantages can be mitigated by transforming the z-scores into another metric that removes decimals and negative values. This is a common practice in standardized testing.

**Procedure:**

1. Transform raw scores in z-scores

\[ z_i = \frac{x_i - \overline{X}}{s} \]

2. Arbitrarily set a mean value that removes negative numbers

3. Arbitrarily set a standard deviation value

4. Calculate transformed z-score using the following formula:

\[ z' = \overline{X}' + z_i \left( s' \right) \]
where
\[ \bar{X}' = \text{Arbitrary mean} \]
\[ S' = \text{Arbitrary standard deviation} \]

**Example**: transformed z-score

Fred took an IQ test and correctly answered 70 out of 100 points. The mean score on this test is 75 with a standard deviation 5. The following transforms Fred’s z-score (-1) to a scale where the mean is 100 and a standard deviation is 15.

Sample statistics: \[ x_i = 70 \quad \bar{X} = 75 \quad s = 5 \]

Fred’s z-score: \[ z_i = -1 \]

Transformation: \[ \bar{X}' = 100 \quad \text{and} \quad s' = 15 \]

\[ z' = \bar{X}' + z_i (s') \quad z' = 100 + (-1)(15) \quad z' = 85 \]

Fred’s transformed IQ test score is 85.
3.4 Bivariate Statistics – Counts and Percents

Bivariate descriptive statistics simultaneously analyze (compare) two variables to explore associations between the variables. When using a contingency table to explore counts, the independent variable (if causation can be meaningfully defined) is represented by the columns and the dependent variable is represented by the rows. This is a common practice that makes interpretation easier. The level of measurement can be nominal or ordinal.

Interpreting Contingency Tables

<table>
<thead>
<tr>
<th>(count)</th>
<th>Female</th>
<th>Male</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support tax</td>
<td>75</td>
<td>50</td>
<td>125</td>
</tr>
<tr>
<td>Do not support tax</td>
<td>25</td>
<td>50</td>
<td>75</td>
</tr>
<tr>
<td>Column Total</td>
<td>100</td>
<td>100</td>
<td>200</td>
</tr>
</tbody>
</table>

Using counts to create more meaningful percentage summary measures.

Column %: Of those who are female, 75% (75/100) support the tax.

Of those who are male, 50% (50/100) support the tax.

Row %: Of those supporting the tax, 60% (75/125) are female.

Of those not supporting the tax, 67% (50/75) are male.

Row Total %: Overall 62.5% (125/200) support the tax.

Overall 37.5% (75/200) do not support the tax

Column Total %: Overall 50% (100/200) are female and 50% (100/200) are male.
Software Output: An example using survey data.

Crosstabulation: EduCat (Rows) by Sex (Columns)

Column Variable Label: Respondent sex
Row Variable Label: Education Level

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>3</td>
<td>4 [a]</td>
<td>7</td>
</tr>
<tr>
<td>Row %</td>
<td>42.86</td>
<td>57.14 [b]</td>
<td></td>
</tr>
<tr>
<td>Col %</td>
<td>11.54</td>
<td>16.67 [c]</td>
<td>14.00</td>
</tr>
<tr>
<td>Total %</td>
<td>6.00</td>
<td>8.00 [d]</td>
<td></td>
</tr>
<tr>
<td>&lt;12 yrs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS Grad</td>
<td>10</td>
<td>11</td>
<td>21</td>
</tr>
<tr>
<td>College</td>
<td>13</td>
<td>9</td>
<td>22</td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>24</td>
<td>50</td>
</tr>
</tbody>
</table>

Interpretation

[a] Count: There are 4 females who have less than 12 years of education.

[b] Row %: Of those who have less than a high school education (<12 yrs), 57.14% (4/7) are female.

[c] Column %: Of those who are female, 16.67% (4/24) have less than a high school education.
Total %: 8% (4/50) of the sample is females with less than a high school education.

Summary Table

The following summary table is based on column percent from the crosstabulation shown on the previous page. Counts are from the column marginals (Totals) for male and female.

<table>
<thead>
<tr>
<th>Education Level (column %)</th>
<th>Total</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observations (n) =</td>
<td>50</td>
<td>26</td>
<td>24</td>
</tr>
<tr>
<td>Less than high school</td>
<td>14.0</td>
<td>11.5</td>
<td>16.7</td>
</tr>
<tr>
<td>High School</td>
<td>42.0</td>
<td>38.5</td>
<td>45.8</td>
</tr>
<tr>
<td>College</td>
<td>44.0</td>
<td>50.0</td>
<td>37.5</td>
</tr>
</tbody>
</table>

Interpretation

There are slightly more men in this sample than women. Using percentages to adjust for differences in sample size, men are more likely (50%) to have a college education than women (37.5%).
3.5 Bivariate Statistics – Means

Bivariate descriptive statistics simultaneously analyze (compare) two variables to explore associations between the variables. At least one of the variables must be interval/ratio data. For most descriptive statistics, the other variable will likely be nominal or ordinal level data. Bivariate comparisons between two interval/ratio variables are discussed in the correlation chapter.

The following example provides separate income means for males and females.

Software Output: Income by Sex

Explore Descriptive Statistics by Subgroup
Analysis Variable: Respondent Income (approximated)

Category: All

<table>
<thead>
<tr>
<th>Count</th>
<th>1439</th>
<th>Pop Std Dev</th>
<th>19498.0888</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0.0000</td>
<td>Sample Std Dev</td>
<td>19504.8672</td>
</tr>
<tr>
<td>Max</td>
<td>75000.0000</td>
<td>Standard Error</td>
<td>514.1769</td>
</tr>
<tr>
<td>Sum</td>
<td>24060750.0000</td>
<td>95% CI (+/-)</td>
<td>1008.6397</td>
</tr>
<tr>
<td>Median</td>
<td>11250.0000</td>
<td>95% Lower Limit</td>
<td>15711.8259</td>
</tr>
<tr>
<td>Mean</td>
<td>16720.4656</td>
<td>95% Upper Limit</td>
<td>17729.1053</td>
</tr>
<tr>
<td>Range</td>
<td>75000.0000</td>
<td>Coeff of Variation</td>
<td>116.6527</td>
</tr>
</tbody>
</table>
By Subgroup Variable: SEX

Category: Male

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>620</td>
<td>22407.0499</td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>0.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td>75000.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>14445000.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>18750.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>23298.3871</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>75000.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pop Std Dev</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample Std Dev</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard Error</td>
<td></td>
<td>901.3426</td>
<td></td>
</tr>
<tr>
<td>95% CI (+/-)</td>
<td></td>
<td>1770.0598</td>
<td></td>
</tr>
<tr>
<td>95% Lower Limit</td>
<td></td>
<td>21528.3273</td>
<td></td>
</tr>
<tr>
<td>95% Upper Limit</td>
<td></td>
<td>25068.4469</td>
<td></td>
</tr>
<tr>
<td>Coeff of Variation</td>
<td></td>
<td>96.2519</td>
<td></td>
</tr>
</tbody>
</table>

Category: Female

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>819</td>
<td>15177.0254</td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>0.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td>75000.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>9615750.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>4500.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>11740.8425</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>75000.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pop Std Dev</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample Std Dev</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard Error</td>
<td></td>
<td>530.9765</td>
<td></td>
</tr>
<tr>
<td>95% CI (+/-)</td>
<td></td>
<td>1042.2369</td>
<td></td>
</tr>
<tr>
<td>95% Lower Limit</td>
<td></td>
<td>10698.6056</td>
<td></td>
</tr>
<tr>
<td>95% Upper Limit</td>
<td></td>
<td>12783.0794</td>
<td></td>
</tr>
<tr>
<td>Coeff of Variation</td>
<td></td>
<td>129.3459</td>
<td></td>
</tr>
</tbody>
</table>

Summary table created from the above output

Mean Income by sex

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observations (n)</td>
<td>1439</td>
<td>620</td>
<td>819</td>
</tr>
<tr>
<td>Mean Income</td>
<td>$16,721</td>
<td>$23,298</td>
<td>$11,741</td>
</tr>
</tbody>
</table>
The chain of reasoning and systematic steps used in hypothesis testing are the backbone of every statistical test regardless of whether one writes out each step in a classroom exercise or uses statistical software to conduct statistical tests on variables stored in a data file.

### 4.1 Chain of reasoning for inferential statistics

- Sample(s) must be randomly selected.
- The sample estimate is compared to the underlying distribution of the same size sampling distribution.
- The probability that a sample estimate reflects the population parameter is determined from the sampling distribution.

The underlying logic of hypothesis testing is based on rejecting a statement of no difference or no association. This is called the null hypothesis. The null hypothesis is only rejected when we have evidence beyond a reasonable doubt that a true difference or association exists in the population(s) from which we drew our random sample(s).
Reasonable doubt is based on probability sampling distributions. The benchmark for reasonable doubt (defined by “alpha”) is established by the researcher. Alpha .05 is a common benchmark for reasonable doubt. At alpha .05 we know from the sampling distribution that a test statistic at or beyond this benchmark will only occur by random chance five times out of 100 (5% probability). Since a test statistic that results in an alpha of .05 could only occur by random chance 5% of the time, we assume that the test statistic resulted because there are true differences between the population parameters, not because we unwittingly drew a biased random sample.

The four possible outcomes in hypothesis testing

The following table summarizes the possible outcomes of hypothesis testing as they relate to “truth” in the underlying population. It is important to remember that in hypothesis testing we are using sample statistics to make predictions about the unknown population values. The orange cells represent erroneous conclusions from hypothesis testing. One error (Type I) is to reject the null hypothesis when there is no difference in the population (i.e., this is being wrong when we conclude significance). The other error (Type II) is to not reject the null hypothesis when there is an actual difference. The blue cells represent correct decisions from hypothesis testing.

<table>
<thead>
<tr>
<th>Hypothesis Testing Outcomes</th>
<th>In the Population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>There is no difference</td>
</tr>
<tr>
<td>Decision From Sample</td>
<td>Null Hyp. True</td>
</tr>
<tr>
<td>Rejected Null Hyp.</td>
<td>Type I error (alpha)</td>
</tr>
<tr>
<td>Did not Reject Null</td>
<td>Correct Decision</td>
</tr>
</tbody>
</table>
When learning statistics we generally conduct statistical tests by hand. In these situations, we establish before the test is conducted what test statistic is needed (called the critical value) to claim statistical significance. So, if we know for a given sampling distribution that a test statistic of plus or minus 1.96 would only occur 5% of the time randomly, any test statistic that is 1.96 or greater in absolute value would be statistically significant. In an analysis where a test statistic was exactly 1.96, there would be a 5% chance of being wrong if statistical significance is claimed. If the test statistic was 3.00, statistical significance could also be claimed but the probability of being wrong would be much less (about .002 if using a 2-tailed test or two-tenths of one percent; 0.2%). Both .05 and .002 are known as alpha; the probability of a Type I error.

When conducting statistical tests with computer software, the exact probability of a Type I error is calculated. It is presented in several formats but is most commonly reported as "p <" or "Sig." or "Signif." or "Significance." Using "p <" as an example, if a priori the threshold for statistical significance is established at alpha .05, any test statistic with significance at or less than .05 would be considered statistically significant and the null hypothesis of no difference must be rejected. The following table links p values with a constant alpha benchmark of .05 (note: alpha remains constant while p-values are a direct result of the analysis on sample data):

<table>
<thead>
<tr>
<th>P &lt;</th>
<th>Alpha</th>
<th>Probability of Making a Type I Error</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>.05</td>
<td>.05</td>
<td>5% chance difference is not significant</td>
<td>Statistically significant</td>
</tr>
<tr>
<td>.10</td>
<td>.05</td>
<td>10% chance difference is not significant</td>
<td>Not statistically significant</td>
</tr>
<tr>
<td>.01</td>
<td>.05</td>
<td>1% chance difference is not significant</td>
<td>Statistically significant</td>
</tr>
</tbody>
</table>
4.2 The Normal Distribution

Although there are numerous sampling distributions used in hypothesis testing, the normal distribution is the most common example of how data would appear if we created a frequency histogram where the x axis represents the values of scores in a distribution and the y axis represents the frequency of scores for each value. Most scores will be similar and therefore will group near the center of the distribution. Some scores will have unusual values and will be located far from the center or apex of the distribution. In hypothesis testing, we must decide whether the unusual values are simply different because of random sampling error or they are in the extreme tails of the distribution because they are truly different from others. Sampling distributions have been developed that tell us exactly what the probability of this sampling error is when data originate from a random sample collected from a population that is normally distributed.

Properties of a normal distribution

★ Forms a symmetric bell-shaped curve.
★ 50% of the scores lie above and 50% lie below the midpoint of the distribution.
★ The curve is asymptotic to the x axis.
★ The mean, median, and mode are located at the midpoint of the x axis.
Using theoretical sampling probability distributions

Sampling distributions approximate the probability that a particular value would occur by chance alone. If the means were collected from an infinite number of repeated random samples of the same sample size from the same population most means will be very similar in value, in other words, they will group around the true population mean. In a normal distribution, most means will collect about a central value or midpoint of a sampling distribution. The frequency of means will decrease as the value of the random sample mean increases its distance from the center of a normal sampling distribution toward the tails. In a normal probability distribution, about 95% of the means resulting from an infinite number of repeated random samples will fall between 1.96 standard errors above and below the midpoint of the distribution which represents the true population mean and only 5% will fall beyond (2.5% in each tail of the distribution).

The following are commonly used points on a distribution for deciding statistical significance.

- 90% of scores $\pm$ 1.65 standard errors
- 95% of scores $\pm$ 1.96 standard errors
- 99% of scores $\pm$ 2.58 standard errors

**Standard error**: The standard error is a mathematical adjustment to the sample standard deviation to account for the effect sample size has on the underlying sampling distribution. It represents the standard deviation of the sampling distribution.

**Alpha and the role of the distribution tails**

The percentage of scores beyond a particular point along the x axis of a sampling distribution represent the percent of the time during an infinite number of repeated samples one would expect to have a score at or beyond that value on the x axis. This
value on the x axis is known as the critical value when used in hypothesis testing. The midpoint represents the actual population value. Most scores will fall near the actual population value but will exhibit some variation due to sampling error. If a score from a random sample falls 1.96 standard errors or farther above or below the mean of the sampling distribution, we know from the probability distribution that there is only a 5% or less chance of randomly selecting a set of scores that would produce a sample mean that far from the true population mean. This area above and below 1.96 standard errors is the region of rejection.

When conducting significance testing, if we have a test statistic that is at least 1.96 standard errors above or below the mean of the sampling distribution, we assume we have a statistically significant difference between our sample mean and the expected mean for the population. Since we know a value that far from the population mean will only occur randomly 5% or less of the time, we assume the difference is the result of a true difference between the sample and the population mean, and is not the result of random sampling error. The 5% is also known as the probability of being wrong when we conclude statistical significance.

1-tailed vs. 2-tailed statistical tests

A 2-tailed test is used when the researcher cannot determine a priori whether a difference between population parameters will be positive or negative. A 1-tailed test is used when it is reasonable to expect a difference will be positive or negative.
4.3 Steps to Hypothesis Testing

Hypothesis testing is used to establish whether the differences exhibited by random samples can be inferred to the populations from which the samples originated. Regardless of whether statistical tests are conducted by hand or through statistical software, there is an implicit understanding that systematic steps are being followed to determine statistical significance. These general steps are 1) assumptions, 2) null and alternative hypotheses, 3) rejection criteria, 4) computation of test statistics, and 5) decision regarding the null hypothesis.

I. General Assumptions

Population is normally distributed

Random sampling

Mutually exclusive comparison samples

Data characteristics match statistical technique

*For interval / ratio data, use the following*

t-tests, Pearson correlation, ANOVA, OLS regression

*For nominal / ordinal data, use the following*

Difference of proportions, chi square and related measures of association, logistic regression

II. State the Hypothesis

Null Hypothesis (Ho): There is no difference between ___ and ___.

Alternative Hypothesis (Ha): There is a difference between ___ and ___.

Note: The alternative hypothesis will indicate whether a 1-tailed or a 2-tailed test is utilized to reject the null hypothesis.

Ha for 1-tail tested: The __ of __ is greater (or less) than the __ of __.

III. Set the Rejection Criteria

This determines how different the parameters and/or statistics must be before the null hypothesis can be rejected. This "region of rejection" is based on alpha (α) -- the error associated with the confidence level. The point of rejection is known as the critical value.

IV. Compute the Test Statistic

The collected data are converted into standardized scores for comparison with the critical value.

V. Decide Results of Null Hypothesis

If the test statistic equals or exceeds the region of rejection bracketed by the critical value(s), the null hypothesis is rejected. In other words, the chance that the difference exhibited between the sample statistics is due to sampling error is remote--there is an actual difference in the population.
Confidence Intervals

For interval estimation, the data must be from a random sample. The following presents interval estimation techniques for proportions (nominal/ordinal data) and also for means (interval/ratio data).

5.1 Interval Estimation for Proportions

Interval estimation (margin of error) uses sample data to determine a range (interval) that, at an established level of confidence, is expected to contain the population proportion.

Steps

Determine the confidence level (alpha is generally .05).

Use the z-distribution table to find the critical value for a 2-tailed test given the selected confidence level (alpha).

Estimate the standard error of the proportion.

\[ s_p = \sqrt{\frac{pq}{n}} \]

where
\[ p = \text{sample proportion} \]

\[ q = 1 - p \]

Estimate the confidence interval.

\[ CV = \text{critical value} \]

\[ \text{CI} = p \pm (CV)(S_p) \]

Interpret

Based on alpha .05, the researcher is 95% confident that the proportion in the population from which the sample was obtained is between __ and __.

Note: Given the sample data and level of error, the confidence interval provides an estimated range of proportions that is most likely to contain the population proportion. The term "most likely" is measured by alpha (i.e., in most cases there is a 5% chance --alpha .05-- that the confidence interval does not contain the true population proportion).
More About the Standard Error of the Proportion

The standard error of the proportion will vary as sample size and the proportion changes. As the standard error increases, so will the margin of error.

<table>
<thead>
<tr>
<th>Proportion (p)</th>
<th>Sample Size (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td>0.9</td>
<td>0.030</td>
</tr>
<tr>
<td>0.8</td>
<td>0.040</td>
</tr>
<tr>
<td>0.7</td>
<td>0.046</td>
</tr>
<tr>
<td>0.6</td>
<td>0.049</td>
</tr>
<tr>
<td>0.5</td>
<td>0.050</td>
</tr>
<tr>
<td>0.4</td>
<td>0.049</td>
</tr>
<tr>
<td>0.3</td>
<td>0.046</td>
</tr>
<tr>
<td>0.2</td>
<td>0.040</td>
</tr>
<tr>
<td>0.1</td>
<td>0.030</td>
</tr>
</tbody>
</table>

**Effect of changes in the proportion**

As a proportion approaches 0.5 the error will be at its greatest value for a given sample size. Proportions close to 0 or 1 will have the lowest error. The error above a proportion of .5 is a mirror reflection of the error below a proportion of .5.

**Effect of changes in sample size**

As sample size increases the error of the proportion will decrease for a given proportion. However, the reduction in error of the proportion as sample size increases is not constant. Using a proportion of 0.9 as an example, increasing the sample size from 100 to 300 cut the standard error by about half (from .03 to .017). Increasing the sample size by another 200 only reduced the standard error by about one quarter (.017 to .013).
**Example:** Interval Estimation for Proportions

**Problem:** A random sample of 500 employed adults found that 23% had traveled to a foreign country. Based on these data, what is the estimate for the entire employed adult population?

\( n=500, \; p = .23, \; q = .77 \)

Use alpha .05 (i.e., the critical value is 1.96)

**Estimate Sampling Error**

\[
s_p = \sqrt{\frac{pq}{n}}
\]

\[
s_p = \sqrt{\frac{.23(.77)}{500}}
\]

\( s_p = .019 \)

**Compute Interval**

\[
CI = p_s \pm CV(s_p)
\]

\[
CI = .23 \pm 1.96(.019)
\]

\( CI = .23 \pm .037 \)

**Interpret**

Based on a 95% confident level, the actual proportion of all employed adults who have traveled to a foreign country is between 19.3% and 26.7%.
Software output:  Summary data from the previous example.

Margin of Error for Proportions

n = 500   Proportion = 0.23   Standard Error = 0.0188

<table>
<thead>
<tr>
<th>Confidence</th>
<th>Error +/-</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>0.0311</td>
<td>0.1989</td>
<td>0.2611</td>
</tr>
<tr>
<td>95%</td>
<td>0.0369</td>
<td>0.1931</td>
<td>0.2669</td>
</tr>
<tr>
<td>99%</td>
<td>0.0486</td>
<td>0.1814</td>
<td>0.2786</td>
</tr>
</tbody>
</table>

Summary Table of Margin of Errors for each proportion:  (Converted to %)

<table>
<thead>
<tr>
<th>The promotion system is ...</th>
<th>Staff</th>
<th>+/- Margin of Error</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fair</td>
<td>50%</td>
<td>9.8% [b]</td>
<td>40.2% - 59.8%</td>
</tr>
<tr>
<td>Sometimes Fair</td>
<td>30%</td>
<td>8.9%</td>
<td>21.0% - 38.9%</td>
</tr>
<tr>
<td>Not Fair</td>
<td>20% [a]</td>
<td>7.8%</td>
<td>12.2% - 27.8%</td>
</tr>
<tr>
<td>Count (n)</td>
<td>100</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Interpretation

[a]  At the 95% confidence level, the percent of all employees in the population who believe the promotion system is not fair is between 12.2% to 27.8%.

[b]  The margin of error is different for each response category due to the change in proportions. Proportions at 0.5 (50%) will have the highest level of error. Using one margin of error for multiple comparisons
To avoid calculating a separate margin of error for each response category, it is common to calculate the most conservative standard error of a proportion (p=0.5) and use this to represent the margin of error for all response options within a specific subgroup. In the table below, a separate margin of error is calculated for the total sample, the male sample, and the female sample.

**Summary Table:** Using one confidence interval (CI) for all data in one subgroup

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Count</strong></td>
<td>1496</td>
<td>641</td>
<td>855</td>
</tr>
<tr>
<td><strong>95% CI (+/-)</strong></td>
<td>2.5%</td>
<td>3.9%</td>
<td>3.4%</td>
</tr>
</tbody>
</table>

**Respondent's highest degree**

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; High school</td>
<td>18.6%</td>
<td>19.5%</td>
<td>18.0%</td>
</tr>
<tr>
<td>High school</td>
<td>52.1%</td>
<td>47.4%</td>
<td>55.7%</td>
</tr>
<tr>
<td>Junior college</td>
<td>6.0%</td>
<td>5.8%</td>
<td>6.2%</td>
</tr>
<tr>
<td>Bachelor</td>
<td>15.6%</td>
<td>16.8%</td>
<td>14.7%</td>
</tr>
<tr>
<td>Graduate</td>
<td>7.6%</td>
<td>10.5%</td>
<td>5.4%</td>
</tr>
</tbody>
</table>

**Interpretation**

[a] Based on the 95% confidence level, the percent of U.S. adults with a junior college education level is between 3.5% and 8.5% (6.0% +/- 2.5%).

[b] Based on the 95% confidence level, the percent of Female U.S. adults with a junior college education level is between 2.8% and 9.6% (6.2% +/- 3.4%).

[c] Based on the 95% confidence level, the percent of U.S. adults with a graduate education level is between 5.1% and 10.1% (7.6% +/- 2.5%).

[d] Based on the 95% confidence level, the percent of Female U.S. adults with a graduate education level is between 2.0% and 8.8% (5.4% +/- 3.4%).
Interval Estimation for the Difference Between Two Proportions

This approach uses sample data to determine a range (interval) that, at an established level of confidence, will contain the difference between two population proportions.

Steps

Determine the confidence level (generally alpha .05).

Use the z distribution table to find the critical value for a 2-tailed test (at alpha .05 the critical value would equal 1.96).

Estimate Sampling Error

\[
s_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p} \hat{q}}{n_1 n_2}}
\]

where

\[
\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}
\]

and

\[
\hat{q} = 1 - \hat{p}
\]

Estimate the Interval

CI = (\hat{p}_1 - \hat{p}_2) ± (CV)(S_{p1-p2})

\(\hat{p}_1 - \hat{p}_2 = \text{difference between two sample proportions}\)

\(CV = \text{critical value}\)

Interpret

Based on alpha .05, the researcher is 95% confident that the difference between the proportions of the two subgroups in the population from which the sample was obtained is between __ and __.

Note: Given the sample data and level of error, the confidence interval provides an estimated range of proportions that is most likely to contain the difference between the
population subgroups. The term "most likely" is measured by alpha or in most cases there is a 5% chance (alpha .05) that the confidence interval does not contain the true difference between the subgroups in the population.

5.2 Interval Estimation For Means

Interval estimation involves using sample data to determine a range (interval) that, at an established level of confidence, is expected to contain the mean of the population.

Steps

1. Determine confidence level (df=n-1; alpha .05, 2-tailed)

2. Use either the z distribution (if n>120) or the t distribution (for all sizes of n).

3. Use the appropriate table to find the critical value for a 2-tailed test

4. Multiple hypotheses can be compared with the estimated interval for the population to determine their significance. In other words, differing values of population means can be compared with the interval estimation to determine if the hypothesized population means fall within the region of rejection.

Estimation Formula

\[
CI = \overline{X} \pm CV \left( \frac{s}{\overline{x}} \right)
\]

where

\[\overline{X}\] = sample mean

CV = critical value (consult z or t distribution table for df=n-1 and chosen alpha-- commonly .05)
Standard error of the mean

\[ s_\bar{x} = \frac{s}{\sqrt{n}} \]

Note: assumes sample standard deviation was calculated using:

\[ S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n - 1}} \]

**Example:** Interval Estimation for Means

**Problem:** A random sample of 30 incoming college freshmen revealed the following statistics: mean age 19.5 years; sample standard deviation 1.2. Based on a 5% chance of error, estimate the range of possible mean ages for all incoming freshmen.

**Estimation**

Critical value (CV)

Df=n-1 or 29

Consult t-distribution for alpha .05, 2-tailed

CV=2.045

Standard error

\[ s_\bar{x} = \frac{1.2}{\sqrt{30}} \quad s_\bar{x} = .219 \]

Estimate Confidence Interval

\[ CI_{95} = \bar{X} \pm CV(s_\bar{x}) \]
\[ CI_{95} = 19.5 \pm 2.045(.219) \]

\[ CI_{95} = 19.5 \pm .448 \]

**Interpretation**

Based on a 95% confidence level, the actual mean age of the all incoming freshmen will be somewhere between 19 years (the lower limit) and 20 years (the upper limit) of age.

**Software Output:** Summary data from previous example.

**Margin of Error for Means**

\[ n = 30 \quad \text{Mean} = 19.5 \quad \text{Standard Deviation} = 1.2 \]

<table>
<thead>
<tr>
<th>Confidence</th>
<th>Error +/-</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>0.3723</td>
<td>19.1277</td>
<td>19.8723</td>
</tr>
<tr>
<td>95%</td>
<td>0.4481</td>
<td>19.0519</td>
<td>19.9481[a]</td>
</tr>
<tr>
<td>99%</td>
<td>0.6039</td>
<td>18.8961</td>
<td>20.1039</td>
</tr>
</tbody>
</table>

**Interpretation**

[a] Based on a 95% confidence level, the actual mean age of the all incoming freshmen will be somewhere between 19 years (the lower limit) and 20 years (the upper limit) of age. Note that both measures are rounded.
This section presents techniques for using counts and proportions to conduct hypothesis testing. Use the techniques presented in this section if the analysis or dependent variable meet the following characteristics.

<table>
<thead>
<tr>
<th>Nominal Data</th>
<th>Ordinal Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifies objects by type or characteristic</td>
<td>Classifies objects by type or characteristic but also has some logical order</td>
</tr>
<tr>
<td>1. Categories are mutually exclusive</td>
<td>1. Categories are mutually exclusive</td>
</tr>
<tr>
<td>2. No logical order</td>
<td>2. Logical order exists</td>
</tr>
<tr>
<td>3. Scaled according to amount of a particular characteristic they possess</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Examples of Nominal Data</th>
<th>Examples of Ordinal Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex, race, ethnicity</td>
<td>income (low, moderate, high)</td>
</tr>
<tr>
<td>government agency</td>
<td>job satisfaction (5 point scale)</td>
</tr>
<tr>
<td>political party preference</td>
<td>military rank (Lieutenant to General)</td>
</tr>
<tr>
<td>religion</td>
<td>course letter grade (A to F)</td>
</tr>
<tr>
<td>neighborhood (urban/suburb)</td>
<td>importance of welfare</td>
</tr>
<tr>
<td>census region</td>
<td>(little, moderate, high)</td>
</tr>
<tr>
<td>yes/no responses</td>
<td></td>
</tr>
<tr>
<td>have children (yes/no)</td>
<td></td>
</tr>
<tr>
<td>voted (yes/no)</td>
<td></td>
</tr>
</tbody>
</table>
6.1 Z-test of Proportions

For hypothesis testing in this section both the independent and dependent variable must be nominal or ordinal. In addition, the data are assumed to be from a random sample. The relevant sampling distribution for this section is the standard normal (Z) distribution. Tables of critical values are included in the Appendix.

**Statistical Techniques Discussed in This Section**

- One-sample test of proportions
- Two-sample test of proportions

**Comparing a Population Proportion to a Sample Proportion (One-Sample Test)**

A one-sample z-test of proportions is used to compare a proportion created by a random sample to a proportion originating from or thought to represent the value for the entire population. As an example, to make sure the random sample of 100 subjects is not biased regarding a person’s sex, one approach would be to compare the proportion of women in the sample to the known proportion of women in the underlying population as reported in census data or by some other reliable source.

**Example**

**Problem:** Historical data indicates that about 10% of the agency's clients believe they were given poor service. Now under new management for six months, a random sample of 110 clients found that 15% believe they were given poor service.

\[ P_u = .10 \quad P_s = .15 \quad n = 110 \]
I. **Assumptions**

Independent random sampling

Nominal level data

Large sample size

II. **State the Hypothesis**

Ho: There is no statistically significant difference between the historical proportion of clients reporting poor service and the current proportion of clients reporting poor service. $p_u = p_c$

If 2-tailed test

Ha: There is a statistically significant difference between the historical proportion of clients reporting poor service and the current proportion of clients reporting poor service. $p_u \neq p_c$

If 1-tailed test

Ha: The proportion of current clients reporting poor service is significantly greater than the historical proportion of clients reporting poor service.

III. **Set the Rejection Criteria**

Use **z-distribution table** to estimate critical value

If 2-tailed test, $\alpha = .05$, $Z_{cv} = 1.96$

If 1-tailed test, $\alpha = .05$, $Z_{cv} = 1.65$
IV. Compute the Test Statistic

*Estimate Standard Error*

\[ p = \text{population proportion} \]

\[ q = 1 - p \]

\[ n = \text{sample size} \]

\[ s_p = \sqrt{\frac{p \cdot q}{n}} \]

\[ s_p = \sqrt{\frac{.10 \cdot .90}{110}} \]

\[ s_p = .029 \]

*Test Statistic*

\[ Z = \frac{p_s - p_u}{s_p} \]

\[ Z = \frac{.15 - .10}{.029} \]

\[ Z = 1.724 \]

V. Decide Results of Null Hypothesis

If a 2-tailed test was used . . .

Since the test statistic of 1.724 did not meet or exceed the critical value of 1.96, there is insufficient evidence to conclude there is a statistically significant difference between the historical proportion of clients reporting poor service and the current proportion of clients reporting poor service.

If a 1-tailed test was used . . .

Since the test statistic of 1.724 exceeds the critical value of 1.65, conclude the proportion of current clients reporting poor service is significantly greater than the historical proportion of clients reporting poor service.
**Software Output:** Summary data from the previous example.

One-Sample Difference of Proportions

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Proportion</td>
<td>0.10</td>
</tr>
<tr>
<td>Sample Proportion</td>
<td>0.15</td>
</tr>
<tr>
<td>Sample Size (n)</td>
<td>110</td>
</tr>
<tr>
<td>Test statistic</td>
<td>1.748</td>
</tr>
</tbody>
</table>

\[ p < 0.0805 \text{ (2tailed)} \] [a]

**Interpretation**

[a] For a 2-tailed test, the p-value represents the probability of making a type 1 error (concluding there is statistical significance when there is none). Since there is about 8% chance of making a type 1 error, which exceeds the 5% error limit established in the rejection criteria of the hypothesis testing process (alpha), do not conclude there is statistical significance.
Comparing Proportions From Two Independent Samples (Two-Sample Test)

A two-sample z-test of proportions is used to compare two proportions created by two random samples or two subgroups of one random sample.

Example

Problem: A survey was conducted of students from the Princeton public school system to determine if the incidence of hungry children was consistent in two schools located in lower-income areas. A random sample of 80 elementary students from school A found that 23% did not have breakfast before coming to school. A random sample of 180 elementary students from school B found that 7% did not have breakfast before coming to school. Before putting more resources into school A, the school superintendent wants to verify there is a statistically significant difference between the schools (i.e., the difference is beyond just random error created from two small samples of students).

I. Assumptions

- Independent random sampling
- Nominal level data
- Large sample size

II. State the Hypothesis

Ho: There is no statistically significant difference between the proportion of students in school A not eating breakfast and the proportion of students in school B not eating breakfast. $P_1 = P_2$
Ha: There is a statistically significant difference between the proportion of students in school A not eating breakfast and the proportion of students in school B not eating breakfast. \( p_1 \neq p_2 \)

III. **Set the Rejection Criteria**

Use *z-distribution table* to estimate critical value

\[ \text{Alpha}.05, Z_{cv} = 1.96 \]

IV. **Compute the Test Statistic**

**Estimate of Standard Error**

\[
\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}
\]

\[
\hat{q} = 1 - \hat{p}
\]

\[
s_{p_1-p_2} = \sqrt{\hat{p} \hat{q} \left( \frac{n_1}{n_1 + n_2} \right) \left( \frac{n_2}{n_1 + n_2} \right)}
\]

\[
p = \frac{80(.23) + 180(.07)}{80 + 180} = .119
\]

\[
q = .881
\]

\[
s_{p_1-p_2} = .043
\]

**Test Statistic**

\[
Z = \frac{p_1 - p_2}{s_{p_1-p_2}}
\]

\[
Z = \frac{.23 - .07}{.043} = 3.721
\]

V. **Decide Results of the Null Hypothesis**

Since the test statistic 3.721 exceeds the critical value of 1.96, conclude there is a statistically significant difference between the proportion of students in school
A not eating breakfast and the proportion of students in school B not eating breakfast.

**Software Output:** Summary data from the previous example.

Two-Sample Difference of Proportions (2-tailed)

<table>
<thead>
<tr>
<th></th>
<th>School A</th>
<th>School B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Proportion</td>
<td>0.23</td>
<td>0.07</td>
</tr>
<tr>
<td>Sample Size (n)</td>
<td>80</td>
<td>180</td>
</tr>
</tbody>
</table>

Z-statistic       p < Significance
3.674             0.0002 [a]

**Interpretation**

[a] For a 2-tailed test, the p-value represents the probability of making a type 1 error (concluding there is statistical significance when there is none). Since there is far less than a 5% chance of making a type 1 error, conclude there is statistical significance.
Appropriate Sample Size

Sample size is an important issue when using statistical tests that rely on the standard normal distribution (z-distribution). As a rule of thumb, sample size is generally considered large enough to use the z-distribution when \( n(p) > 5 \) and \( p \) is less than \( q \) (1-p). If \( p \) is greater than \( q \) (note: \( q = 1-p \)), then \( n(q) \) must be greater than 5. Otherwise use the binomial distribution.

*Example A*

\[ n=60, \ p=.10 \]

\[ n(p) \quad 60(.10) = 6 \]

Conclude sample size is sufficient

If \( n=40, \ p=.10 \) then \( 40(.10) = 4 \)

Conclude sample size may not be sufficient

*Example B*

\[ n=50, \ p=.80 \]

\[ q=1-p \quad \text{or} \quad q=.20 \]

\[ n(q) \quad 60(.20) = 12 \]

Conclude sample size is sufficient
6.2 Chi-Square

Chi-square is the most common technique used to compare two nominal/ordinal variables, especially when at least one of the variables has more than two categories or the sample size is small. In most cases, chi-square is preferred over a difference of proportions z-test. The relevant sampling distribution for this section is the chi-square distribution.

Statistical Techniques Discussed in This Section

★ Goodness of Fit
★ Chi-Square Test of Independence
★ Measuring Association

Chi-Square Goodness-of-Fit Test

The chi-square goodness of fit test is used to compare frequencies (counts) among multiple categories of nominal or ordinal level data for one-sample (univariate analysis).

Problem: Evaluate variations in the proportion of defects produced from five assembly lines. A random sample of 100 defective parts from the five assembly lines produced the following contingency table.

<table>
<thead>
<tr>
<th>Line A</th>
<th>Line B</th>
<th>Line C</th>
<th>Line D</th>
<th>Line E</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>15</td>
<td>22</td>
<td>20</td>
<td>19</td>
</tr>
</tbody>
</table>

I. Assumptions

Independent random sampling
Nominal or Ordinal level data

II. State the Hypothesis

Ho: There is no significant difference among the assembly lines in the observed frequencies of defective parts.

Ha: There is a significant difference among the assembly lines in the observed frequencies of defective parts.

III. Set the Rejection Criteria

Determine degrees of freedom (df) = k – 1 where k equals the number of categories. df=5-1 or df=4

Establish the confidence level (.05, .01, etc.)

Use the chi-square distribution table to establish the critical value

At alpha .05 and 4 degrees of freedom, the critical value from the chi-square distribution is 9.488

IV. Compute the Test Statistic

\[ X^2 = \sum \left( \frac{(Fo - Fe)^2}{Fe} \right) \]

where Fe = Frequency Expected \[ Fe = \frac{n}{k} \]

and . . .

n = sample size

k = number of categories or cells

Fo = observed frequency
<table>
<thead>
<tr>
<th></th>
<th>Line A</th>
<th>Line B</th>
<th>Line C</th>
<th>Line D</th>
<th>Line E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fo</td>
<td>24</td>
<td>15</td>
<td>22</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>Fe (100/5=20)</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Chi-Square</td>
<td>.8</td>
<td>1.25</td>
<td>.2</td>
<td>0</td>
<td>.05</td>
</tr>
</tbody>
</table>

\[ X^2 = 2.30 \]

V. **Decide Results of Null Hypothesis**

Since the chi-square test statistic 2.30 does not meet or exceed the critical value of 9.488, do not conclude there is a statistically significant difference among the assembly lines in the observed frequencies of defective parts.
Chi-Square Test of Independence

The chi-square test of independence is used to compare frequencies (counts) of nominal or ordinal level data for two samples across two or more subgroups displayed in a crosstabulation table. The chi-square test is more common and more flexible than z-tests of proportions.

**Problem**: Evaluate the association between a person's sex and their attitudes toward school spending on athletic programs. A random sample of adults in a school district produced the following table (counts).

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spend more money</td>
<td>15</td>
<td>25</td>
<td>40</td>
</tr>
<tr>
<td>Spend the same</td>
<td>5</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>Spend less money</td>
<td>35</td>
<td>10</td>
<td>45</td>
</tr>
<tr>
<td><strong>Column Total</strong></td>
<td>55</td>
<td>50</td>
<td>105</td>
</tr>
</tbody>
</table>

I. **Assumptions**

- Independent random sampling
- Nominal/Ordinal level data
- No more than 20% of the cells have an expected frequency less than 5
- No empty cells

II. **State the Hypothesis**

**Ho**: There is no association between a person's sex and their attitudes toward spending on athletic programs.

**Ha**: There is an association between a person's sex and their attitudes toward spending on athletic programs.
III. Set the Rejection Criteria

Determine degrees of freedom $df = (#$ of rows - 1)(#$ of columns - 1)

$df = (3 - 1)(2 - 1)$ or $df = 2$

Establish the confidence level (.05, .01, etc.);  $\text{Alpha} = .05$

Based on the chi-square distribution table, the critical value = 5.991

IV. Compute the Test Statistic

$$X^2 = \sum \left( \frac{(Fo - Fe)^2}{Fe} \right)$$

where

Fo= observed frequency

<table>
<thead>
<tr>
<th>Frequency Observed</th>
<th>Female</th>
<th>Male</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spend more</td>
<td>15</td>
<td>25</td>
<td>40</td>
</tr>
<tr>
<td>Spend same</td>
<td>5</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>Spend less</td>
<td>35</td>
<td>10</td>
<td>45</td>
</tr>
<tr>
<td><strong>Column Total</strong></td>
<td>55</td>
<td>50</td>
<td>105</td>
</tr>
</tbody>
</table>

Fe= expected frequency for each cell

$$Fe = \frac{(\text{frequency for the column}) \times (\text{frequency for the row})}{n}$$

<table>
<thead>
<tr>
<th>Frequency Expected</th>
<th>Female</th>
<th>Male</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spend more</td>
<td>55*40/105 = 20.952</td>
<td>50*40/105 = 19.048</td>
<td>40</td>
</tr>
<tr>
<td>Spend same</td>
<td>55*20/105 = 10.476</td>
<td>50*20/105 = 9.524</td>
<td>20</td>
</tr>
<tr>
<td>Spend less</td>
<td>55*45/105 = 23.571</td>
<td>50*45/105 = 21.429</td>
<td>45</td>
</tr>
<tr>
<td><strong>Column Total</strong></td>
<td>55</td>
<td>50</td>
<td>105</td>
</tr>
</tbody>
</table>
Chi-square Calculations

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spend more</td>
<td>(15-20.952)^2/20.952</td>
<td>(25-19.048)^2/19.048</td>
</tr>
<tr>
<td>Spend same</td>
<td>(5-10.476)^2/10.476</td>
<td>(15-9.524)^2/9.524</td>
</tr>
<tr>
<td>Spend less</td>
<td>(35-23.571)^2/23.571</td>
<td>(10-21.429)^2/21.429</td>
</tr>
</tbody>
</table>

Chi-square

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spend more</td>
<td>1.691</td>
<td>1.860</td>
</tr>
<tr>
<td>Spend same</td>
<td>2.862</td>
<td>3.149</td>
</tr>
<tr>
<td>Spend less</td>
<td>5.542</td>
<td>6.096</td>
</tr>
<tr>
<td></td>
<td>21.200</td>
<td></td>
</tr>
</tbody>
</table>

\[ X^2 = 21.2 \]

V. Decide Results of Null Hypothesis

Since the chi-square test statistic 21.2 exceeds the critical value of 5.991, conclude there is a statistically significant association between a person's sex and their attitudes toward spending on athletic programs. As is apparent in the contingency table, males are more likely to support spending on athletic programs than females.

Standardized Residuals

Standardized residuals are used to determine what categories (cells) were major contributors to rejecting the null hypothesis. When the absolute value of the residual (R) is greater than 2.00, the researcher can conclude it was a major influence on a significant chi-square test statistic.
Example using the observed and expected frequencies in the previous example:

\[
R = \frac{Fo - Fe}{\sqrt{Fe}}
\]

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>Absolute R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>Spend More</td>
<td>-1.300</td>
<td>1.364</td>
</tr>
<tr>
<td>Spend Same</td>
<td>-1.692</td>
<td>1.774</td>
</tr>
<tr>
<td>Spend Less</td>
<td><strong>2.354</strong></td>
<td><strong>-2.469</strong></td>
</tr>
</tbody>
</table>

Interpretation

[a] Attitudes toward spending less for both females and males had a major contribution to the chi-square result.
6.3 Coefficients for Measuring Association

The following are a few of several measures of association used with chi-square and other contingency table analyses. When using the chi-square statistic, these coefficients can be helpful in interpreting the strength of the relationship between two variables once statistical significance has been established. The logic for using measures of association is as follows:

Even though a chi-square test may show statistical significance between two variables, the relationship between those variables may not be substantively important. Measures of association are available to help evaluate the relative strength of a statistically significant relationship. In most cases, they are not used in interpreting the data unless the chi-square statistic first shows there is statistical significance (i.e., it doesn’t make sense to say there is a strong relationship between two variables when the statistical test shows this relationship is not statistically significant).

**Nominal and Ordinal Variables**

**Phi**

Phi is only used on 2x2 contingency tables. It is interpreted as a measure of the relative (strength) of an association between two variables ranging from 0 to 1.

\[ Phi = \sqrt{\frac{X^2}{n}} \]

**Pearson’s Contingency Coefficient (C)**

Pearson’s contingency coefficient is interpreted as a measure of the relative (strength) of an association between two variables. The coefficient will always be less than 1 and varies according to the number of rows and columns.
Cramer's V Coefficient (V)

Cramer's V is useful for comparing multiple $X^2$ test statistics and is comparable across contingency tables of varying sizes. It is not affected by sample size and therefore is very useful in situations where a statistically significant chi-square may be the result of large sample size instead of any substantive relationship between the variables. It is interpreted as a measure of the relative (strength) of an association between two variables. The coefficient ranges from 0 to 1 (perfect association). In practice, a Cramer's V of .10 may provide a good minimum threshold for suggesting there may be a substantive relationship between two variables.

$$V = \sqrt{\frac{X^2}{n + X^2}} \quad \text{where} \quad q = \text{smaller \# \ of \ rows \ or \ columns}$$

Describing Strength of Association

**Characterizations**

- >.50  high association
- .30 to .50  moderate association
- .10 to .30  low association
- .01 to .10  little if any association
Proportional Reduction of Error (PRE)

Lambda

Lambda is a proportional reduction in error (PRE) measure that ranges from 0 to 1. Lambda indicates the extent to which the independent variable reduces the error associated with predicting the value of a dependent variable. Multiplied by 100, it represents the percent reduction in error.

Ordinal Variables Only

Gamma

Gamma is another PRE measure ranging from -1 to 1 that estimates the extent errors are reduced in predicting the order of paired cases. Gamma ignores ties.

Kendall’s Tau b

Tau b is similar to Gamma but includes ties. It can range from -1 to 1 but since standardization is different from Gamma, it provides no clear explanation of PRE.

Inter-rater Agreement

Cohen’s Kappa

Cohen’s kappa measures agreement beyond chance. Although a negative value is possible, it commonly ranges from 0 to 1 (perfect agreement). This measure requires a balanced table where the number of rows is the same as the number of columns. The diagonal cells represent agreement.
### Software Output: Survey of U.S. adults.

Crosstabulation: **GUNLAW** (Rows) by **SEX** (Columns)

**Column Variable Label:** Respondent's Sex  
**Row Variable Label:** Gun permits

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favor</td>
<td>314</td>
<td>497</td>
<td>811</td>
</tr>
<tr>
<td></td>
<td>38.72</td>
<td>61.28</td>
<td></td>
</tr>
<tr>
<td></td>
<td>73.88</td>
<td>88.91</td>
<td>82.42</td>
</tr>
<tr>
<td></td>
<td>31.91</td>
<td>50.51</td>
<td></td>
</tr>
<tr>
<td>Oppose</td>
<td>111</td>
<td>62</td>
<td>173</td>
</tr>
<tr>
<td></td>
<td>64.16</td>
<td>35.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>26.12</td>
<td>11.09</td>
<td>17.58</td>
</tr>
<tr>
<td></td>
<td>11.28</td>
<td>6.30</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>425</td>
<td>559</td>
<td>984</td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Favor</td>
<td>314</td>
<td>497</td>
</tr>
<tr>
<td></td>
<td>38.72</td>
<td>61.28</td>
</tr>
<tr>
<td></td>
<td>73.88</td>
<td>88.91</td>
</tr>
<tr>
<td></td>
<td>31.91</td>
<td>50.51</td>
</tr>
<tr>
<td>Oppose</td>
<td>111</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>64.16</td>
<td>35.84</td>
</tr>
<tr>
<td></td>
<td>26.12</td>
<td>11.09</td>
</tr>
<tr>
<td></td>
<td>11.28</td>
<td>6.30</td>
</tr>
<tr>
<td>Total</td>
<td>425</td>
<td>559</td>
</tr>
</tbody>
</table>

### Chi-square Measures

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>DF</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>37.622</td>
<td>1</td>
<td>0.0000 [a]</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>37.417</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>Yate's Correction</td>
<td>36.592</td>
<td>1</td>
<td>0.0000 [b]</td>
</tr>
</tbody>
</table>

**Measures of Association**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cramer's V</td>
<td>.196</td>
<td>[c]</td>
</tr>
<tr>
<td>Pearson C</td>
<td>.192</td>
<td></td>
</tr>
<tr>
<td>Lambda Symmetric</td>
<td>.082</td>
<td>[d]</td>
</tr>
<tr>
<td>Lambda Dependent=Column</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Lambda Dependent=Row</td>
<td>.115</td>
<td>[e]</td>
</tr>
</tbody>
</table>

Note: 00.00% of the cells have an expected frequency <5
Interpretation

[a] The association between opinion on gun control and respondent sex is statistically significant. This is the most common measure used for chi-square significance.

[b] When sample sizes are small, the continuous chi-square value tends to be too large. The Yates continuity correction adjusts for this bias in 2x2 contingency tables. Regardless of sample size, it is a preferred measure for chi-square tests on 2x2 tables.


[d] A symmetric lambda is used when identification of independent and dependent variables is not useful.

[e] Knowing a person’s sex can reduce prediction error by 11.5%.
Summary Table

This table includes descriptive statistics for nominal level data using counts and percentages. It also includes inferential statistics using confidence intervals for proportions, chi-square, and measures of association. Five calculations were required to create the margin of errors. Four crosstabulations are presented in one table using column percent (Sex Education by Sex, Sex Education by Race, Gun Control by Sex, Gun Control by Race).

Table 1: Attitudes toward sex education and gun permit policy issues by sex and race of respondent (percent).

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Sex</th>
<th>Male</th>
<th>Female</th>
<th>Race</th>
<th>White</th>
<th>Non-White</th>
<th>P &lt;** (Cramer's V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>75</td>
<td>36</td>
<td>39</td>
<td>63</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Margin of Error</td>
<td>11.3%</td>
<td>16.3%</td>
<td>15.7%</td>
<td>12.4%</td>
<td>28.3%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sex Education in School

<table>
<thead>
<tr>
<th>Favor</th>
<th>Oppose</th>
<th>P &lt;** (Cramer's V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>72%</td>
<td>28%</td>
<td>.578 [b]</td>
</tr>
<tr>
<td>75%</td>
<td>25%</td>
<td>(.064)</td>
</tr>
<tr>
<td>69%</td>
<td>31%</td>
<td>(.282) [d]</td>
</tr>
</tbody>
</table>

Require Gun Permits

<table>
<thead>
<tr>
<th>Favor</th>
<th>Oppose</th>
<th>P &lt;** (Cramer's V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>48%</td>
<td>52%</td>
<td>.015 [c]</td>
</tr>
<tr>
<td>33%</td>
<td>67%</td>
<td>(.282) [d]</td>
</tr>
<tr>
<td>62%</td>
<td>38%</td>
<td>(.017)</td>
</tr>
<tr>
<td>48%</td>
<td>52%</td>
<td>(.888)</td>
</tr>
<tr>
<td>50%</td>
<td>50%</td>
<td>(.017)</td>
</tr>
</tbody>
</table>

Source: 1993 General Social Survey of U.S. Adults

* Based on alpha .05 (95% confidence interval)

** Significance is based on the chi-square test of independence.
Interpretation

[a] Based on a 95% confidence level, the proportion of all U.S. adults who favor requiring gun permits is 48% plus or minus 11.3% or between 37% to 59%.

[b] A statistically significant relationship between sex and attitudes toward sex education in school is not evident (p > .05). Based on this random sample, there is a lack of convincing evidence for a relationship in the underlying population.

[c] There is a statistically significant relationship between sex and attitudes toward gun permits (p < .05). Women are significantly more likely (62%) to favor gun permits than men (33%).

[d] There is a weak to moderate association between sex and attitudes toward gun permits based on a Cramer’s V of .282.
This section presents hypothesis testing techniques for interval/ratio data. The type of analytical technique used will depend on the level of the independent variable.

### Examples of Inferential Statistics using Interval/Ratio Data

<table>
<thead>
<tr>
<th>Statistical Technique</th>
<th>Dependent Variable</th>
<th>Independent Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confidence interval</td>
<td>miles to work [c]</td>
<td>None</td>
</tr>
<tr>
<td>t-test</td>
<td>annual income [c]</td>
<td>Sex [a]</td>
</tr>
<tr>
<td>ANOVA</td>
<td>physicians per capita [c]</td>
<td>census region [a]</td>
</tr>
<tr>
<td>ANOVA</td>
<td>years education [c]</td>
<td>Military rank [b]</td>
</tr>
<tr>
<td>Correlation, regression</td>
<td>annual income [c]</td>
<td>education [c]</td>
</tr>
<tr>
<td>Correlation, regression</td>
<td>weight (kilograms) [c]</td>
<td>height (centimeters) [c]</td>
</tr>
</tbody>
</table>

[a] Nominal  [b] Ordinal  [c] Interval/Ratio

This chapter focuses on t-tests and ANOVA. Later chapters will address correlation and regression.
7.1 T-test of Means

For inferential statistics (confidence intervals and significance tests), the data are assumed to be from a random sample. To conduct significance tests, the test statistic is compared to a critical value from a sampling distribution. Tables of critical values are included in the Appendix. The relevant sampling distributions for this section are the t distribution and f distribution.

Statistical Techniques

★ Population Mean to a Sample Mean (One-Sample Test)

★ Independent Samples With Equal Variance (Two-Sample Test)

★ Independent Samples Without Equal Variance (Two-Sample Test)

Comparing a Population Mean to a Sample Mean (One-Sample Test)

A one-sample t-test is used to compare a mean from a random sample to the mean (mu) of a population. It is especially useful to compare a mean from a random sample to an established data source such as census data to determine if a sample is unbiased (i.e., representative of the underlying population). Examples include comparing mean age and education levels of survey respondents to known values in the population.

Problem: Compare the mean age of incoming students to the known mean age for all previous incoming students. A random sample of 30 incoming college freshmen revealed the following statistics: mean age 19.5 years, standard deviation 1 year. The college database shows the mean age for previous incoming students was 18.
I. **Assumptions**

- Interval/ratio level data
- Random sampling
- Normal distribution in population

II. **State the Hypothesis**

**Ho:** There is no significant difference between the mean age of past college students and the mean age of current incoming college students.

\[ \mu = \bar{X} \]

**Ha:** There is a significant difference between the mean age of past college students and the mean age of current incoming college students.

\[ \mu \neq \bar{X} \]

III. **Set the Rejection Criteria**

- Significance level .05 alpha, 2-tailed test
- Degrees of Freedom = n-1 or 29
- Critical value from **t-distribution** = 2.045

IV. **Compute the Test Statistic**

*Standard error of the sample mean*

\[ s_{\bar{x}} = \frac{1}{\sqrt{30}} \]

\[ s_{\bar{x}} = .183 \]
Test statistic

\[ t = \frac{19.5 - 18}{1.83} \quad t = 8.197 \]

V. Decide Results of Null Hypothesis

Given that the test statistic (8.197) exceeds the critical value (2.045), the null hypothesis is rejected in favor of the alternative. There is a statistically significant difference between the mean age of the current class of incoming students and the mean age of freshman students from past years. In other words, this year’s freshman class is on average older than freshmen from prior years.

If the results had not been significant, the null hypothesis would not have been rejected. This would be interpreted as the following: There is insufficient evidence to conclude there is a statistically significant difference in the ages of current and past freshman students.

Software Output: Summary data from previous example

One-Sample Difference of Means

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Mean</td>
<td>18.0000</td>
</tr>
<tr>
<td>Sample Mean</td>
<td>19.5000</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>1.0000</td>
</tr>
<tr>
<td>Sample Size (n)</td>
<td>30</td>
</tr>
<tr>
<td>Test statistic</td>
<td>8.2160</td>
</tr>
<tr>
<td>p &lt; (0.0001) (2-tailed)</td>
<td>[a]</td>
</tr>
</tbody>
</table>

Interpretation

[a] For a 2-tailed test, the p-value represents the probability of making a type 1 error (concluding there is statistical significance when there is none). Since there is less than a .01 % (p < .0001) chance of making a type 1 error, there is statistical significance.
Comparing Two Independent Sample Means (Two-Sample Test)

With Homogeneity of Variance $\sigma_1^2 = \sigma_2^2$

A two-sample t-test is used to compare two sample means. The independent variable is nominal level data and the dependent variable is interval/ratio level data.

**Problem:** The number of years of education were collected from one random sample of 38 police officers from City A and a second random sample of 30 police officers from City B. The average years of education for the sample from City A is 15 years with a standard deviation of 2 years. The average years of education for the sample from City B is 14 years with a standard deviation of 2.5 years. Is there a statistically significant difference between the education levels of police officers in City A and City B?

I. **Assumptions**

   - Random sampling
   - Independent samples
   - Interval/ratio level data

II. **State Hypotheses**

   Ho: There is no statistically significant difference between the mean education level of police officers working in City A and the mean education level of police officers working in City B. $\mu_1 = \mu_2$
For a 2-tailed hypothesis test

Ha: There is a statistically significant difference between the mean education level of police officers working in City A and the mean education level of police officers working in City B. $\mu_1 \neq \mu_2$

For a 1-tailed hypothesis test

Ha: The mean education level of police officers working in City A is significantly greater than the mean education level of police officers working in City B.

III. Set the Rejection Criteria

Determine the degrees of freedom (df) = (n1+n2)-2  \hspace{1cm} df = 38+30-2=66

Determine level of confidence -- alpha (1 or 2-tailed test)

Use the *t*-distribution table to determine the critical value

- If using 2-tailed test, Alpha.05, $t_{cv} = 1.997$
- If using 1-tailed test, Alpha.05, $t_{cv} = 1.668$

IV. Compute Test Statistic

*Standard error*

$$S_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{n_1 + n_2}{n_1n_2}}$$

$$S_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{37(4) + 29(6.25)}{38 + 30 - 2}} \sqrt{\frac{38 + 30}{38(30)}} \Rightarrow S_{\bar{x}_1-\bar{x}_2} = .545$$
Test Statistic

\[ t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} \]

\[ t = \frac{15 - 14}{.545} \]

\[ t = 1.835 \]

V. Decide Results

If using 2-tailed test . . .

There is no statistically significant difference between the mean years of education for police officers in City A and mean years of education for police officers in City B. The test statistic 1.835 does not meet or exceed the critical value of 1.997 for a 2-tailed test.

If using 1-tailed test . . .

Police officers in City A have significantly more years of education than police officers in City B. The test statistic 1.835 exceeds the critical value of 1.668 for a 1-tailed test.
**Software Output:** Uses survey data to compare the mean incomes of males and females.

Two-Sample Difference of Means

Independent Variable: SEX  
Dependent Variable: INCOME

Sample One: 0 Female  
Sample Two: 1 Male

<table>
<thead>
<tr>
<th></th>
<th>Sample 1</th>
<th>Sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

| Sample Mean     | 12279.4667 | 15266.4706 |
| Std Deviation   | 4144.6323  | 3947.7352  |
| Sample Size (n) | 15         | 17         |

**Homogeneity of Variance**

<table>
<thead>
<tr>
<th>F-ratio</th>
<th>DF (14, 16)</th>
<th>p &lt; 0.4283</th>
</tr>
</thead>
</table>

**T-statistic**

| Equal Variance | -2.0870 | 30 | 0.0455 |
| Unequal Variance | -2.0800 | 31.05 | 0.0459 |

**Interpretation**

[a] The F-ratio is not statistically significant. Therefore, use the equal variance test statistic.

[b] For a 2-tailed test, the p-value represents the probability of making a type 1 error (concluding there is statistical significance when there is none). Since there is about 4.6% (p < .0455) chance of making a type 1 error, which does not exceed the 5% error limit (p=.05) established in the rejection criteria of the hypothesis testing process (alpha), conclude there is statistical significance.
Computing F-ratio

The F-ratio is used to determine whether the variances in two independent samples are equal. If the F-ratio is not statistically significant, assume there is homogeneity of variance and employ the standard t-test for the difference of means. If the F-ratio is statistically significant, use an alternative t-test computation such as the Cochran and Cox method.

**Problem**: Given the following summary statistics, are the variances equal?

Sample A  \( s^2 = 20 \)  \( n=10 \)

Sample B  \( s^2 = 30 \)  \( n=30 \)

**Set the Rejection Criteria**

Determine the degrees of freedom (df) for each sample

* **Numerator of the ratio is the sample with the larger variance (Sample B).**

  Numerator \( df = n - 1 \)  \( df \) for numerator (Sample B) = 29

* **Denominator of the ratio is the sample with the smaller variance (Sample A).**

  Denominator \( df = n - 1 \)  \( df \) for denominator (Sample A) = 9

Determine the level of confidence -- alpha

Consult [F-Distribution table](#) for \( df = (29,9), \alpha = 0.05 \)

\( F_{cv} = 2.70 \)
Compute the Test Statistic

\[ Fratio = \frac{S_1^2}{S_2^2} \]

where

\[ S_1^2 = \text{largest variance (Sample B)} \]

\[ S_2^2 = \text{smallest variance (Sample A)} \]

\[ F = \frac{30}{20} \quad F = 1.50 \]

Decide Results

Compare the test statistic with the f critical value (Fcv) listed in the F distribution. If the F-ratio equals or exceeds the critical value, the null hypothesis (Ho) \( \sigma_1^2 = \sigma_2^2 \) (there is no difference between the sample variances) is rejected. If there is a difference in the sample variances, the comparison of two independent means should involve the use of the Cochran and Cox method or one of several alternative techniques.

The test statistic (1.50) did not meet or exceed the critical value (2.70). Therefore, there is no statistically significant difference between the variance exhibited in Sample A and the variance exhibited in Sample B. Assume homogeneity of variance for tests of the difference between sample means.
Software Output: Comparing the mean incomes of high school graduates and high school dropouts from survey data.

Two-Sample Difference of Means

Independent Variable: EduCat Education Level
Dependent Variable: Income Respondent individual income

Sample One: 1 <12 yrs
Sample Two: 2 HS Grad

<table>
<thead>
<tr>
<th></th>
<th>Sample 1</th>
<th>Sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;12 yrs</td>
<td>23042.5000</td>
<td>31000.8235</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>4450.5413</td>
<td>9442.4477</td>
</tr>
<tr>
<td>Sample Size (n)</td>
<td>6</td>
<td>17</td>
</tr>
</tbody>
</table>

Homogeneity of Variance

<table>
<thead>
<tr>
<th></th>
<th>F-ratio</th>
<th>DF (16, 5)</th>
<th>p &lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal Variance</td>
<td>-1.9660</td>
<td>21</td>
<td>0.0627 [b]</td>
</tr>
<tr>
<td>Unequal Variance</td>
<td>-2.7220</td>
<td>21.67</td>
<td>0.0124 [c]</td>
</tr>
</tbody>
</table>

Interpretation

[a] The F-ratio is statistically significant. Use the unequal variance test statistic.

[b] If homogeneity of variance had not been considered, the research may have erroneously failed to find statistical significance.

[c] Since there is only a 1.2% (p < .0124) chance of making a type 1 error, which does not exceed the 5% error limit (p=.05, alpha), conclude there is statistical significance.
7.2 One-Way Analysis of Variance (ANOVA)

One-way analysis of variance (ANOVA) is used to evaluate means from two or more subgroups. A statistically significant ANOVA indicates there is more variation between subgroups than would be expected by chance. It does not identify which subgroup pairs are significantly different from each other. ANOVA is used to evaluate multiple means of one independent variable to avoid conducting multiple t-tests (see multiple comparison problem in section 6.3).

Problem: The number of years of education were obtained from one random sample of 38 police officers from City A, a second random sample of 30 police officers from City B, and a third random sample of 45 police officers from City C. The average years of education for the sample from City A is 15 years with a standard deviation of 2 years. The average years of education for the sample from City B is 14 years with a standard deviation of 2.5 years. The average years of education for the sample from City C is 16 years with a standard deviation of 1.2 years.

Is there a statistically significant difference between the education levels of police officers in City A, City B, and City C?

Data For Computations

<table>
<thead>
<tr>
<th></th>
<th>City A</th>
<th>City B</th>
<th>City C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (years)</td>
<td>15</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>S (Standard Deviation)</td>
<td>2</td>
<td>2.5</td>
<td>1.2</td>
</tr>
<tr>
<td>S² (Variance)</td>
<td>4</td>
<td>6.25</td>
<td>1.44</td>
</tr>
<tr>
<td>n (number of cases)</td>
<td>38</td>
<td>30</td>
<td>45</td>
</tr>
<tr>
<td>Sum of Squares</td>
<td>152</td>
<td>187.5</td>
<td>64.8</td>
</tr>
</tbody>
</table>
I. Assumptions

Independent Random sampling

Interval/ratio level data

Population variances equal

Groups are normally distributed

II. State the Hypothesis

Ho: There is no statistically significant difference among the three cities in the mean years of education for police officers. $\mu_1 = \mu_2 = \mu_3$

Ha: There is a statistically significant difference among the three cities in the mean years of education for police officers. $\mu_1 \neq \mu_2 \neq \mu_3$

III. Set the Rejection Criteria

Determine the degrees of freedom for the F Distribution

Numerator Degrees of Freedom

df=k-1 where k=3 (number of independent samples/groups)  df=2

Denominator Degrees of Freedom

df=n-k where n=113 (sum of all independent samples)  df=110

Establish Critical Value

Determine the level of confidence -- alpha

At alpha.05, df=(2,110)
Consult \textit{f-distribution}, Fcv = 3.072

IV. Compute the Test Statistic

\[
F = \frac{\sum n_k (\bar{X}_i - \bar{X}_g)^2 / (k-1)}{\left[ \sum (X_i - \bar{X}_1)^2 + \sum (X_i - \bar{X}_2)^2 + \sum (X_i - \bar{X}_3)^2 \right] / (n-k)}
\]

where

\( \bar{X}_i \) = each group mean

\( \bar{X}_g \) = grand mean for all the groups = (sum of all scores)/N

\( n_k \) = number in each group

Note: \( F = \) Mean Squares between groups

\( \) Mean Squares within groups

Estimate Grand Mean

\[
\bar{X}_g = \frac{\sum X_1 + \sum X_2 + \sum X_3}{n_1 + n_2 + n_3} \quad \bar{X}_g = \frac{570 + 420 + 720}{38 + 30 + 45} \quad \bar{X}_g = \frac{1710}{113} \quad \bar{X}_g = 15.133
\]

Estimate F Statistic

\[
F = \frac{\sum n_k (\bar{X}_i - \bar{X}_g)^2 / (k-1)}{\left[ \sum (X_i - \bar{X}_1)^2 + \sum (X_i - \bar{X}_2)^2 + \sum (X_i - \bar{X}_3)^2 \right] / (n-k)}
\]

\[
F = \frac{38(15 - 15.133)^2 + 30(14 - 15.133)^2 + 45(16 - 15.133)^2 / (3 - 1)}{152 + 187.5 + 64.8} / (113 - 3)
\]
\[ F = \frac{(0.672 + 38.51 + 33.826)/2}{3.676} \]
\[ F = 9.931 \]

V. **Decide Results of Null Hypothesis**

Compare the F statistic to the F critical value. If the F statistic equals or exceeds the Fcv, the null hypothesis is rejected. This suggests that the population means of the groups sampled are not equal -- there is a difference between the group means.

Since the F-statistic (9.931) exceeds the F critical value (3.072), we reject the null hypothesis and conclude there is a statistically significant difference between the three cities in the mean years of education for police officers.
**Software Output:** Comparing amount of time spent watching TV per day by age group.

Analysis of Variance (ANOVA): TVHOURS (Means) by AGECAT4 (Groups)

**Independent Variable Label:** Age categories
**Dependent Variable Label:** Hours watch TV per day

**Group Summary Statistics**

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>18-29</td>
<td>185</td>
<td>3.0108</td>
<td>2.2698</td>
</tr>
<tr>
<td>30-39</td>
<td>243</td>
<td>2.6008</td>
<td>2.0331</td>
</tr>
<tr>
<td>40-49</td>
<td>198</td>
<td>2.5859</td>
<td>1.7712</td>
</tr>
<tr>
<td>50+</td>
<td>364</td>
<td>3.2665</td>
<td>2.6385</td>
</tr>
</tbody>
</table>

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>92.09</td>
<td>3 [a]</td>
</tr>
<tr>
<td>Within Groups</td>
<td>5093.45</td>
<td>986 [c]</td>
</tr>
<tr>
<td>Total</td>
<td>5185.54</td>
<td>989 [e]</td>
</tr>
</tbody>
</table>

**F Statistic**

<table>
<thead>
<tr>
<th>F Statistic</th>
<th>p &lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.94 [f]</td>
<td>0.0005 [g]</td>
</tr>
</tbody>
</table>

**Interpretation**

[a] k-1, where k = 4 group means
[b] Between Groups Sum of Squares ÷ degrees of freedom
[c] n-k, where n = 990
[d] Within Groups Sum of Squares ÷ degrees of freedom
[e] n-1
Between Groups Mean Squares ÷ Within Groups Mean Squares

The F-Statistic is statistically significant. The p-value represents the probability of making a type 1 error (concluding there is statistical significance when there is none). Since there is about .05% (p < .0005) chance of making a type 1 error, conclude there is statistical significance; ie, there is variation among the groups.
7.3 Multiple Comparison Problem
(Post hoc comparisons)

As statistical tests are repeatedly run between subgroups within the same variable, the probability of making a type one error increases (i.e., if 100 t-tests were conducted, by random chance 5 out of 100 may be statistically significant when in fact they are not). Post hoc comparisons adjust for the problem of multiple comparisons and provide information regarding which subgroup pairs have a statistically significant difference. There are several alternative approaches to conducting post hoc comparisons. Bonferroni is one of the most conservative (less likely to find statistical significance).

**Software Output:** The following presents post hoc comparisons for the preceding ANOVA output.

**Bonferroni Post Hoc Comparison of Means**

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>G1-G2</th>
<th>SE</th>
<th>p&lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>18–29</td>
<td>30–39</td>
<td>0.410</td>
<td>0.209</td>
<td>0.3005</td>
</tr>
<tr>
<td>18–29</td>
<td>40–49</td>
<td>0.425</td>
<td>0.207</td>
<td>0.2464</td>
</tr>
<tr>
<td>18–29</td>
<td>50+</td>
<td>-0.256</td>
<td>0.228</td>
<td>1.0000</td>
</tr>
<tr>
<td>30–39</td>
<td>18–29</td>
<td>-0.410</td>
<td>0.209</td>
<td>0.3005</td>
</tr>
<tr>
<td>30–39</td>
<td>40–49</td>
<td>0.015</td>
<td>0.184</td>
<td>1.0000</td>
</tr>
<tr>
<td>30–39</td>
<td>50+</td>
<td>-0.666[a]</td>
<td>0.200</td>
<td>0.0056[b]</td>
</tr>
<tr>
<td>40–49</td>
<td>18–29</td>
<td>-0.425</td>
<td>0.207</td>
<td>0.2464</td>
</tr>
<tr>
<td>40–49</td>
<td>30–39</td>
<td>-0.015</td>
<td>0.184</td>
<td>1.0000</td>
</tr>
<tr>
<td>40–49</td>
<td>50+</td>
<td>-0.681</td>
<td>0.209</td>
<td>0.0073[c]</td>
</tr>
<tr>
<td>50+</td>
<td>18–29</td>
<td>0.256</td>
<td>0.228</td>
<td>1.0000</td>
</tr>
<tr>
<td>50+</td>
<td>30–39</td>
<td>0.666</td>
<td>0.200</td>
<td>0.0056</td>
</tr>
<tr>
<td>50+</td>
<td>40–49</td>
<td>0.681</td>
<td>0.209</td>
<td>0.0073</td>
</tr>
</tbody>
</table>
Interpretation

[a] The difference between the mean of those 30-39 (G1) and the mean of those 50 or older (G2) is a -0.666. A negative mean indicates those 50 or older have a greater mean number of hours watching television than those 30-39.

[b] There is a statistically significant difference between the number of hours watching television for those between the ages of 30 and 39 and the number of hours watching television for those 50 and older.

[c] There is a statistically significant difference between the number of hours watching television for those between the ages of 40 and 49 and the number of hours watching television for those 50 and older.
Summary Report

This table uses descriptive statistics (means), t-tests, and ANOVA to compare city manager characteristics. The independent variables are sex and city size (nominal/ordinal). The dependent variables are age, education, and years as a manager (interval/ratio).

Table 2: City manager characteristics (means) by sex and city size category.

<table>
<thead>
<tr>
<th>Sex</th>
<th>City Size Category</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
</tr>
<tr>
<td>Sample size &gt;</td>
<td>30</td>
</tr>
</tbody>
</table>

Manager Characteristics

<table>
<thead>
<tr>
<th></th>
<th>Sex</th>
<th>City Size Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>46.80</td>
<td>49.68</td>
</tr>
<tr>
<td>Years of Education</td>
<td>16.63</td>
<td>16.68</td>
</tr>
<tr>
<td>Years as Manager</td>
<td>16.08</td>
<td>18.40</td>
</tr>
</tbody>
</table>

Skill Preference Score

<table>
<thead>
<tr>
<th>Skill Preference Score</th>
<th>Sex</th>
<th>City Size Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negotiate-Analytical</td>
<td>-0.15</td>
<td>-0.03</td>
</tr>
</tbody>
</table>

Source: 1996 Survey of U.S. City Managers

* T-test of means, 2-tailed
** ANOVA
*** Based on two measures of skill importance (larger score indicates higher importance). The preference score was created by subtracting the analytical score (management science) from the negotiation score (political savvy). A positive preference score indicates a city manager finds negotiation skills more important than analytical skills. A negative preference score indicates a city manager finds analytical skills more important than negotiation skills.
Interpretation

[a] There is a significant difference between male and female city managers in mean age. Male city managers (49.7 years) are significantly older than female city managers (41.8 years).

[b] There is a significant difference among small, medium, and large cities in the mean ages of city managers. It appears that larger size cities are likely to have older city managers than smaller cities. This conclusion should be confirmed with a post-host analysis.

[c] Males have significantly greater years experience (about 6 years more on average) as city managers than females.

[d] There is a significant difference among small, medium, and large cities in the mean skill preferences of city managers. It appears that large cities are likely to have city managers who place more importance on negotiation skills than analytical skills. In contrast, small and medium size cities prefer analytical skills over negotiation. This conclusion should be confirmed with a post-host analysis.
8 Correlation

8.1 Pearson's Product Moment Correlation Coefficient

Correlation coefficients estimate strength and direction of association between two interval/ratio level variables. The Pearson Correlation Coefficient presented here can range from a -1.00 to 1.00. A positive coefficient indicates the values of variable A vary in the same direction as variable B. A negative coefficient indicates the values of variable A and variable B vary in opposite directions.

Example: The following data were collected to estimate the correlation between years of formal education and income at age 35.

<table>
<thead>
<tr>
<th>Susan</th>
<th>Bill</th>
<th>Bob</th>
<th>Tracy</th>
<th>Joan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education (years)</td>
<td>12</td>
<td>14</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>Income ($1000)</td>
<td>25</td>
<td>27</td>
<td>32</td>
<td>44</td>
</tr>
</tbody>
</table>

Verify Conditions for using Pearson r

Interval/ratio data must be from paired observations.

A linear relationship should exist between the variables.

No extreme values in the data.
The following scattergram assists in evaluating linearity and also helps to identify problems related to extreme values.

Y: Income

<table>
<thead>
<tr>
<th></th>
<th>*</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>44.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25.0</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>12.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

X: Education

Compute Pearson's r

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>X</td>
<td>Y</td>
<td>XY</td>
<td>X²</td>
<td>Y²</td>
</tr>
<tr>
<td>-------</td>
<td>---</td>
<td>---</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Susan</td>
<td>12</td>
<td>25</td>
<td>300</td>
<td>144</td>
<td>625</td>
</tr>
<tr>
<td>Bill</td>
<td>14</td>
<td>27</td>
<td>378</td>
<td>196</td>
<td>729</td>
</tr>
<tr>
<td>Bob</td>
<td>16</td>
<td>32</td>
<td>512</td>
<td>256</td>
<td>1024</td>
</tr>
<tr>
<td>Tracy</td>
<td>18</td>
<td>44</td>
<td>792</td>
<td>324</td>
<td>1936</td>
</tr>
<tr>
<td>Joan</td>
<td>12</td>
<td>26</td>
<td>312</td>
<td>144</td>
<td>676</td>
</tr>
<tr>
<td>Σ</td>
<td>72</td>
<td>154</td>
<td>2294</td>
<td>1064</td>
<td>4990</td>
</tr>
<tr>
<td>n</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
\[
 r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left[n\sum X^2 - (\sum X)^2\right]\left[n\sum Y^2 - (\sum Y)^2\right]}} 
\]

\[
 r_{xy} = \frac{5(2294) - 72(154)}{\sqrt{\left[5(1064) - 72^2\right]\left[5(4990) - 154^2\right]}} 
\]

\[
 r_{xy} = \frac{382}{\sqrt{167824}} \quad r_{xy} = \frac{382}{409.663} \quad r_{xy} = .933 
\]

**Interpret**

A positive coefficient indicates the values of variable A vary in the same direction as variable B. A negative coefficient indicates the values of variable A and variable B vary in opposite directions.

**Characterizations of Pearson r**

.9 to 1 very high correlation

.7 to .9 high correlation

.5 to .7 moderate correlation

.3 to .5 low correlation

.0 to .3 little if any correlation

In this example, there is a very high positive correlation between the variation of education and the variation of income. Individuals with higher levels of education earn more than those with comparably lower levels of education.
Coefficient of Determination

The coefficient of determination represents the proportion of variation in the dependent variable (Y) explained by the independent variable (X). In the case of Pearson r, the coefficient of determination can be obtained by squaring the Pearson r coefficient.

Example: From previous example

\[
\text{Coefficient} = r^2 \quad r^2 = .933^2 \quad r^2 = .871
\]

Eighty-seven percent of the variance displayed in the income variable is associated with the variance displayed in the education variable.

Hypothesis Testing for Pearson r

Problem: Determine statistical significance based on a Pearson r of .933 for annual income and education obtained from a national random sample of 20 employed adults.

I. Assumptions

   Data originated from a random sample

   Data are interval/ratio

   Both variables are distributed normally

   Linear relationship and homoscedasticity

II. State the Hypothesis
Ho: There is no association between annual income and education for employed adults.

Ha: There is an association between annual income and education for employed adults.

III. Set the Rejection Criteria

Determine the degrees of freedom (df) \( df=n – 2 \) or 20-2=18

Determine the confidence level, alpha (1-tailed or 2-tailed)

Use the critical values from the t distribution at df=18

\[ t_{cv} @ .05 \text{ alpha (2-tailed)} = 2.101 \]

IV. Compute Test Statistic

\[ t = r \sqrt{ \frac{n-2}{1-r^2} } \]

\[ t = .933 \sqrt{ \frac{20-2}{1-.871} } = 11.022 \]

V. Decide Results

Since the test statistic 11.022 exceeds the critical value 2.101, there is a statistically significant association in the national population between an employed adult’s education and their annual income.
**Software Output:** Example of one bivariate comparison with a scattergram. Comparing individual income in U.S. dollars to years of education.

Pearson's Correlation

**INCOME:** [a]

- 42779.0
- 28030.0
- 12281.0

**EDUC:** [b]

Number of cases: 32
Missing: 0 [c]
Pearson Correlation: **0.751** [d]
p < (2-tailed signif.): **0.0000** [e]

**Interpretation**

[a] The Y axis of the scattergram. If theory suggests cause and effect, the Y axis is commonly used for the dependent (response) variable.

[b] The X axis of the scattergram. If theory suggests cause and effect, the X axis is commonly used for the independent variable.

[c] Since each observation (case) must have values for both income and education, any observations where one or both of these variables have no data will be removed from the analysis.

[d] Pearson correlation coefficient representing a high positive correlation between education and income. Interpretation: As years of education increases so does personal income.

[e] There is a statistically significant association between income and education.
**Software Output:** Example of bivariate comparisons displayed in a correlation matrix. Comparing individual income, education, and months of work experience.

### Correlation: Pearson (R) Coefficients

<table>
<thead>
<tr>
<th>Coeff</th>
<th>Correlation Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>INCOME</td>
</tr>
<tr>
<td>p &lt;</td>
<td>INCOME</td>
</tr>
<tr>
<td>INCOME</td>
<td>1.000 [a]</td>
</tr>
<tr>
<td></td>
<td>32 [b]</td>
</tr>
<tr>
<td></td>
<td>. [c]</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.751</td>
</tr>
<tr>
<td></td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>0.000</td>
</tr>
<tr>
<td>WORKEXP</td>
<td>-0.160</td>
</tr>
<tr>
<td></td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>0.381</td>
</tr>
</tbody>
</table>

2-tailed significance tests
'. ' p-value not computed

**Interpretation**

[a] The diagonal in the matrix represents Pearson correlations between the same variable which will always be 1 since the variables are identical. The correlations above the diagonal are a mirror reflection of these below the diagonal, so only interpret half of the matrix (in this case three correlation coefficients).

[b] The number of paired observations for this comparison.

[c] A statistical test is not performed when comparing a variable to itself. Mathematically this will always equal zero.
Pearson correlation coefficient representing a high positive correlation between education and income. Interpretation: As years of education increases so does personal income.

There is a statistically significant association between income and education.

Pearson correlation coefficient representing a very weak negative correlation between work experience and income. Interpretation: As years of work experience increases personal income decreases.

There is not a statistically significant association between work experience and income.

There is a statistically significant association between work experience and education.
8.2 Spearman Rho Coefficient

Spearman’s Rho is used to estimate strength and direction of association between two ordinal level variables (paired observations that are ranked). The Spearman Rho Coefficient presented here can range from a -1.00 to 1.00. A positive coefficient indicates the values of variable A vary in the same direction as variable B. A negative coefficient indicates the values of variable A and variable B vary in opposite directions.

Verify the conditions are appropriate

Scores of two variables are ranks

Problem: Five college students' have the following rankings in math and philosophy courses. Is there an association between student rankings in math and philosophy courses?

<table>
<thead>
<tr>
<th>Student</th>
<th>Alice</th>
<th>Jordan</th>
<th>Dexter</th>
<th>Betty</th>
<th>Corina</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math class rank</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Philosophy class rank</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

Compute Spearman Rho

\[
\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}
\]

where

\( n = \) number of paired ranks

\( d = \) difference between the paired ranks
Note: When two or more observations of one variable are the same, ranks are assigned by averaging positions occupied in their rank order.

Example:

<table>
<thead>
<tr>
<th>Score</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>6</th>
<th>6</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td>1</td>
<td>2</td>
<td>3.5</td>
<td>3.5</td>
<td>5</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Math Rank</th>
<th>Philosophy Rank</th>
<th>X-Y</th>
<th>(X-Y)^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Y</td>
<td>D</td>
<td>d^2</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>-4</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>9</td>
</tr>
</tbody>
</table>

\[
\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} \quad \rho = 1 - \frac{6(30)}{5(5^2 - 1)} \quad \rho = 1 - \frac{180}{5(25 - 1)} \quad \rho = 1 - \frac{180}{120} \\
\]

\[
\rho = -.500 
\]

Interpret Coefficient

There is a moderate negative correlation between the math and philosophy course rankings of students. Students who rank high as compared to other students in their math course generally have lower philosophy course ranks and those with low math rankings have higher philosophy course rankings than those with high math rankings.

Note: The formulas for Pearson r and Spearman rho are equivalent when there are no tied ranks.
Hypothesis Testing for Spearman Rho

Significance testing for inference to the population.

Problem: Based on a Spearman Rho of .70 and a sample size of 20, is there an association between the music and physics rankings of students statistically significant?

I. Assumptions

Data originated from a random sample

Data are ordinal

Both variables are distributed normally

Linear relationship and homoscedasticity

Sample size is $\geq 10$

II. State the Hypothesis

Null Hypothesis (Ho): There is no association between music and physics class rankings for all students in the population. $\rho = 0$

Alternative Hypothesis (Ha): There is association between music and physics class rankings for all students in the population. $\rho \neq 0$

III. Set the Rejection Criteria

Determine the degrees of freedom $n - 2$ or $20 - 2 = 18$

Determine the confidence level, alpha (1-tailed or 2-tailed)

Use the critical values from the $t$ distribution
Note: The t distribution should only be used when the sample size is 10 or more.

tcv @ .05 alpha (2-tailed) = 2.101

IV. Compute the Test Statistic

Convert $\rho$ into

$$t = \rho \sqrt{\frac{n - 2}{1 - \rho^2}} \quad t = .70 \sqrt{\frac{20 - 2}{1 - .70^2}} \quad t = \rho \sqrt{\frac{18}{1 - .49}} \quad t = 4.159$$

V. Decide Results of Null Hypothesis

Reject if the t-observed is equal to or greater than the critical value. The variables are/are not related in the population. In other words, the association displayed from the sample data can/cannot be inferred to the population.

Since the test statistic 4.159 exceeds the critical value 2.101, there is a statistically significant association in the national population between student music and physics rankings. Students who rank high in musical ability will also likely rank high in physics.
Simple OLS Regression

Linear ordinary least squares regression involves predicting the score for a dependent variable (Y) based on the score of an independent variable (X). Data are tabulated for two variables X and Y. If a linear relationship exists between the variables, it is appropriate to use linear regression to base predictions of the Y variable from values of the X variable. See the multiple regression chapter for further discussion of assumptions.

9.1 Procedure

1. Determine the regression line with an equation of a straight line.

2. Use the equation to predict scores.

3. Determine the "standard error of the estimate" (Sxy) to evaluate the distribution of scores around the predicted Y score.

Determine the regression line

Data are tabulated for two variables X and Y. Use Pearson’s r to help determine if there is a linear relationship between the variables. If a significant linear relationship
exists between the variables, it is appropriate to use linear regression to base predic-
tions of the Y variable on the relationship developed from the original data.

Equation for a straight line

\[ \hat{Y} = a + bx \]

where

\[ \hat{Y} = \text{predicted score} \]

b = slope of a regression line (regression coefficient)

\[
 b = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2}
\]

x = individual score from the X distribution

a = Y intercept (regression constant)

\[
a = \overline{Y} - b\overline{X}
\]

where

\[ \overline{X} = \text{mean of variable X} \]

\[ \overline{Y} = \text{mean of variable Y} \]

Use the equation to predict scores

Insert the values for the intercept (a) and slope (b) of the completed equation for a
straight line and select an individual X score to predict a Y score.

\[ \hat{Y} = a + bx \]
Determine the Standard Error of the Estimate

The standard error of the estimate is used to provide a confidence interval around a point estimate for predicted Y in the underlying population. In relative terms, larger standard errors indicate less prediction accuracy for the population value than equations with smaller standard errors.

\[ s_e = \sqrt{\frac{\sum \left( Y_i - \hat{Y} \right)^2}{n - 2}} \]

Confidence Interval for the Predicted Score

Sometimes referred to as the conditional mean of Y given a specific score of X.

\[ CI = \hat{Y} \pm t_{cv} \left( s_e \right) \]

**Determine critical value (t_{cv})**

Degrees of Freedom (df) = n-2

Select alpha (generally based on .05 for a 2-tailed test)

Obtain the critical value (t_{cv}) from the **t-distribution**
Problem: The following data were collected to estimate the correlation between years of formal education and income at age 35 and are the same data used in an earlier example to estimate Pearson r.

<table>
<thead>
<tr>
<th></th>
<th>Susan</th>
<th>Bill</th>
<th>Bob</th>
<th>Tracy</th>
<th>Joan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education (years)</td>
<td>12</td>
<td>14</td>
<td>16</td>
<td>18</td>
<td>12</td>
</tr>
<tr>
<td>Income ($1000)</td>
<td>25</td>
<td>27</td>
<td>32</td>
<td>44</td>
<td>26</td>
</tr>
</tbody>
</table>

Determine Regression Line

\[ b = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2} \]
\[ b = \frac{5(2294) - (72)(154)}{5(1064) - (5184)} \]
\[ b = 2.809 \]

\[ a = \overline{Y} - b \overline{X} \]
\[ a = 30.8 - 2.809(14.4) \]
\[ a = -9.650 \]

Given x (education) of 15 years, estimate predicted Y
\[ \hat{Y} = a + bx \]

\[ \hat{Y} = -9.650 + 2.809(15) \]

\[ \hat{Y} = 32.485 \] or an estimated annual income of $32,485 for a person with 15 years of education (this is the point estimate).

**Determine the Standard Error of the Estimate**

Standard error of the estimate of \( Y \)

\[
se = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}}
\]

\[
s_e = \sqrt{\frac{32.205}{3}}
\]

\[ s_e = 3.276 \]

**Determine the Confidence interval for predicted \( Y \)**

\[
CI = \hat{Y} \pm t_{cv} (s_e)
\]

where

Degrees of Freedom (df) = 5-2 or 3

Alpha .05, 2-tailed

Based on the **t-distribution** \( t_{cv} = 3.182 \)
Confidence interval for predicted $Y$

$$Y_c = 32.485 \pm 3.182(3.276)$$

$$Y_c = 32.485 \pm 10.424$$

Given a 5% chance of error, the estimated income for a person with 15 years of education will be $32,485$ plus or minus $10,424$ or somewhere between $22,060$ and $42,909$ (the interval estimate).

**Standard error of the slope estimate**

To develop confidence intervals or test hypotheses, we need to estimate the standard error of the slope estimate ($s_b$).

$$s_b = \sqrt{\frac{\sum(y_i - \hat{y})^2}{(n-2)\sum(x_i - \bar{X})^2}}$$

Based on the example data . . .

$$s_b = \sqrt{\frac{32.2/3}{27.4}} \quad s_b = .625$$

**Calculating Confidence Interval for the slope ($b$)**

$$CI = b \pm CV(s_b)$$
Where CV = critical value (t-distribution, 2-tailed, .05 alpha, df=n-2) and 2.8 is the point estimate for the slope.

\[
CI = 2.8 \pm 3.182(0.625) \quad \text{or simplify to} \quad CI = 2.8 \pm 1.998 \quad CI = 2.8 \pm 2.0
\]

**Interpretation**

The slope for education is between 0.8 and 4.8.
9.2 Hypothesis testing

The null hypothesis states that X is not related with Y (i.e., the slope is 0)

Ho: $\beta = 0$  (slope coefficient is zero)

Ha: $\beta \neq 0$  (slope coefficient is not zero)

Determine Critical Value

**t-distribution**, 2-tail, alpha .05, n-2=3

$b_{cv} = 3.182$

$t = \frac{\beta - 0}{s_\beta} = \frac{2.8}{.625} = 4.48$

Decision

Reject the null hypothesis. The slope for education is statistically significant (i.e., t test statistic of 4.48 is beyond the critical value of 3.182). As education in years increases so will income.
9.3 Evaluating the power of the regression model

If we only had information on Y (Income), our best guess of an individual's income would be the mean income. However, if we have a paired X variable (Education) that is related to Y, we can use this additional variable to improve our ability to predict an individual's income.

The independent variable's ability to model variations in Y can be evaluated by comparing the amount of deviation explained by our model using X to the total amount of deviation in Y. This ratio is known as the Coefficient of Determination or $R^2$ which represents the proportion of variation in Y explained by X. It can range from 0 to 1.

### Components of Deviation ($R^2$) \(y=\text{income}; \ x=\text{education}\)

The components of deviation for one observation are as follows:

\[
y_i - \bar{Y} = \text{the deviation of the Y observation from the mean of Y} \quad \text{(Total Dev.)}
\]

\[
y_i - \hat{Y} = \text{deviation explained by X} \quad \text{(Explained Dev.)}
\]

\[
y_i - \hat{y}_i = \text{deviation not explained by X} \quad \text{(Unexplained Dev.)}
\]
Example of Explained and Unexplained Deviations Using One Observation

\[ \bar{Y} = \text{mean income } $30.8k \]

\[ y_i = \text{Tracy’s income } $44k \]

\[ x_i = \text{Tracy’s education } 18 \text{ years} \]

\[ \hat{y} = \text{Tracy’s predicted income is } $40.9 \]

\[ \hat{Y} = -9.5 + 2.8(18) \]

Example of Deviations in a Two Dimensional Plot
The formula for estimating deviations for all observations is as follows:

**TSS (Total Sum of Squares)**

\[ \sum (Y_i - \bar{Y})^2 = \text{the total deviation of } Y \]

**RSS (Regression Explained Sum of Squares)**

\[ \sum (\hat{Y}_i - \bar{Y})^2 = \text{deviation explained by } X \]

**ESS (Error Sum of Squares)**

\[ \sum (Y_i - \hat{Y}_i)^2 = \text{deviation not explained by } X \]

\[ R^2 = \frac{RSS}{RSS + ESS} \]

or

\[ R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} \]
Example:

\[
R^2 = \frac{213.3}{213.3 + 32.2} \quad R^2 = .87
\]

**Impact of \( R^2 \) on predictions:**

A relatively high \( R^2 \) is required to make accurate predictions (.90 or better). It is very unlikely in social science that we will obtain \( R^2 \) this high, thus we focus more on explaining relationships.

**\( R^2 \) is sample specific:**

Two samples with the same variables, slope, and intercept could have different \( R^2 \) because of the fit between the data and the regression line (different variation in \( Y \); see formula).
**Software Output:** Regressing individual income on education from survey data.

---

**Simple OLS Regression**

<table>
<thead>
<tr>
<th>Cases(n)</th>
<th>R</th>
<th>R Square</th>
<th>Adj Rsq</th>
<th>SE Est.</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>0.751</td>
<td>0.564</td>
<td>0.550</td>
<td>2854.601 [c]</td>
</tr>
</tbody>
</table>

**ANOVA Statistics**

<table>
<thead>
<tr>
<th>Sum of Sqs</th>
<th>df</th>
<th>Mean Sq</th>
<th>F</th>
<th>p &lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>316481758.400</td>
<td>1</td>
<td>316481758.400</td>
<td>38.838</td>
</tr>
<tr>
<td>Residual</td>
<td>244462446.475</td>
<td>30</td>
<td>8148748.216</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>560944204.875</td>
<td>31</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Coefficients**

<table>
<thead>
<tr>
<th>Variable</th>
<th>b Coeff.</th>
<th>std. error</th>
<th>BetaWgt</th>
<th>t</th>
<th>p &lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>5077.512</td>
<td>1497.830</td>
<td>3.390</td>
<td>0.0020</td>
<td></td>
</tr>
<tr>
<td>EDUC</td>
<td>732.400</td>
<td>117.522</td>
<td>0.751</td>
<td>6.232</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

**Estimated Model**

\[ \text{INCOME} = 5077.513 + 732.400(\text{EDUC}) \]

**Interpretation**

[a] The correlation coefficient.

[b] Education explains 56% of the variation in income.

[c] Standard error of the estimate. Used for creating confidence intervals for predicted Y.

[d] The regression model is statistically significant.
[e] The Y intercept and mean value of income if education is equal to 0.

[f] The impact of a one-unit change in education on income. Interpretation: One additional year of education will on average result in an increase in income of $732.

[g] Standard error of the slope coefficient. Used for creating confidence intervals for b and for calculating a t-statistic for significance testing.

[h] Standardized regression partial slope coefficients. Used only in multiple regression models (more than one independent variable), beta-weights in the same regression model can be compared to one another to evaluate relative effect on the dependent variable. Beta-weights will indicate the direction of the relationship (positive or negative). Regardless of whether negative or positive, larger beta-weight absolute values indicate a stronger relationship.

[i] The t-test statistic. Calculated by dividing the b coefficient by the standard error of the slope.

[j] Education has a statistically significant positive association with income.
Multiple OLS Regression

Multiple regression is an extension of the simple regression model that allows the incorporation of more than one independent variable into the explanation of $Y$. Multiple regression helps clarify the relationship between each independent variable and $Y$ by holding constant the effects of the other independent variables in the model.

**General equation:**

$$Y = a_0 + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_kX_k + e$$

**Interpretation:**

The value of $Y$ is determined by a linear combination of the independent variables ($X_k$) plus an error term ($e$).

We continue to use the least squares approach for fitting the data based on the formula for the straight line; this as the least sum of the squared differences between observed $Y$ and predicted $Y$.

$$ESS = \sum(Y_i - \hat{Y}_i)^2$$

However, the fit cannot be visualized graphically on a two dimensional scattergram. The line is best visualized on three or more dimensional planes.
10.1 Partial slope equation

(Three variable example)

\[ Y = a_0 + b_1 X_1 + b_2 X_2 + e \]

Process:

1. Assume there is some correlation between the independent variables.

\[ r_{12} \neq 0 \]

2. Measure portion of \( X_1 \) not explained by \( X_2 \), which is a simple application of the bi-variate model through the following process:

   Use equation for a straight line to predict \( X_1 \) given \( X_2 \).

   \[ X_1 = c_1 + c_2 X_2 + u \quad \text{analogous to} \quad X_1 = a_0 + b_2 X_2 + e \]

   Compute the prediction error where \( u \) represents the portion of \( X_1 \) which \( X_2 \) cannot explain.

   \[ u = \hat{X}_1 - \tilde{X}_1 \]

2. Repeat steps to measure portion of \( Y \) which is not explained by \( X_2 \):

   Use equation for a straight line to predict \( Y \) given \( X_2 \).

   \[ Y = d_1 + d_2 X_2 + v \quad \text{analogous to} \quad Y = a_0 + b_2 X_2 + e \]

   Compute the prediction error where \( v \) represents the portion of \( Y \) which \( X_2 \) cannot explain.
3. The above is incorporated into the formula for $b_1$

$$b_1 = \frac{\sum (X_1 - \hat{X}_1)(Y - \hat{Y})}{\sum (X_1 - \hat{X}_1)^2}$$

or

$$b_1 = \frac{\sum (u)(v)}{\sum u^2}$$

This equation results in a measure of the slope for $X_1$ that is independent of the linear effect of $X_2$ on $X_1$ and $Y$. Put in more applied terms, a multiple regression model allows us to evaluate the spuriousness of relationships. As an example. We found in our bivariate model that education has a significant causal relationship with income. Given the simplicity of this model, we don't know if the relationship might disappear if we controlled for employee experience. In adding experience to our model, we face four possible outcomes:

1. Education is significant but experience is not (evidence bivariate relationship between education and income is not spurious for experience).

2. Experience is significant but education is not (evidence bivariate relationship between education and income is spurious).

3. Both education and experience are significant (evidence bivariate relationship between education and income is not spurious).

4. Neither are significant (evidence of high correlation between education and experience).
Partial slope example

<table>
<thead>
<tr>
<th>Name</th>
<th>Income ($1000s)</th>
<th>Education (Years)</th>
<th>Experience (Years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Susan</td>
<td>25</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>Bill</td>
<td>27</td>
<td>14</td>
<td>5</td>
</tr>
<tr>
<td>Bob</td>
<td>32</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>Tracy</td>
<td>44</td>
<td>18</td>
<td>8</td>
</tr>
<tr>
<td>Joan</td>
<td>26</td>
<td>12</td>
<td>10</td>
</tr>
</tbody>
</table>

Data

\[ Y = a_0 + b_1 X_1 + b_2 X_2 + e \]

Model Fit Statistics

<table>
<thead>
<tr>
<th>Cases(n)</th>
<th>R</th>
<th>R Square</th>
<th>Adj Rsq</th>
<th>SE Est.</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.974</td>
<td>0.948</td>
<td>0.896</td>
<td>2.538</td>
</tr>
</tbody>
</table>

ANOVA Statistics

<table>
<thead>
<tr>
<th>Sum of Sqs</th>
<th>df</th>
<th>Mean Sq</th>
<th>F</th>
<th>p &lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>233.915</td>
<td>2</td>
<td>116.957</td>
<td>18.153</td>
</tr>
<tr>
<td>Residual</td>
<td>12.885</td>
<td>2</td>
<td>6.443</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>246.800</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Coefficients

<table>
<thead>
<tr>
<th>Variable</th>
<th>b Coeff.</th>
<th>std. error</th>
<th>BetaWgt</th>
<th>t</th>
<th>p &lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-14.981</td>
<td>7.739</td>
<td>-1.936</td>
<td>0.1925</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>2.900</td>
<td>0.490</td>
<td>0.962</td>
<td>5.925</td>
<td>0.0273</td>
</tr>
<tr>
<td>Experience</td>
<td>0.692</td>
<td>0.400</td>
<td>0.281</td>
<td>1.732</td>
<td>0.2255</td>
</tr>
</tbody>
</table>

Equation after regression analysis:

\[ Y = -14.981 + 2.9(X_1) + 0.692(X_2) + e \]
Interpreting the Parameter Estimates

**Intercept** \((a_0)\): The average value of income when all independent variables are equal to "0" (-14.981 or -$14,981).

**Education** \((b_1)\): When we hold experience constant, a one unit change in education results in an average change in income of 2.9k ($2,900).

**Experience** \((b_2)\): When we hold education constant, a one unit change in experience results in an average change in income of .692k ($692).

Confidence Intervals for Slope Coefficients

This uses the same general procedure as bivariate (simple) regression model.

\[
CI = b_1 \pm t_{cv} \left( s_{b_1} \right)
\]

where \(t_{cv} = \text{critical value from the } t \text{ distribution}\)

- \(n = \text{sample size}\)
- \(k = \text{number of independent variables}\)

\[
df = n - k - 1 \quad \quad \quad \quad df = 5 - 2 - 1 \quad \quad \quad \quad df = 2
\]

- alpha .05, 2-tailed with 2 degrees of freedom
- \(t_{cv,} = 4.303\)
The equation for the 95% confidence interval for the education slope is the following:

\[ CI = 2.900 \pm 4.303(490) \]

This is interpreted as there is a probability of .95 that the slope of education in the population is between .800 and 5 ($800 and $5,000).

**Significance Tests (2-tailed)**

The significance test for the two slope coefficients uses the same degrees of freedom, sampling distribution, and critical value (tcv=4.303) as applied to the confidence interval calculation.

\[ t_{b1} = \frac{b_1}{s_{b1}} \quad t_{b1} = \frac{2.900}{.490} \quad t_{b1} = 5.918 \quad \text{significant; } t_{b1} > t_{cv} \text{ of 4.303} \]

\[ t_{b2} = \frac{b_2}{s_{b2}} \quad t_{b2} = \frac{.692}{.400} \quad t_{b2} = 1.73 \quad \text{(not significant; } t_{b2} < t_{cv} \text{ of 4.303)} \]

**10.2 Predicting Y**

Based on the following equation . . .

\[ \hat{Y} = -14.981 + 2.9(X_1) + .692(X_2) \]

and data for Jim who has 13 years of education and 10 years experience.

\[ \hat{Y} = -14.981 + 2.9(13) + .692(10) \quad \hat{Y} = -14.981 + 37.7 + 6.92 \quad \hat{Y} = 29.639 \]
Jim's predicted average income is $29,639 (this is the predicted point estimate for Jim).

**Confidence Interval for Predicted Y**

Again using the same degrees of freedom, sampling distribution, and critical value.

\[ CI = a_0 \pm t_{cv} \left( s_{a_0} \right) \]

\[
CI = 29.639 \pm 4.303(2.538)
\]

\[
CI = 29.639 \pm 10.92
\]

There is a probability of .95 that someone in the population with 13 years of education and 10 years experience would earn between $18,720 and $40,560. Obviously this model would not be very useful for predicting incomes in the population. This is not surprising given the very small sample size (n=5).

**10.3 Adjusted R²**

Adjusted R² is used to compensate for the addition of variables to the model. As more independent variables are added to the regression model, unadjusted R² will generally increase (there will never be a decrease). This will occur even when the additional variables do little to help explain the dependent variable. To compensate for this, adjusted R² is corrected for the number of independent variables in the model. The result is an adjusted R² than can go up or down depending on whether the addition of another variable adds or does not add to the explanatory power of the model. Adjusted R² will always be lower than unadjusted.
It has become standard practice to report the adjusted $R^2$, especially when there are multiple models presented with varying numbers of independent variables.

*Adjusted $R^2$ formula*

\[
\overline{R^2} = \left( R^2 - \frac{k}{n-1} \right) \left( \frac{n-1}{n-k-1} \right)
\]

\[
\overline{R^2} = \left( .948 - \frac{2}{5-1} \right) \left( \frac{5-1}{5-2-1} \right)
\]

\[
\overline{R^2} = .896
\]

**10.4 Standardized Partial Slope Coefficients (beta weights)**

Partial slope coefficients are not directly comparable to one another. This is because each variable is likely based on a different metric (age, income, sex). Beta weights (standardized slope coefficients) are used to compare the relative importance of regression coefficients in a model. They are not useful for comparisons among models from other samples.

\[
\beta_i = b_i \left( \frac{s_{x_i}}{s_y} \right)
\]

Formula: $\beta_i$ or partial slope * (std dev. of $X_i$ ÷ std dev. of $Y$)

**Example:**
Education partial slope coefficient \( (b_i) = 1878.2 \)

Education standard deviation \( s_{x_i} = 2.88 \)

Begin salary standard deviation \( s_y = 7870.64 \)

\[
\beta_i = b_i \left( \frac{s_{x_i}}{s_y} \right) \quad \beta_i = 1878.2 \left( \frac{2.88}{7870.64} \right) \quad \beta_i = .687
\]

This is interpreted as the average standard deviation change in Y (salary) associated with one standard deviation change in X (education), when all other independent variables are held constant. Or a one standard deviation change in education results in an average change of .687 standard deviations in income.

### 10.5 F-statistic

The F-statistic represents a hypothesis test to determine if the independent variables taken as a whole help explain the dependent variable. This is a test of the model, not the partial regression coefficients.

I. Assumptions

See regression assumptions

II. State the Hypothesis

Ho: The regression coefficients taken together simultaneously are equal to zero.

\[
H_0 = \beta_1 = \beta_2 = \beta_k = 0
\]
Ha: The regression coefficients taken together simultaneously are equal to zero.

\[ H_0 \neq \beta_1 \neq \beta_2 \neq \beta_k \neq 0 \]

III. Set the Rejection Criteria

*where*

- degrees of freedom for the numerator = \( k \)
- degrees of freedom for the denominator = \( n-k-1 \)

At alpha .05, DF = (2, 471)

**F distribution** critical value = 3.02

IV. Compute Test Statistic

*Formula (A)*: This shows that the F-ratio is related to the explanatory power of the entire regression equation

\[
F = \left( \frac{R^2}{1-R^2} \right) \left( \frac{n-k-1}{k} \right)
\]

*Formula (B)*: This formula is more relevant for common regression output.

\[
F = \frac{\sum \left( \hat{Y}_i - \bar{Y} \right)^2 / k}{\sum \left( Y_i - \hat{Y}_i \right)^2 / (n-k-1)}
\]

The mean sum of squares for what the model explains divided by the mean sum of squares for what the model does not explain.
Example:

Calculate F-test statistic -- Formula (A) and Formula (B)

\[ F = \frac{R^2 (n - k - 1)}{(1 - R^2) k} \]

\[ F = \left( \frac{.446}{1 - .446} \right) \left( \frac{471}{2} \right) \]

\[ F = 189.59 \]

or

\[ F = \frac{\sum (Y_i - \bar{Y})^2 / k}{\sum (Y_i - \hat{Y}_i)^2 / (n - k - 1)} \]

\[ F = \frac{1306195553 / 2}{16238949412 / 471} \]

\[ F = 189.43 \]

V. Decision

Reject null hypothesis. The model of independent variables is significantly related to the dependent variable.

Note: It is possible to have independent variables that individually are not statistically significant but as a group do have a significant relationship with the dependent variable. Lack of significance for each independent variable may be due to high correlation with other independent variables in the model (multicolinearity).

10.6 Dummy Variables

It is possible to include nominal and ordinal level independent variables in a regression model by dichotomous coding of variables into integers 0 and 1. These are commonly called dummy variables. The interpretation of a dummy variable is similar to other independent variables except as follows:
The partial slope coefficient represents the on average change in the dependent variable between the coded value of 1 for the independent variable and the reference group of 0.

Creating Dummy Variables

Always use k-1 dummy variables in a regression model, where k is equal to the number of categories in the variable. In the following example, a dummy variable called “Male” is created where 1=Male and 0=Female, the reference group. There is no need to create a dummy variable for Female (k=2; or, K-1 = 1 dummy variable).

Example A: Creating a dummy variable from a two-category variable

Original Variable

<table>
<thead>
<tr>
<th>Gender</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Percent</td>
<td></td>
</tr>
<tr>
<td>Valid</td>
<td>Female</td>
<td>216</td>
<td>45.6</td>
<td>45.6</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>258</td>
<td>54.4</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>474</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Male Dummy: recoded into 0/1 variable (0=Female and 1=Male)

<table>
<thead>
<tr>
<th>MALE</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Percent</td>
<td></td>
</tr>
<tr>
<td>Valid</td>
<td>0.00</td>
<td>216</td>
<td>45.6</td>
<td>45.6</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>258</td>
<td>54.4</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>474</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>
Example B: Creating a dummy variable from a three-category variable

For a three-category variable, two dummy variables are created to represent three categories.

Original Variable

<table>
<thead>
<tr>
<th>Employment Category</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clerical</td>
<td>363</td>
<td>76.6</td>
<td>76.6</td>
<td>76.6</td>
</tr>
<tr>
<td>Custodial</td>
<td>27</td>
<td>5.7</td>
<td>5.7</td>
<td>82.3</td>
</tr>
<tr>
<td>Manager</td>
<td>84</td>
<td>17.7</td>
<td>17.7</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>474</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Clerical Dummy: recoded into 0/1 variable (0=Custodial and Manager; 1=Clerical)

<table>
<thead>
<tr>
<th>Clerical Employee</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.00</td>
<td>111</td>
<td>23.4</td>
<td>23.4</td>
<td>23.4</td>
</tr>
<tr>
<td>1.00</td>
<td>363</td>
<td>76.6</td>
<td>76.6</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>474</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Custodian Dummy: recoded into 0/1 variable (0=Clerical and Manager; 1=Custodian)

<table>
<thead>
<tr>
<th>Custodian Employee</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.00</td>
<td>447</td>
<td>94.3</td>
<td>94.3</td>
<td>94.3</td>
</tr>
<tr>
<td>1.00</td>
<td>27</td>
<td>5.7</td>
<td>5.7</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>474</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Managers are the reference group, represented in the model when the clerical and custodian dummy variables are both equal to zero in the data.
**Software Output**: A multiple regression example regressing individual income on education in years, work experience in months, and a dummy variable representing sex (0=Female, 1=Male).

---

### Multiple OLS Regression

**Explanatory Model**

\[
\text{INCOME} = \text{Constant} + \text{EDUC} + \text{WORKEXP} + \text{SEX} + e
\]

**Model Fit Statistics**

<table>
<thead>
<tr>
<th>Cases(n)</th>
<th>R</th>
<th>R Square</th>
<th>Adj Rsq</th>
<th>SE Est.</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>0.865</td>
<td>0.748</td>
<td>0.721 [a]</td>
<td>2246.736 [b]</td>
</tr>
</tbody>
</table>

**ANOVA Statistics**

<table>
<thead>
<tr>
<th>Sum of Sqs</th>
<th>df</th>
<th>Mean Sq</th>
<th>F</th>
<th>p &lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>419605192.680</td>
<td>3</td>
<td>139868397.560</td>
<td>27.709</td>
</tr>
<tr>
<td>Residual</td>
<td>141339012.195</td>
<td>28</td>
<td>5047821.864</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>560944204.875</td>
<td>31</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Coefficients**

<table>
<thead>
<tr>
<th>Variable</th>
<th>b Coeff.</th>
<th>std. error</th>
<th>BetaWgt</th>
<th>t</th>
<th>p &lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-338.10507</td>
<td>2027.91657</td>
<td>-0.167</td>
<td>-0.167</td>
<td>0.8688</td>
</tr>
<tr>
<td>EDUC</td>
<td>870.14643 [d]</td>
<td>108.50630</td>
<td>0.89240 [e]</td>
<td>8.019</td>
<td>0.0000 [f]</td>
</tr>
<tr>
<td>WORKEXP</td>
<td>193.70166 [g]</td>
<td>82.88935</td>
<td>0.26208</td>
<td>2.337</td>
<td>0.0268</td>
</tr>
<tr>
<td>SEX</td>
<td>2821.21837 [h]</td>
<td>803.62244</td>
<td>0.33626</td>
<td>3.511</td>
<td>0.0015</td>
</tr>
</tbody>
</table>

**Estimated Model**

\[
\text{INCOME} = -338.105 + 870.146(\text{EDUC}) + 193.702(\text{WORKEXP}) + 2821.218(\text{SEX})
\]

**Interpretation**

[a] Adjusted R². Education, experience, and a person’s sex explain 72% of the variation in individual income.
[b] Standard error of the estimate. Used for creating confidence intervals for predicted Y.

c] The regression model is statistically significant.

d] The impact of a one-unit change in education on income. One additional year of education will on average result in an increase in income of $870 holding experience and sex constant.

e] Standardized regression partial slope coefficients. Used only in multiple regression models (more than one independent variable), beta-weights in the same regression model can be compared to one another to evaluate relative effect on the dependent variable. Beta-weights will indicate the direction of the relationship (positive or negative). Regardless of whether negative or positive, larger beta-weight absolute values indicate a stronger relationship. In this model, education has the greatest effect on income followed in order by sex and experience.

h] Education has a statistically significant association with income. The probability of a type one error is less than .0001 which far exceeds alpha of p = .05. Also, experience and sex are statistically significant.

i] One additional month of experience will on average result in an increase in income of $194 holding education and sex constant.

j] Males will on average earn an income of $2,821 more than women holding education and experience constant.
11

Regression Assumptions

Although ordinary least squares multiple regression is a very robust statistical technique, careful consideration should be given to the assumptions underlying obtaining the Best Linear Unbiased Estimates (BLUE) from a regression model. In practice, it is very difficult to completely satisfy all the assumptions but there are techniques to help adjust the affects on the model. A few suggested corrections are provided in the following brief review.

1. **No measurement error**

The independent (X) and dependent (Y) variables are accurately measured. Non-random error (e.g., measuring the wrong income for a defined group of people) will seriously bias the model. Random error is the more common assumption violation. Random error in an independent variable may lower $R^2$ and the partial slope coefficients can vary dramatically depending on the amount of error. Other independent variables that do not have random measurement error will be biased if they are correlated with another independent variable with measurement error. Error in measuring the dependent variable many not bias the estimates if the error is random.

2. **No specification error**

The theoretical model is assumed to be a) linear, b) additive, and c) includes the correct variables.

**Non-Linearity**
Linear implies that the average change in the dependent variable associated with a one unit change in an independent variable is constant regardless of the level of the independent variable. If the partial slope for $X$ is not constant for differing values of $X$, $X$ has a nonlinear relationship with $Y$ and results in biased partial slopes. There are two basic types of nonlinearity discussed in the following.

*Not Linear but slope does not change direction.*

In this example the variable’s slope is positive but the steepness of the curve increases as the value of $X$ increases.

Correction: A log-log model takes a nonlinear specification where the slope changes as the value of $X$ increases and makes it linear in terms of interpreting the parameter estimates. It accomplishes this by transforming the dependent and all independent variables by taking their log and replacing the original variables with the logged variables. The result is coefficients that are interpreted as a % change in $Y$ given a 1% change in $X$.

**Example:**

\[
\log Y = a + b_1 \log X_1 + b_2 \log X_2 + b_3 \log X_3 + \log e
\]

Note: All data must have positive values. A log of a negative value will equal 0 and as a result biases the model. After modeling, the anti log of the coefficients can be used to estimate $y$. 
**Not linear and the slope changes direction (pos or neg)**

In this example the variable’s slope changes direction depending on the value of X.

Correction: A polynomial model may be used to correct for changes in the slope coefficient sign (positive or negative): This is accomplished by adding additional variables that are incremental powers of the independent variable to model bends in the slope.

**Example:**

\[ Y = a + b_1X_1 + b_2X_2^2 + b_3X_3^3 + e \]

**Non-Additivity**

Additive implies that the average change in the dependent variable associated with a one unit change in an independent variable \((X_1)\) is constant regardless of the value of another independent variable \((X_2)\) in the model. If this assumption is violated, we can no longer interpret the slope by saying "holding other variables constant" since the values of the other variables may possibly change the slope coefficient and therefore its interpretation.
Nonadditive Relationship between X1 & X2

The figure above displays a non-additive relationship when (X1) is interval/ratio and (X2) is a dummy variable. If the partial slope for (X1) is not constant for differing values of (X2), (X1) and (X2) do not have an additive relationship with Y.

**Correction**: An interaction term may be added to the model using a dummy variable where the slope of X1 is thought to depend on the value of a dummy variable X2. The final model will look like the following:

**Model**: \[ Y = a + b_1 X_1 + b_2 X_1 X_2 + e \]

**where** \( X_1 X_2 = \)the interaction between X1 and X2 or \( X_1 * X_2 \)

**Interpretation**:

- \( b_1 \) is interpreted as the slope for X1 when the dummy variable (X2) is 0
- \( b_1 + b_2 \) is interpreted as the slope for X1 when the dummy variable (X2) is 1

**Correction for two interval independent variables**:
Model without interaction term  
\[ Y = a + b_1 X_1 + b_2 X_2 + e \]

Model with interaction term  
\[ Y = a + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2 + e \]

where \( b_3 = X_1 X_2 \) (the interactive term)

**Incorrect Independent Variables**

Including the correct independent variables implies that an irrelevant variable has not been included in the model and all theoretically important relevant variables are included. Failing to include the correct variables in the model will bias the slope coefficients and may increase the likelihood of improperly finding statistical significance. Including irrelevant variables will make it more difficult to find statistical significance.

*Correction:* Remove irrelevant variables and, if possible, include missing relevant variables.

**3. Mean of errors equals zero**

When the mean error (reflected in residuals) is not equal to zero, the y intercept may be biased. Violation of this assumption will not affect the slope coefficients. The partial slope coefficients will remain Best Unbiased Linear Estimates (BLUE).

**4. Error term is normally distributed**

The distribution of the error term closely reflects the distribution of the dependent variable. If the dependent variable is not normally distributed the error term may not be normally distributed. Violation of this assumption will not bias the partial slope coefficients but may affect significance tests.
Correction: Always correct other problems first and then re-evaluate the residuals

★ If the distribution of residuals is skewed to the right (higher values), try using the natural log of the dependent variable

★ If the distribution of residuals is skewed to the left (lower values), try squaring the dependent variable.

Always check the distribution of the residuals again after trying a correction.

5. Homoskedasticity

The variance of the error term is constant (homoskedastic) for all values of the independent variables. Heteroskedasticity occurs when the variance of the error term does not have constant variance. The parameter estimates for partial slopes and the intercept are not biased if this assumption is violated; however, the standard errors are biased and hence significance tests may not be valid.

Heteroskedasticity: Error gets larger as X increases

Diagnosis Of Heteroskedasticity
Plot the regression residuals against the values of the independent variable(s). If there appears an even pattern about a horizontal axis, heteroskedasticity is unlikely.

For small samples there may be some tapering at each end of the horizontal distribution.
If there is a cone or bow tie shaped pattern, heteroskedasticity is suspected.

**Correction**: If an excluded independent variable is suspected, including this variable in the model may correct the problem. Otherwise, it may be necessary to use generalized least squares (GLS) or weighted least squares (WLS) models to create coefficients that are BLUE.

6. **No autocorrelation**

No autocorrelation assumes that the error terms are not correlated across observations. Violation of this assumption is likely to be a problem with time-series data where the value of one observation is not completely independent of another observation. (Example: A simple two-year time series of the same individuals is likely to find that a person's income in year 2 is correlated with their income in the prior year.) If there is autocorrelation, the parameter estimates for the partial slopes and the intercept are not biased but the standard errors are biased and hence significance tests may not be valid.

**Diagnosis**

Suspect autocorrelation with any time series/longitudinal data
Use the Durbin-Watson (d) statistic

★ d=2 > no correlation between error terms
★ d=0 > perfect positive correlation between error terms
★ d=4 > perfect negative correlation between error terms

Correction: Use generalized least squares (GLS) or weighted least squares (WLS) models to create coefficients that are BLUE.

7. **No Multicolinearity**

The assumption of no multicolinearity is an issue only for multiple regression models. Multicolinearity occurs when one of the independent variables has a substantial linear relationship with another independent variable in the equation. It occurs to some extent in any model and is more a matter of degree of colinearity rather than whether it exists or not. Multicolinearity will result in variability in the partial slope coefficients from one sample to the next or when models are changed slightly. In addition, standard errors are increased which reduces the likelihood of finding statistical significance. The result of the above two situations is an unbiased estimator that is very inefficient.

*Diagnosis*

★ Failing to find any variables statistically significant yet the F-statistic shows the model is significant.

★ Dramatic changes in coefficients as independent variables are added or deleted from the model.
Examine covariation among the independent variables by calculating all possible bivariate combinations of the Pearson correlation coefficient. Generally a high correlation coefficient (say .80 or greater) suggests a problem. This is imperfect since multicollinearity may not be reflected in a bivariate correlation matrix.

Regress each independent variable on the other independent variables. If any of the $R^2$s are near 1.0 (this indicates one independent variable is almost entirely explained by the other independent variables in the model), there is a high degree of multicollinearity.

**Correction:**

- Increase sample size to lower standard errors. This doesn’t always work and is normally not feasible since adding more cases is not a simple exercise in most studies.
- Combine two or more variables that are highly correlated into a single indicator of an abstract concept.
- Delete one of the variables that are highly correlated. This may result in a poorly specified model.
- Leave all the variables in model and rely on the joint hypothesis F-test to evaluate the significance of the model. This is especially useful if multicollinearity is causing most if not all of the independent variables to be not significant.

**Summary**

As noted at the beginning of this section, it is very difficult to completely satisfy all of the assumptions required in regression modeling. The best starting point in designing an adequate model is theory. Theory should drive what variables are included in
the regression model and it will also help inform the researcher on the possible inter-
actions among the independent variables. A model that does not perform as antici-
pated by theory suggests either the theory was wrong or that the regression esti-
mates and significance tests are being biased by one or more assumption violations.

The following page has a brief summary of the assumption violations, affect on the
model, and possible corrective action.
# BLUE Check List: Diagnosing and correcting violations of regression assumptions

<table>
<thead>
<tr>
<th>Assumption Violation</th>
<th>Consequences</th>
<th>Diagnosis</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random error in DV</td>
<td>+</td>
<td>data errors; unreliable measure</td>
<td>clean errors find better measure</td>
</tr>
<tr>
<td>Random error in IV</td>
<td>B</td>
<td>data errors; unreliable measure</td>
<td>clean errors find better measure</td>
</tr>
<tr>
<td>Nonlinear model</td>
<td>B</td>
<td>review theoretical relationships; transform variable(s)</td>
<td>(transformation will depend on type of nonlinear relationship)</td>
</tr>
<tr>
<td>(slope not constant for all X1 values)</td>
<td></td>
<td>break X1 into 2+ dummy variables and evaluate consistency of slopes</td>
<td></td>
</tr>
<tr>
<td>Nonadditive model</td>
<td>B</td>
<td>review theoretical relationships; create interaction term</td>
<td>(X1*X2)</td>
</tr>
<tr>
<td>(slope for X1 not constant for values of X2)</td>
<td></td>
<td>break X2 into 2+ dummy variables and evaluate consistency of slopes</td>
<td></td>
</tr>
<tr>
<td>Irrelevant IV in model</td>
<td>+</td>
<td>coefficient is unlikely to be statistically significant; removal won’t lower Adj R²</td>
<td>remove</td>
</tr>
<tr>
<td>Relevant IV excluded</td>
<td>B</td>
<td>theory; adding should increase Adj R²</td>
<td>add to model</td>
</tr>
<tr>
<td>Residuals not normally distributed</td>
<td>B</td>
<td>plot histogram of residuals</td>
<td>positive skew &gt; log DV negative skew &gt; sqr DV</td>
</tr>
<tr>
<td>Homoskedasticity</td>
<td>B</td>
<td>plot residuals against each IV</td>
<td>try including additional IV</td>
</tr>
<tr>
<td>(variance not constant for all values of Xi)</td>
<td></td>
<td>(funnel, cone, or bow tie shape indicative of heteroskedasticity)</td>
<td>use GLS or WLS models</td>
</tr>
<tr>
<td>Autocorrelation</td>
<td>B</td>
<td>always suspect when using time series data; check Durbin-Watson statistic</td>
<td>use GLS or WLS models</td>
</tr>
<tr>
<td>(error terms correlated across observations)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multicolinearity</td>
<td>Erratic</td>
<td>no IV significant; dramatic coefficient change when IV are added or deleted; check correlation among IV; regress each IV on other IV; R² near 1.0 suggests multicolinearity</td>
<td>increase sample size; combine highly correlated IVs into a single measure; delete one of the correlated IVs; use F-test to evaluate model signif.</td>
</tr>
</tbody>
</table>

B=Biased; ‘+’ =Increases; ‘-’ =Decreases; b=slope coefficient; Se=Standard Error; DV=Dependent Variable; IV=Independent Variable
Logistic Regression

Logistic regression is used for multiple regression models where the dependent variable is dichotomous (0 or 1 values) or ordinal if conducting multinomial logistic regression. By convention, the dependent variable should be coded 0 and 1 where 1 represents the value of greatest interest.

Assumptions:

See multiple regression for a more complete list of ordinary least squares (OLS) regression assumptions. In addition to the key characteristic that the dependent variable is discrete rather than continuous, the following assumptions for logistic regression differ from those for OLS regression:

1. Specification error - Linearity

Linearity is not required in logistic regression between the independent and dependent variables. It does require that the logits of the independent and dependent variables are linear. If not linear, the model may not find statistical significance when it actually exists (Type II error).

2. Large Sample Size

Unlike OLS regression, large sample size is needed for logistic regression. If there is difficulty in converging on a solution, a large number of iterations, or very large regression coefficients, there may be insufficient sample size.
Software Output: Evaluating the affect of age, education, sex, and race on whether or not a person voted in the presidential election.

------------------------ Logistic Regression -------------------

Dependent Y: VOTE

Cases with Y=0: 417 28.86% No, did not vote
Cases with Y=1: 1028 71.14% Yes, voted
Total Cases: 1445

Explanatory Model

VOTE92 = Constant + AGE + EDUC + FEMALE + WHITE + e

Model Fit Statistics

Initial -2 Log Likelihood 1736.533
Model -2 Log Likelihood 1544.154 Iteration (4)
Cox & Snell Rsq 0.125 [a]
McFadden Rsq 0.111 [b]
Model Chi-Square 192.380 df 4 p< 0.0000 [c]

Coefficients

<table>
<thead>
<tr>
<th>Variable</th>
<th>b Coeff.</th>
<th>SE</th>
<th>Wald</th>
<th>p</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-4.1620</td>
<td>0.4184</td>
<td>98.9755</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td>0.0335 [d]</td>
<td>0.0040</td>
<td>69.1386 [e]</td>
<td>0.0000</td>
<td>1.0340 [g]</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.2809</td>
<td>0.0244</td>
<td>132.4730</td>
<td>0.0000</td>
<td>1.3243</td>
</tr>
<tr>
<td>FEMALE</td>
<td>-0.1249</td>
<td>0.1269</td>
<td>0.9686</td>
<td>0.3250</td>
<td>0.8826</td>
</tr>
<tr>
<td>WHITE</td>
<td>0.1000</td>
<td>0.1671</td>
<td>0.3581</td>
<td>0.5495</td>
<td>1.1052</td>
</tr>
</tbody>
</table>

Estimated Model

VOTE92 = -4.1620 + .0335(AGE) + .2809(EDUC) + -.1249(FEMALE) + .1000(WHITE)

Classification Table (.50 cutpoint)

<table>
<thead>
<tr>
<th>Observed Y</th>
<th>Predicted Y</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>103</td>
<td>314</td>
</tr>
<tr>
<td>1</td>
<td>64</td>
<td>964</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Interpretation

[a] This ranges from 0 to less than 1 and indicates the explanatory value of the model. The closer the value is to 1 the greater the explanatory value. Analogous to $R^2$ in OLS regression.

[b] Indicates explanatory value of the model. Ranges from 0 to 1 and adjusts for the number of independent variables. Analogous to Adjusted $R^2$ in OLS regression. In this model, the explanatory value is relatively low.

[c] Tests the combined significance of the model in explaining the dependent (response) variable. The significance test is based on the Chi-Square sampling distribution. Analogous to the F-test in OLS regression. The independent variables in this model have a statistically significant combined effect on voting behavior.

[d] A one unit increase in age (1 year) results in an average increase of .0335 in the log odds of vote equaling 1 (voted in election).

[e] The test statistic for the null hypothesis that the $b$ coefficient is equal to 0 (no effect). It is calculated by $(B/SE)^2$.

[f] Based on the Chi-Square sampling distribution, there is a statistically significant relationship between age and voting behavior. The probability voting increases with age, holding education, sex, and race constant.

[g] OR=Odds Ratio. The odds of a change in the dependent variable (vote) given a one unit change in age is 1.03. Note: if the OR is >1, odds increase; if OR <1, odds decrease; if OR =1, odds are unchanged by this variable. In this example, the odds of voting increases with a one unit increase in age.

[h] Of those who did not vote, the model correctly predicted non-voting in 24.7% of the non-voters included in the sample data.

[i] Of those who did vote, the model correctly predicted voting in 93.8% of the voters included in the sample data.

[j] Of those who did or did not vote, the model correctly predicted 73.8% of the voting decisions.
13

Distribution Tables

Critical Values for Z Distributions

Critical Values for T Distributions

Critical Values for Chi-square Distributions

Critical Values for F Distributions
### Z Distribution Critical Values

<table>
<thead>
<tr>
<th>Z</th>
<th>Area Beyond</th>
<th>Z</th>
<th>Area Beyond</th>
<th>Z</th>
<th>Area Beyond</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.150</td>
<td>0.440</td>
<td>1.300</td>
<td>0.097</td>
<td>2.450</td>
<td>0.007</td>
</tr>
<tr>
<td>0.200</td>
<td>0.421</td>
<td>1.350</td>
<td>0.089</td>
<td>2.500</td>
<td>0.006</td>
</tr>
<tr>
<td>0.250</td>
<td>0.401</td>
<td>1.400</td>
<td>0.081</td>
<td>2.550</td>
<td>0.005</td>
</tr>
<tr>
<td>0.300</td>
<td>0.382</td>
<td>1.450</td>
<td>0.074</td>
<td>2.580</td>
<td>0.005</td>
</tr>
<tr>
<td>0.350</td>
<td>0.363</td>
<td>1.500</td>
<td>0.067</td>
<td>2.650</td>
<td>0.004</td>
</tr>
<tr>
<td>0.400</td>
<td>0.345</td>
<td>1.550</td>
<td>0.061</td>
<td>2.700</td>
<td>0.004</td>
</tr>
<tr>
<td>0.450</td>
<td>0.326</td>
<td>1.600</td>
<td>0.055</td>
<td>2.750</td>
<td>0.003</td>
</tr>
<tr>
<td>0.500</td>
<td>0.309</td>
<td>1.650</td>
<td>0.050</td>
<td>2.800</td>
<td>0.003</td>
</tr>
<tr>
<td>0.550</td>
<td>0.291</td>
<td>1.700</td>
<td>0.045</td>
<td>2.850</td>
<td>0.002</td>
</tr>
<tr>
<td>0.600</td>
<td>0.274</td>
<td>1.750</td>
<td>0.040</td>
<td>2.900</td>
<td>0.002</td>
</tr>
<tr>
<td>0.650</td>
<td>0.258</td>
<td>1.800</td>
<td>0.036</td>
<td>2.950</td>
<td>0.002</td>
</tr>
<tr>
<td>0.700</td>
<td>0.242</td>
<td>1.850</td>
<td>0.032</td>
<td>3.000</td>
<td>0.001</td>
</tr>
<tr>
<td>0.750</td>
<td>0.227</td>
<td>1.900</td>
<td>0.029</td>
<td>3.050</td>
<td>0.001</td>
</tr>
<tr>
<td>0.800</td>
<td>0.212</td>
<td>1.960</td>
<td>0.025</td>
<td>3.100</td>
<td>0.001</td>
</tr>
<tr>
<td>0.850</td>
<td>0.198</td>
<td>2.000</td>
<td>0.023</td>
<td>3.150</td>
<td>0.001</td>
</tr>
<tr>
<td>0.900</td>
<td>0.184</td>
<td>2.050</td>
<td>0.020</td>
<td>3.200</td>
<td>0.001</td>
</tr>
<tr>
<td>0.950</td>
<td>0.171</td>
<td>2.100</td>
<td>0.018</td>
<td>3.250</td>
<td>0.001</td>
</tr>
<tr>
<td>1.000</td>
<td>0.159</td>
<td>2.150</td>
<td>0.016</td>
<td>3.300</td>
<td>0.001</td>
</tr>
<tr>
<td>1.050</td>
<td>0.147</td>
<td>2.200</td>
<td>0.014</td>
<td>3.350</td>
<td>0.000</td>
</tr>
<tr>
<td>1.100</td>
<td>0.136</td>
<td>2.250</td>
<td>0.012</td>
<td>3.400</td>
<td>0.000</td>
</tr>
<tr>
<td>1.150</td>
<td>0.125</td>
<td>2.300</td>
<td>0.011</td>
<td>3.450</td>
<td>0.000</td>
</tr>
<tr>
<td>1.200</td>
<td>0.115</td>
<td>2.350</td>
<td>0.009</td>
<td>3.500</td>
<td>0.000</td>
</tr>
</tbody>
</table>

This is only a portion of a much larger table.

The area beyond Z is the proportion of the distribution beyond the critical value (region of rejection). Example: If a test at .05 alpha is conducted, the area beyond Z for a two-tailed test is .025 (Z-value 1.96). For a one-tailed test the area beyond Z would be .05 (Z-value 1.65).
### T Distribution Critical Values (2-Tailed)

<table>
<thead>
<tr>
<th>DF</th>
<th>0.10</th>
<th>0.05</th>
<th>0.01</th>
<th>DF</th>
<th>0.10</th>
<th>0.05</th>
<th>0.01</th>
<th>DF</th>
<th>0.10</th>
<th>0.05</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.314</td>
<td>12.706</td>
<td>63.657</td>
<td>10</td>
<td>1.612</td>
<td>2.228</td>
<td>3.169</td>
<td>19</td>
<td>1.729</td>
<td>2.093</td>
<td>2.861</td>
</tr>
<tr>
<td>2</td>
<td>2.920</td>
<td>4.303</td>
<td>9.291</td>
<td>11</td>
<td>1.796</td>
<td>2.201</td>
<td>3.106</td>
<td>20</td>
<td>1.725</td>
<td>2.086</td>
<td>2.845</td>
</tr>
<tr>
<td>3</td>
<td>2.353</td>
<td>3.182</td>
<td>5.841</td>
<td>12</td>
<td>1.782</td>
<td>2.179</td>
<td>3.055</td>
<td>21</td>
<td>1.721</td>
<td>2.080</td>
<td>2.831</td>
</tr>
<tr>
<td>4</td>
<td>2.132</td>
<td>2.776</td>
<td>4.604</td>
<td>13</td>
<td>1.771</td>
<td>2.160</td>
<td>3.012</td>
<td>22</td>
<td>1.717</td>
<td>2.074</td>
<td>2.819</td>
</tr>
<tr>
<td>5</td>
<td>2.015</td>
<td>2.571</td>
<td>4.032</td>
<td>14</td>
<td>1.761</td>
<td>2.145</td>
<td>2.977</td>
<td>23</td>
<td>1.714</td>
<td>2.069</td>
<td>2.807</td>
</tr>
<tr>
<td>6</td>
<td>1.943</td>
<td>2.447</td>
<td>3.707</td>
<td>15</td>
<td>1.753</td>
<td>2.131</td>
<td>2.947</td>
<td>24</td>
<td>1.711</td>
<td>2.064</td>
<td>2.797</td>
</tr>
<tr>
<td>7</td>
<td>1.895</td>
<td>2.365</td>
<td>3.499</td>
<td>16</td>
<td>1.746</td>
<td>2.120</td>
<td>2.921</td>
<td>25</td>
<td>1.708</td>
<td>2.060</td>
<td>2.877</td>
</tr>
<tr>
<td>8</td>
<td>1.860</td>
<td>2.306</td>
<td>3.355</td>
<td>17</td>
<td>1.740</td>
<td>2.110</td>
<td>2.898</td>
<td>26</td>
<td>1.706</td>
<td>2.056</td>
<td>2.779</td>
</tr>
<tr>
<td>9</td>
<td>1.833</td>
<td>2.262</td>
<td>3.250</td>
<td>18</td>
<td>1.734</td>
<td>2.101</td>
<td>2.878</td>
<td>27</td>
<td>1.703</td>
<td>2.052</td>
<td>2.771</td>
</tr>
<tr>
<td>10</td>
<td>1.812</td>
<td>2.228</td>
<td>3.169</td>
<td>19</td>
<td>1.729</td>
<td>2.093</td>
<td>2.861</td>
<td>28</td>
<td>1.701</td>
<td>2.048</td>
<td>2.763</td>
</tr>
<tr>
<td>11</td>
<td>1.796</td>
<td>2.201</td>
<td>3.106</td>
<td>20</td>
<td>1.725</td>
<td>2.086</td>
<td>2.845</td>
<td>29</td>
<td>1.699</td>
<td>2.045</td>
<td>2.756</td>
</tr>
<tr>
<td>12</td>
<td>1.782</td>
<td>2.179</td>
<td>3.055</td>
<td>30</td>
<td>1.697</td>
<td>2.042</td>
<td>2.750</td>
<td>31</td>
<td>1.696</td>
<td>2.040</td>
<td>2.744</td>
</tr>
<tr>
<td>13</td>
<td>1.771</td>
<td>2.160</td>
<td>3.012</td>
<td>32</td>
<td>1.694</td>
<td>2.037</td>
<td>2.738</td>
<td>33</td>
<td>1.692</td>
<td>2.035</td>
<td>2.733</td>
</tr>
<tr>
<td>14</td>
<td>1.761</td>
<td>2.145</td>
<td>2.977</td>
<td>34</td>
<td>1.691</td>
<td>2.032</td>
<td>2.728</td>
<td>35</td>
<td>1.690</td>
<td>2.030</td>
<td>2.724</td>
</tr>
<tr>
<td>15</td>
<td>1.753</td>
<td>2.131</td>
<td>2.947</td>
<td>36</td>
<td>1.688</td>
<td>2.028</td>
<td>2.719</td>
<td>37</td>
<td>1.687</td>
<td>2.026</td>
<td>2.715</td>
</tr>
<tr>
<td>16</td>
<td>1.746</td>
<td>2.120</td>
<td>2.921</td>
<td>38</td>
<td>1.686</td>
<td>2.024</td>
<td>2.712</td>
<td>39</td>
<td>1.685</td>
<td>2.023</td>
<td>2.708</td>
</tr>
<tr>
<td>17</td>
<td>1.740</td>
<td>2.110</td>
<td>2.898</td>
<td>40</td>
<td>1.684</td>
<td>2.021</td>
<td>2.704</td>
<td>41</td>
<td>1.683</td>
<td>2.020</td>
<td>2.701</td>
</tr>
</tbody>
</table>

This is only a portion of a much larger table.
### Chi-square Distribution Critical Values

<table>
<thead>
<tr>
<th>df</th>
<th>.10</th>
<th>.05</th>
<th>.02</th>
<th>.01</th>
<th>.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.706</td>
<td>3.841</td>
<td>5.412</td>
<td>6.635</td>
<td>10.827</td>
</tr>
<tr>
<td>3</td>
<td>6.251</td>
<td>7.815</td>
<td>9.837</td>
<td>11.345</td>
<td>16.266</td>
</tr>
<tr>
<td>5</td>
<td>9.236</td>
<td>11.070</td>
<td>13.388</td>
<td>15.086</td>
<td>20.515</td>
</tr>
<tr>
<td>6</td>
<td>10.645</td>
<td>12.592</td>
<td>15.033</td>
<td>16.812</td>
<td>22.457</td>
</tr>
<tr>
<td>7</td>
<td>12.017</td>
<td>14.067</td>
<td>16.622</td>
<td>18.475</td>
<td>24.322</td>
</tr>
<tr>
<td>11</td>
<td>17.275</td>
<td>19.675</td>
<td>22.618</td>
<td>24.725</td>
<td>31.264</td>
</tr>
<tr>
<td>13</td>
<td>19.812</td>
<td>22.362</td>
<td>25.472</td>
<td>27.688</td>
<td>34.528</td>
</tr>
<tr>
<td>15</td>
<td>22.307</td>
<td>24.996</td>
<td>28.259</td>
<td>30.578</td>
<td>37.697</td>
</tr>
<tr>
<td>16</td>
<td>23.542</td>
<td>26.296</td>
<td>29.633</td>
<td>32.000</td>
<td>39.252</td>
</tr>
<tr>
<td>17</td>
<td>24.769</td>
<td>27.587</td>
<td>30.995</td>
<td>33.409</td>
<td>40.790</td>
</tr>
<tr>
<td>18</td>
<td>25.989</td>
<td>28.869</td>
<td>32.346</td>
<td>34.805</td>
<td>42.312</td>
</tr>
<tr>
<td>19</td>
<td>27.204</td>
<td>30.144</td>
<td>33.687</td>
<td>36.191</td>
<td>43.820</td>
</tr>
<tr>
<td>20</td>
<td>28.412</td>
<td>31.410</td>
<td>35.020</td>
<td>37.566</td>
<td>45.315</td>
</tr>
<tr>
<td>21</td>
<td>29.615</td>
<td>32.671</td>
<td>36.343</td>
<td>38.932</td>
<td>46.797</td>
</tr>
<tr>
<td>22</td>
<td>30.813</td>
<td>33.924</td>
<td>37.656</td>
<td>40.289</td>
<td>48.268</td>
</tr>
<tr>
<td>23</td>
<td>32.007</td>
<td>35.172</td>
<td>38.968</td>
<td>41.638</td>
<td>49.728</td>
</tr>
<tr>
<td>24</td>
<td>33.196</td>
<td>36.415</td>
<td>40.270</td>
<td>42.980</td>
<td>51.179</td>
</tr>
<tr>
<td>25</td>
<td>34.382</td>
<td>37.652</td>
<td>41.566</td>
<td>44.314</td>
<td>52.620</td>
</tr>
<tr>
<td>26</td>
<td>35.563</td>
<td>38.885</td>
<td>42.856</td>
<td>45.642</td>
<td>54.052</td>
</tr>
<tr>
<td>27</td>
<td>36.741</td>
<td>40.113</td>
<td>44.140</td>
<td>46.963</td>
<td>55.476</td>
</tr>
<tr>
<td>28</td>
<td>37.916</td>
<td>41.337</td>
<td>45.419</td>
<td>43.278</td>
<td>56.893</td>
</tr>
<tr>
<td>29</td>
<td>39.087</td>
<td>42.557</td>
<td>46.693</td>
<td>49.588</td>
<td>58.302</td>
</tr>
<tr>
<td>30</td>
<td>40.256</td>
<td>43.773</td>
<td>47.962</td>
<td>50.892</td>
<td>59.703</td>
</tr>
</tbody>
</table>

This is only a portion of a much larger table.
## F Distribution Critical Values

<table>
<thead>
<tr>
<th>Denominator DF</th>
<th>Numerator DF</th>
<th>Alpha .05</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6.608</td>
<td>4.373</td>
</tr>
<tr>
<td>7</td>
<td>5.591</td>
<td>3.329</td>
</tr>
<tr>
<td>10</td>
<td>4.965</td>
<td>2.548</td>
</tr>
<tr>
<td>13</td>
<td>4.667</td>
<td>2.218</td>
</tr>
<tr>
<td>15</td>
<td>4.543</td>
<td>2.078</td>
</tr>
<tr>
<td>20</td>
<td>4.351</td>
<td>1.856</td>
</tr>
<tr>
<td>25</td>
<td>4.242</td>
<td>1.725</td>
</tr>
<tr>
<td>30</td>
<td>4.171</td>
<td>1.637</td>
</tr>
<tr>
<td>35</td>
<td>4.121</td>
<td>1.574</td>
</tr>
<tr>
<td>40</td>
<td>4.085</td>
<td>1.526</td>
</tr>
<tr>
<td>45</td>
<td>4.057</td>
<td>1.488</td>
</tr>
<tr>
<td>50</td>
<td>4.034</td>
<td>1.457</td>
</tr>
<tr>
<td>55</td>
<td>4.016</td>
<td>1.431</td>
</tr>
<tr>
<td>60</td>
<td>4.001</td>
<td>1.409</td>
</tr>
<tr>
<td>120</td>
<td>3.920</td>
<td>1.280</td>
</tr>
<tr>
<td>500</td>
<td>3.860</td>
<td>1.159</td>
</tr>
</tbody>
</table>

This is only a portion of a much larger table.

<table>
<thead>
<tr>
<th>Denominator DF</th>
<th>Numerator DF</th>
<th>Alpha .01</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>16.258</td>
<td>9.042</td>
</tr>
<tr>
<td>7</td>
<td>12.246</td>
<td>5.671</td>
</tr>
<tr>
<td>10</td>
<td>10.044</td>
<td>3.930</td>
</tr>
<tr>
<td>13</td>
<td>9.074</td>
<td>3.187</td>
</tr>
<tr>
<td>15</td>
<td>8.683</td>
<td>2.891</td>
</tr>
<tr>
<td>20</td>
<td>8.096</td>
<td>2.445</td>
</tr>
<tr>
<td>25</td>
<td>7.770</td>
<td>2.194</td>
</tr>
<tr>
<td>30</td>
<td>7.562</td>
<td>2.032</td>
</tr>
<tr>
<td>35</td>
<td>7.419</td>
<td>1.918</td>
</tr>
<tr>
<td>40</td>
<td>7.314</td>
<td>1.833</td>
</tr>
<tr>
<td>45</td>
<td>7.234</td>
<td>1.767</td>
</tr>
<tr>
<td>50</td>
<td>7.171</td>
<td>1.713</td>
</tr>
<tr>
<td>55</td>
<td>7.119</td>
<td>1.670</td>
</tr>
<tr>
<td>60</td>
<td>7.077</td>
<td>1.633</td>
</tr>
<tr>
<td>120</td>
<td>6.851</td>
<td>1.421</td>
</tr>
<tr>
<td>500</td>
<td>6.686</td>
<td>1.232</td>
</tr>
</tbody>
</table>

This is only a portion of a much larger table.
Appendix

Basic Formulas

Glossary of Abbreviated Terms

Order of Mathematical Operations
Basic Formulas

Population Mean

\[ \mu = \frac{\Sigma X_i}{N} \]

Sample Mean

\[ \overline{X} = \frac{\Sigma X_i}{n} \]

Population Variance

\[ \sigma^2 = \frac{\Sigma (X_i - \mu)^2}{N} \]

Sample Variance

\[ S^2 = \frac{\Sigma (X_i - \overline{X})^2}{n - 1} \]

Population Standard Deviation

\[ \sigma = \sqrt{\frac{\Sigma (X_i - \mu)^2}{N}} \]

Sample Standard Deviation

\[ S = \sqrt{\frac{\Sigma (X_i - \overline{X})^2}{n - 1}} \]
Z Score

\[ Z = \frac{X_i - \bar{X}}{S} \]

Standard Error--Mean

\[ s_x = \frac{s}{\sqrt{n}} \]

Standard Error--Proportion

\[ s_p = \sqrt{\frac{pq}{n}} \]
Glossary of Abbreviated Terms

\( \bar{x} \)  Mean of a sample

\( Y \)  Any dependent variable

\( F \)  The F ratio

\( Fe \)  Expected frequency

\( Fo \)  Observed frequency

\( Ho \)  Null Hypothesis

\( Ha \)  Alternate Hypothesis

\( N \)  Number of cases in the population

\( n \)  Number of cases in the sample

\( P \)  Proportion

\( p < \)  Probability of a Type I error

\( r \)  Pearson's correlation coefficient

\( r^2 \)  Coefficient of determination

\( s \)  Sample standard deviation

\( s^2 \)  Sample variance

\( X \)  Any independent variable

\( X_i \)  Any individual score
\( \alpha \)  
Alpha-- probability of Type I error

\( \mu \)  
Mean (mu) of a population

\( \sigma \)  
Population standard deviation

\( \sigma^2 \)  
Population variance

\( \Sigma \)  
"Summation of"

\( \chi^2 \)  
Chi-square statistic

\( \infty \)  
Infinity
Order of Mathematical Operations

First
All operations in parentheses, brackets, or braces, are calculated from the inside out
\[(1 + 1) \times 2 = 4\]

Second
All operations with exponents
\[(1 + 1)^2 \times 2 = 8\]

Third
Multiply and divide before adding and subtracting
\[2 \times 2 + 1 = 5\]
\[2 \div 2 - 1 = 0\]
A priori

Decisions that are made before the results of an inquiry are known. In hypothesis testing, testable hypotheses and decisions regarding alpha and whether a one-tailed or two-tailed test will be used should be established before collecting or analyzing the data.

Related Glossary Terms
Drag related terms here
Abstract concept

The starting point for measurement. Abstract concepts are best understood as general ideas in linguistic form that help us describe reality. They range from the simple (hot, heavy, fast) to the more difficult (responsive, effective, fair). Abstract concepts should be evident in the research question and/or purpose statement.

An example of a research question is given below. In this example, the abstract concepts are sector of employment (e.g., private, non-profit, public) and employee quality.

*Research Question:* Is the quality of public sector and private sector employees different?

**Related Glossary Terms**

Research question
Addition rule of probability

The probability of outcome A or outcome B occurring is equal to the sum of their respective probabilities minus the probability of their joint occurrence. In the following example, a joint occurrence is not possible.

Example

Out of a deck of 52 playing cards, what is the probability of picking an Ace or a Jack? Assume a well shuffled deck and mutually exclusive event possibilities (i.e., you cannot pull a Jack and an Ace at the same time).

Addition Rule: When two events, A and B, are mutually exclusive, the probability that A or B will occur is the sum of the probability of each event.

\[ P(A \text{ or } B) = P(A) + P(B) \]

or

\[ P(\text{Ace or Jack}) = \frac{4}{52} + \frac{4}{52} = \frac{8}{52} \]

There is about a 15% chance of pulling an Ace or Jack with one draw from a deck of playing cards.

Related Glossary Terms

Multiplication rule of probability
**Alpha**

The probability of a Type I error. Alpha represents the threshold for claiming statistical significance.

When conducting statistical tests with computer software, the exact probability of a Type I error is calculated. It is presented in several formats but is most commonly reported as "p <" or "Sig." or "Signif." or "Significance." Using "p <" as an example, if a priori you established a threshold for statistical significance at alpha .05, any test statistic with significance at or less than .05 would be considered statistically significant and you would be required to reject the null hypothesis of no difference.

The following table links p values with a constant alpha benchmark of .05.

<table>
<thead>
<tr>
<th>P &lt;</th>
<th>Alpha</th>
<th>Probability of Making a Type I Error</th>
<th>Significant?</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.05</td>
<td>5% chance not statistically significant</td>
<td>Yes</td>
</tr>
<tr>
<td>0.10</td>
<td>0.05</td>
<td>10% chance not statistically significant</td>
<td>No</td>
</tr>
<tr>
<td>0.01</td>
<td>0.05</td>
<td>1% chance not statistically significant</td>
<td>Yes</td>
</tr>
<tr>
<td>0.96</td>
<td>0.05</td>
<td>96% chance not statistically significant</td>
<td>No</td>
</tr>
</tbody>
</table>

**Related Glossary Terms**

P-value, Type I error, Type II error
Alternative hypothesis

This is a statement that asserts there is a difference between two population parameters. It has the opposite assertion of a null hypothesis. An alternative hypothesis is also known as the research hypothesis.

Example

Alternate Hypothesis: There is a statistically significant difference between the historical proportion of clients reporting poor service and the current proportion of clients reporting poor service.

Null Hypothesis: There is no statistically significant difference between the historical proportion of clients reporting poor service and the current proportion of clients reporting poor service.

The four possible outcomes in hypothesis testing

<table>
<thead>
<tr>
<th>DECISION</th>
<th>Actual Population Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Null Hyp. True</td>
</tr>
<tr>
<td></td>
<td>(there is no difference)</td>
</tr>
<tr>
<td>Rejected Null Hyp</td>
<td>Type I error (alpha)</td>
</tr>
<tr>
<td>Did not Reject Null</td>
<td>Correct Decision</td>
</tr>
</tbody>
</table>

(Alpha = probability of making a Type I error)

Related Glossary Terms

Hypothesis, Null hypothesis
Antecedent variable

An independent variable that comes before other independent variables.

Related Glossary Terms
Dependent variable, Independent variable, Intervening variable
Simultaneously analyzing two variables to determine a relationship between the variables.

Software Output Example: Education level by individual’s sex.

Crosstabulation: EduCat (Rows) by Sex (Columns)

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;12 yrs</td>
<td>3</td>
<td>4 [a]</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>42.86</td>
<td>57.14 [b]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11.54</td>
<td>16.67 [c]</td>
<td>14.00</td>
</tr>
<tr>
<td></td>
<td>6.00</td>
<td>8.00 [d]</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HS Grad</th>
<th>10</th>
<th>11</th>
<th>21</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>47.62</td>
<td>52.38</td>
<td></td>
</tr>
<tr>
<td></td>
<td>38.46</td>
<td>45.83</td>
<td>42.00</td>
</tr>
<tr>
<td></td>
<td>20.00</td>
<td>22.00</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>College</th>
<th>13</th>
<th>9</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>59.09</td>
<td>40.91</td>
<td></td>
</tr>
<tr>
<td></td>
<td>50.00</td>
<td>37.50</td>
<td>44.00</td>
</tr>
<tr>
<td></td>
<td>26.00</td>
<td>18.00</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>26</th>
<th>24</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>52.00</td>
<td>48.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Interpretation

[a] Count: There are 4 females who have less than 12 years of education.

[b] Row %: Of those who have less than a high school education (<12 yrs), 57.14% (4/7) are female.

[c] Column %: Of those who are female, 16.67% (4/24) have less than a high school education.

[d] Total %: 8% (4/50) of the sample is female with less than a high school education.

Related Glossary Terms
Applied Stat: Counts-Chi-square, Bivariate table, Contingency table, Univariate analysis
Simultaneously analyzing two variables to determine a relationship between the variables. Examples include t-tests of means and Analysis of Variance.

**Software Output Example:** Individual annual mean income by education level.

<table>
<thead>
<tr>
<th>EduCat</th>
<th>Count</th>
<th>Mean</th>
<th>Median</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;12 yrs</td>
<td>7</td>
<td>23188.4286</td>
<td>24064.0000</td>
<td>4081.0739</td>
</tr>
<tr>
<td>HS Grade</td>
<td>21</td>
<td>30731.2381</td>
<td>28086.0000</td>
<td>8521.1830</td>
</tr>
<tr>
<td>College</td>
<td>22</td>
<td>40829.7273</td>
<td>35739.0000</td>
<td>21211.0346</td>
</tr>
</tbody>
</table>

**Related Glossary Terms**

Drag related terms here
Applied Stat: Correlation

Used to explain the relationship between variables. Represents the notion that one variable will vary systematically depending on the values of another variable.

Correlation coefficients estimate strength and direction of association between two interval/ratio level variables. Used to create a summary measure that reflects the covariation between two interval/ratio variables, the Pearson Correlation Coefficient presented here can range from a -1.00 to 1.00. A positive coefficient indicates the values of variable A vary in the same direction as variable B. A negative coefficient indicates the values of variable A and variable B vary in opposite directions.

Assumptions

Interval/ratio data from paired observations.
A linear relationship should exist between the variables -- verified by plotting the data on a scattergram.
No extreme values in the data.

Probability Distribution

T Distribution where ...

Degrees of Freedom

df= n-2

Formula

\[
\rho = \frac{n \sum xy - \sum x \sum y}{\sqrt{(n \sum x^2 - (\sum x)^2) (n \sum y^2 - (\sum y)^2)}}
\]

Software Output Example: Comparing individual income in U.S. dollars to years of education.

Pearson's Correlation

INCOME: [a]

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>21779.0</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>14030.0</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>6281.0</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

EDUC: [b]

---------------|--------------|
4.0          12.0          20.0

Number of cases: 32
Missing: 0 [c]
Pearson Correlation: 0.751 [d]
p < (2-tailed signif.): 0.0000 [e]

Interpretation

[a] The Y axis of the scattergram. If theory suggests cause and effect, the Y axis is commonly used for the dependent (response) variable.

[b] The X axis of the scattergram. If theory suggests cause and effect, the X axis is commonly used for the independent variable.

[c] Since each observation (case) must have values for both income and education, any observations where one or both of these variables have no data will be removed from the analysis.

[d] Pearson correlation coefficient representing a high positive correlation between education and income. Interpretation: As years of education increases so does personal income.

[e] There is a statistically significant association between income and education.

Related Glossary Terms

Applied Stat: Correlation Matrix, Association, Covariation, Covary, Zero-order correlation
Applied Stat: Correlation Matrix

Used to explain the relationship among multiple variables. Represents the notion that one variable will vary systematically depending on the values of another variable. The example below shows Pearson Product Moment Correlations in matrix form. The cells above the diagonal are a mirror reflection of the cells below the diagonal.

Software Output Example: Multiple bivariate comparisons displayed in a correlation matrix. Comparing individual income, education, and months of work experience.

Correlation: Pearson (R) Coefficients

<table>
<thead>
<tr>
<th>Coeff</th>
<th>Correlation Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>-------------------</td>
</tr>
<tr>
<td>p &lt;</td>
<td>INCOME</td>
</tr>
</tbody>
</table>

| INCOME | 1.000 [a] | 0.751 [d] | -0.160 [f] |
|        | 32 [b]    | 32     | 32       |
|        | . [c]     | 0.000  | 0.381    |

| EDUC   | 0.751      | 1.000  | -0.520 |
|        | 32         | 32     | 32     |
|        | 0.000      | .      | 0.002  |

| WORKEXP | -0.160 | -0.520 | 1.000 |
|         | 32     | 32     | 32    |
|         | 0.381  | 0.002  | .     |

2-tailed significance tests
'. ' p-value not computed

Interpretation

[a] The diagonal in the matrix represents Pearson correlations between the same variable which will always be 1 since the variables are identical. The correlations above the diagonal are a mirror reflection of these below the diagonal, so you only interpret half of the matrix (in this case three correlation coefficients).

[b] The number of paired observations for this comparison.

[c] A statistical test is not performed when you are comparing a variable to itself. Mathematically this will always equal zero.

[d] Pearson correlation coefficient representing a high positive correlation between education and income. Interpretation: As years of education increases so does personal income.

[e] There is a statistically significant association between income and education.

[f] Pearson correlation coefficient representing a very weak negative correlation between work experience and income. Interpretation: As years of work experience increases personal income decreases.

[g] There is not a statistically significant association between work experience and income.

Related Glossary Terms

Applied Stat: Correlation
Chi-square compares the observed and expected counts of two or more subgroups across two or more criteria. By convention, subgroups of the independent variable are placed in the columns and the subgroups of the dependent variable are represented by the rows.

**Assumptions**

- Independent random sampling
- Nominal/Ordinal level data
- No more than 20% of the cells have an expected frequency less than 5
- No empty cells

**Probability Distribution**

Chi-square Distribution where...

**Degrees of Freedom**

\[ df = (\text{# of rows} - 1)(\text{# of columns} - 1) \]

**Formula**

Chi-square Test Statistic

\[ \chi^2 = \sum \frac{(O - E)^2}{E} \]

where

- \( O \) = observed frequency
- \( E \) = expected frequency for each cell
- \( E = (\text{frequency for the column}) \times (\text{frequency for the row}) / n \)

**Software Output Example:** Data from survey of U.S. Adults.

<table>
<thead>
<tr>
<th>Gun Law (Rows)</th>
<th>SEX (Columns)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
</tr>
<tr>
<td>Favor</td>
<td>314</td>
</tr>
<tr>
<td></td>
<td>38.72</td>
</tr>
<tr>
<td></td>
<td>73.88</td>
</tr>
<tr>
<td>Oppose</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td>64.16</td>
</tr>
<tr>
<td></td>
<td>26.12</td>
</tr>
<tr>
<td></td>
<td>11.28</td>
</tr>
<tr>
<td>Total</td>
<td>425</td>
</tr>
<tr>
<td></td>
<td>43.19</td>
</tr>
</tbody>
</table>

**Chi-square**

- Pearson: 37.622, 1, 0.0000
- Likelihood Ratio: 37.417, 1, 0.0000
- Yate's Correction: 36.592, 1, 0.0000

**Measures of Association**

- Cramer's V: 0.196
- Pearson C: 0.192
- Lambda Symmetric: 0.002
- Lambda Dependent x Column: 0.000
- Lambda Dependent x Row: 0.115

**Note:** 0.00% of the cells have an expected frequency <5

**Interpretation**

[a] Statistically significant (most common measure used for significance).

[b] When sample sizes are small, the continuous chi-square value tends to be too large. The Yate continuity correction adjusts for this bias in 2x2 contingency tables. Regardless of sample size, it is a preferred measure for chi-square tests on 2x2 tables.


[d] A symmetric lambda is used when identification of independent and dependent variables is not useful.

[e] Knowing a person's sex can reduce prediction error by 11.5%.

**Related Glossary Terms**

- Bivariate Analysis-Counts
- Counts-Cramer’s V
- Counts-Goodness of fit
- Counts-Lambda
- Counts-Yates continuity correction
- Fisher’s Exact Test
- Gamma
- Nonparametric
- Pearson’s C
- Phi
- Proportional reduction in error (PRE)
When using the chi-square statistic, Cramer’s V coefficients can be helpful in interpreting the strength of a relationship between two variables once statistical significance has been established. Cramer’s V is also useful for comparing multiple $X^2$ test statistics and is comparable across contingency tables of varying sizes. It is not affected by sample size and therefore is very useful in situations where you suspect a statistically significant chi-square was the result of large sample size instead of any substantive relationship between the variables. It is interpreted as a measure of the relative (strength) of an association between two variables from 0 (no association) to 1 (very strong association). In practice, a Cramer’s V of .10 may provide a good minimum threshold for suggesting there is a substantive relationship between two variables.

**Formula**

$$V = \sqrt{\frac{X^2}{(n-1)q}}$$

where $q =$ smaller # of rows or columns

**Software Output Example:** Data from survey of U.S. Adults.

Crosstabulation: GUNLAW (Rows) by SEX (Columns)

Column Variable Label: Respondent’s Sex
Row Variable Label: Gun permits

<table>
<thead>
<tr>
<th>Count</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>314</td>
<td>497</td>
<td>811</td>
<td></td>
</tr>
<tr>
<td>38.72</td>
<td>61.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>73.88</td>
<td>88.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31.91</td>
<td>50.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oppose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>111</td>
<td>62</td>
<td>173</td>
<td></td>
</tr>
<tr>
<td>64.16</td>
<td>35.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26.12</td>
<td>11.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.28</td>
<td>6.30</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Total |
|-------|------|-------|
|  425 |  559 |  984  |
|  43.19 |  56.81 |

Chi-square

<table>
<thead>
<tr>
<th>Value</th>
<th>DF</th>
<th>p &lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>37.622</td>
<td>1</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>37.417</td>
<td>1</td>
</tr>
<tr>
<td>Yates’s Correction</td>
<td>36.592</td>
<td>1</td>
</tr>
</tbody>
</table>

Measures of Association

| Cramer’s V | .196 | [a] |
| Pearson C | .192 |
| Lambda Symmetric | .082 |
| Lambda Dependent=Column | .000 |
| Lambda Dependent=Row | .115 |

Note: 0.00% of the cells have an expected frequency <5

**Interpretation**

[a] Weak association (both Cramer’s V and Pearson C).

**Related Glossary Terms**

Applied Stat: Counts-Chi-square, Association
Applied Stat: Counts-Goodness of fit

Uses chi-square to test the significance of the distribution of observations in a set of categories in one variable.

Also used in logistic regression to estimate the overall performance of the regression model.

**Assumptions**

Independent random sampling
Nominal or Ordinal level data

**Probability Distribution**

Chi-square Distribution where ...

**Degrees of Freedom**

\[ (df) = k - 1 \text{ where } k \text{ equals the number of categories} \]

**Formula**

\[ \chi^2 = \sum \left( \frac{(F_o - F_e)^2}{F_e} \right) \]

where \( F_e = \text{Frequency Expected} \)

\[ F_e = \frac{n}{k} \]

and ...

\( n = \text{sample size} \)
\( k = \text{number of categories or cells} \)
\( F_o = \text{observed frequency} \)

**Software Output Example: Education Level of a small sample of U.S. Adults.**

**Frequencies**

<table>
<thead>
<tr>
<th>Variable: EduCat</th>
<th>Education Level</th>
<th>Count</th>
<th>Percent</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>14.00</td>
<td>14.00</td>
</tr>
<tr>
<td>&lt;12 yrs</td>
<td></td>
<td>21</td>
<td>42.00</td>
<td>56.00</td>
</tr>
<tr>
<td>College</td>
<td></td>
<td>22</td>
<td>44.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

---------------------------------------------
Total         50                     Missing  0
Value Count   3

**Goodness-of-Fit**

<table>
<thead>
<tr>
<th>Value</th>
<th>DF</th>
<th>p &lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-Sample Chi-square</td>
<td>8.440</td>
<td>2</td>
</tr>
</tbody>
</table>

**Note:** Frequency Expected = 16.67

**Interpretation**

[a] There is a statistically significant difference among the education categories.

**Related Glossary Terms**

Applied Stat: Counts-Chi-square
Applied Stat: Counts-Lambda

Lambda is a proportional reduction in error (PRE) measure. It represents the proportion by which you reduce the error in predicting the dependent variable by incorporating values from the independent variable. The value of lambda can range from 0 (no improvement in prediction) to 1 (knowing one variable means you can predict without error). A symmetric lambda is used when identification of independent/dependent variables is not useful.

Software Output Example: Data from survey of U.S. Adults.

Crosstabulation: GUNLAW (Rows) by SEX (Columns)

Column Variable Label: Respondent's Sex
Row Variable Label: Gun permits

<table>
<thead>
<tr>
<th>Count</th>
<th>Row %</th>
<th>Col %</th>
<th>Total %</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favor</td>
<td>314</td>
<td>497</td>
<td>811</td>
<td>38.72</td>
<td>61.28</td>
<td></td>
</tr>
<tr>
<td></td>
<td>73.88</td>
<td>88.91</td>
<td>82.42</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>31.91</td>
<td>50.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oppose</td>
<td>111</td>
<td>62</td>
<td>173</td>
<td>64.16</td>
<td>35.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>26.12</td>
<td>11.09</td>
<td>17.58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11.28</td>
<td>6.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>425</td>
<td>559</td>
<td>984</td>
<td>43.19</td>
<td>56.81</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Chi-square

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>DF</th>
<th>p &lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>37.622</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>37.417</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>Yate's Correction</td>
<td>36.592</td>
<td>1</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Measures of Association

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cramer's V</td>
<td>.196</td>
</tr>
<tr>
<td>Pearson C</td>
<td>.192</td>
</tr>
<tr>
<td>Lambda Symmetric</td>
<td>.082</td>
</tr>
<tr>
<td>Lambda Dependent=Column</td>
<td>.000</td>
</tr>
<tr>
<td>Lambda Dependent=Row</td>
<td>.115</td>
</tr>
</tbody>
</table>

Note: 00.00% of the cells have an expected frequency <5

Interpretation

[a] A symmetric lambda is used when identification of independent and dependent variables is not useful.

[b] Knowing a person's sex can reduce prediction error by 11.5%.

Related Glossary Terms
Applied Stat: Counts-Chi-square, Association, Measurement scale
Applied Stat: Counts-Yates continuity correction

When sample sizes are small, the continuous chi-square tends to be too large. The continuity correction adjusts for this bias in 2x2 contingency tables. Regardless of sample size, it is a preferred measure for chi-square tests on 2x2 tables. If a cell has zero observations, Fisher’s exact test is more appropriate for chi-square significance tests.

Software Output Example: Data from survey of U.S. Adults.

Crosstabulation: GUNLAW (Rows) by SEX (Columns)

Column Variable Label: Respondent’s Sex
Row Variable Label: Gun permits

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favor</td>
<td>314</td>
<td>497</td>
<td>811</td>
</tr>
<tr>
<td></td>
<td>38.72</td>
<td>61.28</td>
<td></td>
</tr>
<tr>
<td></td>
<td>73.88</td>
<td>88.91</td>
<td>82.42</td>
</tr>
<tr>
<td></td>
<td>31.91</td>
<td>50.51</td>
<td></td>
</tr>
<tr>
<td>Oppose</td>
<td>111</td>
<td>62</td>
<td>173</td>
</tr>
<tr>
<td></td>
<td>64.16</td>
<td>35.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>26.12</td>
<td>11.09</td>
<td>17.58</td>
</tr>
<tr>
<td></td>
<td>11.28</td>
<td>6.30</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>425</td>
<td>559</td>
<td>984</td>
</tr>
<tr>
<td></td>
<td>43.19</td>
<td>56.81</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Chi-square

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>DF</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>37.622</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>37.417</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>Yate’s Correction</td>
<td>36.592</td>
<td>1</td>
<td>0.0000 [a]</td>
</tr>
</tbody>
</table>

Note: 0.00% of the cells have an expected frequency <5

Interpretation

[a] When sample sizes are small, the continuous chi-square value tends to be too large. The Yates continuity correction adjusts for this bias in 2x2 contingency tables. Regardless of sample size, it is a preferred measure for chi-square tests on 2x2 tables.

Related Glossary Terms
Applied Stat: Counts-Chi-square
Applied Stat: Frequencies

Used to represent the relative distribution of values in one variable. The categories can be any level of measurement and are presented as counts, percents, and cumulative percents.

Software Output Example: Education level.

<table>
<thead>
<tr>
<th>Value</th>
<th>Count</th>
<th>Percent</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>08</td>
<td>1</td>
<td>2.00</td>
<td>2.00</td>
</tr>
<tr>
<td>09</td>
<td>2</td>
<td>4.00</td>
<td>6.00</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>4.00</td>
<td>10.00</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>4.00</td>
<td>14.00</td>
</tr>
<tr>
<td>12</td>
<td>21</td>
<td>42.00</td>
<td>56.00</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>4.00</td>
<td>60.00</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>8.00</td>
<td>68.00</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>6.00</td>
<td>74.00</td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>12.00</td>
<td>86.00</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>6.00</td>
<td>92.00</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>2.00</td>
<td>94.00</td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>4.00</td>
<td>98.00</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>2.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total 50
Missing 0

Interpretation

[a] 2 out of 50 have 9 years of education
[b] 4.00% have 9 years of education (2/50=.04; .04 * 100=4.00%)
[c] 6.00% have 9 years or less education (3/50=.06; .06*100 = 6.00%)

Related Glossary Terms
Cumulative frequency, Cumulative percentage, Frequency distribution, Relative frequency
Cohen's kappa is a common technique for estimating paired interrater agreement for nominal and ordinal level data. Kappa is a coefficient that represents agreement obtained between two readers beyond what would be expected by chance alone. A value of 1.0 represents perfect agreement. A value of 0.0 represents no agreement.

**Assumptions**

Elements being rated (images, diagnoses, clinical indications, etc.) are independent of each other.

One rater's classifications are made independently of the other rater's classifications.

The same two raters provide the classifications used to determine kappa.

The rating categories are independent of each other.

**Formula**

\[
\hat{k} = \frac{\hat{p}_c - \hat{p}_e}{1 - \hat{p}_e}
\]

where

\(\hat{p}_c\) = Proportion of cases where there is agreement between two raters

\(\hat{p}_e\) = Proportion of cases where raters would agree by chance

**Software Output Example: Breast Imaging by Two Radiologists.**

Crosstabulation: Radiologist A by Radiologist B

<table>
<thead>
<tr>
<th></th>
<th>Not Diseased</th>
<th>Diseased</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>50</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>Row %</td>
<td>50.00</td>
<td>50.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Col %</td>
<td>50.00</td>
<td>50.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

**Agreement Statistics**

<table>
<thead>
<tr>
<th>Proportion Agree</th>
<th>0.800 (a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kappa Statistic</td>
<td>0.600 (b)</td>
</tr>
<tr>
<td>Kappa Standard Error</td>
<td>0.100</td>
</tr>
<tr>
<td>Kappa Significance (p&lt;)</td>
<td>0.000 (c)</td>
</tr>
</tbody>
</table>

**Interpretation**

[a] The radiologist interpretations agree in 80% of the patients (note they are reviewing imaging of the same patients independently).

[b] Kappa indicates a moderate level of agreement.

[c] The agreement between the two radiologists is statistically significant.

**Related Glossary Terms**

Applied Stat: Counts-Chi-square, Association
**Applied Stat: Mean-1 sample**

Compares a mean obtained from a random sample to a population mean. It is generally used to determine if there is a difference between what was obtained in a random sample and an established benchmark that either originated from all members of a comparable population or is assumed to represent the population.

**Assumptions**

- Interval/ratio level data
- Random sampling
- Normal distribution in population

**Probability Distribution**

- T- Distribution
- Degrees of Freedom = n-1

**Formula**

\[ t = \frac{\bar{X} - \mu}{S_x} \]

Where

- Standard error of the sample mean

\[ S_x = \frac{s}{\sqrt{n}} \]

**Software Output Example: Starting income after completing two years of college.**

<table>
<thead>
<tr>
<th>One-Sample Difference of Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Mean</td>
</tr>
<tr>
<td>Sample Mean</td>
</tr>
<tr>
<td>Std Deviation</td>
</tr>
<tr>
<td>Sample Size (n)</td>
</tr>
<tr>
<td>Test statistic</td>
</tr>
<tr>
<td>p &lt;</td>
</tr>
</tbody>
</table>

**Interpretation**

[a] Represents the historical mean starting income of all graduates of two year technical colleges in the United States (in constant dollars).

[b] Represents the mean starting income of a random sample of 30 graduates from this year’s technical colleges in the United States.

[c] The standard deviation for the sample of 30 graduates (you generally use the sample standard deviation unless the population standard deviation is known).

[d] There is a statistically significant difference between the starting income of this year’s graduating class and the historical average starting pay. This year’s graduates have a mean starting pay that is significantly less than the mean pay for prior graduates.

Note: The above example uses StatCalc output.

**Related Glossary Terms**

- Applied Stat: Mean-2 sample
Compared two means from two random samples: such as mean incomes of males and females, Republicans and Democrats, Urban and Rural residents. If there are more than two comparable groups, ANOVA is a more appropriate technique.

Assumptions

Random sampling

Independent samples

Interval/ratio level data

Homogeneity of variance

Probability Distribution

T distribution

Degrees of Freedom = (n1+n2)-2

Formula

\[ t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \]

Where the Standard Error

\[ s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}} \]

Software Output Example: Comparing the mean income of males and females.

Two-Sample Difference of Means

Independent Variable: SEX
Dependent Variable: INCOME

Sample One: 0 Female
Sample Two: 1 Male

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>Sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>Sample Mean</td>
<td>12279.4667</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>4144.6323</td>
</tr>
<tr>
<td>Sample Size (n)</td>
<td>15</td>
</tr>
</tbody>
</table>

Homogeneity of Variance

F-ratio 1.09  DF (14, 16)  p < 0.4283 [a]

T-statistic  DF  p < (2-tailed)

Equal Variance -2.0870  30  0.0455 [b]
Unequal Variance -2.0800  31.05  0.0459

Interpretation

[a] The F-ratio is not statistically significant. Therefore, use the equal variance test statistic.

[b] For a 2-tailed test, the p-value represents the probability of making a type 1 error (concluding there is statistical significance when there is none). Since there is about 4.6% (p < .0455) chance of making a type 1 error, which does not exceed the 5% error limit (p=.05) established in the rejection criteria of the hypothesis testing process (alpha), you conclude there is statistical significance.

Related Glossary Terms

Applied Stat: Mean-1 sample

Index
Applied Stat: Mean-Analysis of Variance

Used to evaluate means from two or more subgroups. A statistically significant ANOVA indicates there is more variation between subgroups than would be expected by chance. It does not identify which subgroup pairs are significantly different from each other. Used to evaluate multiple means of one independent variable to avoid conducting multiple t-tests.

Assumptions

- Independent Random sampling
- Interval/ratio level data
- Population variances equal
- Groups are normally distributed

Probability Distribution

F-Distribution where . . .

Numerator Degrees of Freedom
df=k-1 where k=number of independent samples/groups

Denominator Degrees of Freedom
df=n-k where n=sum of all observations from the independent samples/groups

Formula

F-Test Statistic

\[ F = \frac{\sum (X_i - \overline{X})^2 / (k-1)}{\sum (X_i - \overline{X})^2 / (n-k)} \]

where

\( \overline{X} \) = each group mean
\( \overline{X} \) = grand mean for all the groups = (sum of all scores)/N
\( n \) = number in each group

Note:  
\[ F = \frac{\text{Mean Squares between groups}}{\text{Mean Squares within groups}} \]

Software Output Example: Comparing amount of time spent watching TV per day by four age groups (n=990).

Analysis of Variance (ANOVA): TVHOURS (Means) by AGECAT4 (Groups)

Independent Variable Label: Age categories
Dependent Variable Label: Hours watch TV per day

Analysis of Variance

<table>
<thead>
<tr>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>92.09</td>
<td>3 [a]</td>
</tr>
<tr>
<td>Within Groups</td>
<td>5185.54</td>
<td>986 [c]</td>
</tr>
<tr>
<td>Total</td>
<td>5277.63</td>
<td>989 [e]</td>
</tr>
</tbody>
</table>

F Statistic: 5.94 [f]  p < 0.0005 [g]

Interpretation

[a]  k-1, where k = 4 group means
[b]  Between Groups Sum of Squares + degrees of freedom
[c]  n-k, where n = 990
[d]  Within Groups Sum of Squares + degrees of freedom
[e]  n-1
[f]  Between Groups Mean Squares + Within Groups Mean Squares
[g]  The F-Statistic is statistically significant. The p-value represents the probability of making a type 1 error (concluding there is statistical significance when there is none). Since there is about 0.05% (p < .0005) chance of making a type 1 error, you conclude there is statistical significance; i.e., there is variation among the groups.

Related Glossary Terms

Applied Stat: Mean-Bonferroni post hoc

Index
The Bonferroni post hoc test adjusts alpha to compensate for multiple comparisons. In this situation, one may falsely conclude a significant effect where there is none. In particular, use of a .05 cut-off value for significance theoretically guarantees that if there were 20 pairwise comparisons, there will by chance alone appear to be one with significance at the .05 level. The Bonferroni adjustment is often offered as an additional test for ANOVA models to guide post ANOVA exploration and interpretation regarding which paired mean comparisons are likely to have statistically significant differences. There are other corrections for this problem. Some are useful for unordered groups, such as patient height versus sex, while others are applied to ordered groups (to evaluate a trend), such as patient height versus sex when stratified by age. Consequently, when multiple statistical tests are conducted between the same variables, the significance cut-off value is often adjusted to represent a more conservative estimate of statistical significance. There is a debate about which post hoc method to use, and some hold the view that this correction is overused. The Bonferroni correction adjusts the threshold for significance, which is equal to the desired p-value (e.g., .05, .01) divided by the number of paired outcome variables being examined. One limitation of the Bonferroni correction is that by reducing the level of significance associated with each test, we have reduced the power of the test, thereby increasing the chance of incorrectly retaining the null hypothesis.

Software Output Example: Comparing amount of time spent watching TV per day by age group.

Analysis of Variance (ANOVA): TVHOURS (Means) by AGECAT4 (Groups)

Independent Variable Label: Age categories
Dependent Variable Label: Hours watch TV per day

Analysis of Variance

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>92.09</td>
<td>3</td>
<td>30.70</td>
</tr>
<tr>
<td>Within Groups</td>
<td>5093.45</td>
<td>986</td>
<td>5.17</td>
</tr>
<tr>
<td>Total</td>
<td>5185.54</td>
<td>989</td>
<td></td>
</tr>
</tbody>
</table>

F Statistic

p <

5.94 0.0005

Bonferroni Post Hoc Comparison of Means

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>G1-G2</th>
<th>SE</th>
<th>p&lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>18-29</td>
<td>30-39</td>
<td>0.410</td>
<td>0.209</td>
<td>0.3005</td>
</tr>
<tr>
<td>18-29</td>
<td>40-49</td>
<td>0.425</td>
<td>0.207</td>
<td>0.2464</td>
</tr>
<tr>
<td>18-29</td>
<td>50+</td>
<td>-0.256</td>
<td>0.228</td>
<td>1.0000</td>
</tr>
<tr>
<td>30-39</td>
<td>18-29</td>
<td>-0.410</td>
<td>0.209</td>
<td>0.3005</td>
</tr>
<tr>
<td>30-39</td>
<td>40-49</td>
<td>0.015</td>
<td>0.184</td>
<td>1.0000</td>
</tr>
<tr>
<td>30-39</td>
<td>50+</td>
<td>-0.666</td>
<td>0.200</td>
<td>0.0056 [a]</td>
</tr>
<tr>
<td>40-49</td>
<td>18-29</td>
<td>-0.425</td>
<td>0.207</td>
<td>0.2464</td>
</tr>
<tr>
<td>40-49</td>
<td>30-39</td>
<td>-0.015</td>
<td>0.184</td>
<td>1.0000</td>
</tr>
<tr>
<td>40-49</td>
<td>50+</td>
<td>-0.681</td>
<td>0.209</td>
<td>0.0073 [e]</td>
</tr>
<tr>
<td>50+</td>
<td>18-29</td>
<td>0.256</td>
<td>0.228</td>
<td>1.0000</td>
</tr>
<tr>
<td>50+</td>
<td>30-39</td>
<td>0.666</td>
<td>0.200</td>
<td>0.0056</td>
</tr>
<tr>
<td>50+</td>
<td>40-49</td>
<td>0.681</td>
<td>0.209</td>
<td>0.0073</td>
</tr>
</tbody>
</table>

Interpretation

[a] The difference between the mean of those 30-39 (G1) and the mean of those 50 or older (G2) is a -0.666. A negative mean indicates those 50 or older have a greater mean number of hours watching television than those 30-39.

[b] There is a statistically significant difference between the number of hours watching television for those between the ages of 30 and 39 and the number of hours watching television for those 50 and older.

[c] There is a statistically significant difference between the number of hours watching television for those between the ages of 40 and 49 and the number of hours watching television for those 50 and older.

Related Glossary Terms

Applied Stat: Mean-Analysis of Variance, Post hoc test
Applied Stat: Mean-Confidence Interval

Interval estimation involves using sample data to determine a range (interval) that, at an established level of confidence, is expected to contain the mean of the population.

Steps
1. Determine confidence level (df=n-1; alpha .05, 2-tailed)
2. Use either z distribution (if n>120) or t distribution (for all sizes of n).
3. Use the appropriate table to find the critical value for a 2-tailed test
4. Multiple hypotheses can be compared with the estimated interval for the population to determine their significance. In other words, differing values of population means can be compared with the interval estimation to determine if the hypothesized population means fall within the region of rejection.

Estimation Formula
\[ CI = \bar{x} \pm CV \left( \frac{s}{\sqrt{n}} \right) \]

where
\[ \bar{x} = \text{sample mean} \]
\[ CV = \text{critical value (consult distribution table for df=n-1 and chosen alpha--commonly .05)} \]
\[ s = \text{Standard error of the mean} \]

Note: assumes sample standard deviation was calculated using:
\[ s = \sqrt{\frac{(n-1)s^2}{n}} \]

Example: Interval Estimation for Means

Problem: A random sample of 30 incoming college freshmen revealed the following statistics: mean age 19.5 years; sample standard deviation 1.2. Based on a 5% chance of error, estimate the range of possible mean ages for all incoming freshmen.

Estimation

Critical value (CV)
\[ \text{Df=n-1 or 29} \]
Consult t-distribution for alpha .05, 2-tailed
\[ CV=2.045 \]

Standard error
\[ s = \frac{1.2}{\sqrt{30}} \]
\[ \bar{x} = 19.5 \]

Estimate Confidence Interval
\[ CI = \bar{x} \pm CV \left( \frac{s}{\sqrt{n}} \right) \]
\[ CI_{95} = 19.5 \pm 2.045(2.19) \]
\[ CI_{95} = 19.5 \pm 4.448 \]

Software Output Example: Summary data from the previous example

<table>
<thead>
<tr>
<th>Margin of Error for Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>n  = 30</td>
</tr>
<tr>
<td>Mean = 19.5000</td>
</tr>
<tr>
<td>Standard Deviation = 1.2000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Confidence</th>
<th>Error +/-</th>
<th>Margin Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower</td>
<td>Upper</td>
</tr>
<tr>
<td>90%</td>
<td>0.3723</td>
<td>19.1277</td>
</tr>
<tr>
<td>95%</td>
<td>0.4481</td>
<td>19.0519</td>
</tr>
<tr>
<td>99%</td>
<td>0.6039</td>
<td>18.8961</td>
</tr>
<tr>
<td></td>
<td>19.1277</td>
<td>19.8723</td>
</tr>
<tr>
<td></td>
<td>19.0519</td>
<td>19.9481</td>
</tr>
<tr>
<td></td>
<td>18.8961</td>
<td>20.1039</td>
</tr>
</tbody>
</table>

Note: The above example uses StatCalc output.

Related Glossary Terms
Confidence interval, Confidence limits
Applied Stat: Percentage

The number of cases in a category of a variable divided by the number of cases in all categories of the same variable. The quantity is then multiplied by 100 to convert the proportion into a percentage.

Software Output Example: Education level.

Frequencies
Variable: Edu Years of education

<table>
<thead>
<tr>
<th>Value</th>
<th>Count</th>
<th>Percent</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>08</td>
<td>1</td>
<td>2.00</td>
<td>2.00</td>
</tr>
<tr>
<td>09</td>
<td>2 [a]</td>
<td>4.00 [b]</td>
<td>6.00 [c]</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>4.00</td>
<td>10.00</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>4.00</td>
<td>14.00</td>
</tr>
<tr>
<td>12</td>
<td>21</td>
<td>42.00</td>
<td>56.00</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>4.00</td>
<td>60.00</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>8.00</td>
<td>68.00</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>6.00</td>
<td>74.00</td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>12.00</td>
<td>86.00</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>6.00</td>
<td>92.00</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>2.00</td>
<td>94.00</td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>4.00</td>
<td>98.00</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>2.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total 50
Missing 0

Interpretation

[a] Count: 2 out of 50 have 9 years of education

[b] Percentage: 4% have 9 years of education (2/50=.04; .04 * 100=4.00%)

[c] Cumulative Percentage: 6% have 9 years or less education (3/50=.06; .06*100 = 6.00%)

Related Glossary Terms
Applied Stat: Proportion, Relative frequency
Applied Stat: Proportion

A frequency proportion represents a subfrequency divided by the total frequency. This weights the count in the subgroup by the total. When multiplied by 100, a proportion becomes a percentage.

<table>
<thead>
<tr>
<th>Proportion Formula</th>
<th>Percent Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{\text{Occurrence}}{\text{Total}} )</td>
<td>( \frac{\text{Occurrence}}{\text{Total}} \times 100 )</td>
</tr>
</tbody>
</table>

Related Glossary Terms

Applied Stat: Percentage, Relative frequency
Applied Stat: Proportion-1 sample

Compares a proportion obtained from a random sample to a population proportion. It is generally used to determine if there is a difference between what was obtained in a random sample and an established benchmark that either originated from all members of a comparable population or is assumed to represent the population.

Assumptions

- Independent random sampling
- Nominal level data
- Large sample size

Probability Distribution

Use z-distribution table

Formula

\[ Z = \frac{p_s - p_u}{s_p} \]

\[ p_s = \text{sample proportion} \]

\[ p_u = \text{population proportion} \]

Where

**Standard Error of the proportion**

\[ q_u = 1 - p_u \]

\[ n = \text{sample size} \]

\[ s_p = \sqrt{\frac{p_u q_u}{n}} \]

Software Output Example: Comparing historical proportion of clients reporting poor service and the current proportion of clients reporting poor service.

One-Sample Difference of Proportions

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Proportion</td>
<td>0.10</td>
</tr>
<tr>
<td>Sample Proportion</td>
<td>0.15</td>
</tr>
<tr>
<td>Sample Size (n)</td>
<td>110</td>
</tr>
<tr>
<td>Test statistic</td>
<td>1.748</td>
</tr>
<tr>
<td>( p &lt; 0.0805 ) (2tailed) [a]</td>
<td></td>
</tr>
</tbody>
</table>

Interpretation

\[ \text{[a]} \quad \text{For a 2-tailed test, the p-value represents the probability of making a type 1 error (concluding there is statistical significance when there is none). Since there is about 8% chance of making a type 1 error, which exceeds the 5% error limit established in the rejection criteria of the hypothesis testing process (alpha), you would not conclude there is statistical significance.} \]

Note: The above example uses StatCalc output.

**Related Glossary Terms**

Applied Stat: Proportion-2 sample
Applied Stat: Proportion-2 sample

Compares two proportions from two random samples; such as males and females, Republicans and Democrats, Urban and Rural residents. If there are more than two comparable groups, chi-square is a more appropriate technique.

**Assumptions**

- Independent random sampling
- Nominal level data
- Large sample size

**Probability Distribution**

Use z-distribution

**Formula**

\[ Z = \frac{p_1 - p_2}{\sqrt{pq/n_1 + n_2}} \]

\[ p_1 = \text{Proportion from sample one} \]

\[ p_2 = \text{Proportion from sample two} \]

**Standard Error of the difference in proportions**

\[ s_{p_{12}} = \sqrt{pq/n_1 + n_2} \]

\[ p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \]

\[ q = 1 - p \]

**Software Output Example:** Comparing the difference between the proportion of students in school A not eating breakfast before coming to school and the proportion of students in school B not eating breakfast before coming to school.

Two-Sample Difference of Proportions (2-tailed)

<table>
<thead>
<tr>
<th></th>
<th>School A</th>
<th>School B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Proportion</td>
<td>0.23</td>
<td>0.07</td>
</tr>
<tr>
<td>Sample Size (n)</td>
<td>80</td>
<td>180</td>
</tr>
<tr>
<td>Z-statistic</td>
<td>3.674</td>
<td></td>
</tr>
<tr>
<td>p &lt; Significance</td>
<td></td>
<td>0.0002  [a]</td>
</tr>
</tbody>
</table>

**Interpretation**

[a] For a 2-tailed test, the p-value represents the probability of making a type 1 error (concluding there is statistical significance when there is none). Since there is far less than a 5% chance of making a type 1 error, you would conclude there is statistical significance between the schools in the proportion of children who need nutrition support at school.

**Note:** The above example uses StatCalc output.

**Related Glossary Terms**

Applied Stat: Proportion-1 sample
Interval estimation (margin of error) uses sample data to determine a range (interval) that, at an established level of confidence, is expected to contain the population proportion.

Steps

Determine the confidence level (alpha is generally .05)

Use the z-distribution table to find the critical value for a 2-tailed test given the selected confidence level (alpha)

Estimate the standard error of the proportion

\[ \sqrt{\frac{pq}{n}} \]

where

\( p \) = sample proportion

\( q = 1 - p \)

Estimate the confidence interval

\[ CI = p \pm (CV)(\sqrt{\frac{pq}{n}}) \]

Interpret

Based on alpha .05, you are 95% confident that the proportion in the population from which the sample was obtained is between __ and __.

Note: Given the sample data and level of error, the confidence interval provides an estimated range of proportions that is most likely to contain the population proportion. The term "most likely" is measured by alpha (i.e., in most cases there is a 5% chance -- alpha .05 -- that the confidence interval does not contain the true population proportion).

More About the Standard Error of the Proportion

The standard error of the proportion will vary as sample size and the proportion changes. As the standard error increases, so will the margin of error.

Sample Size (n)

<table>
<thead>
<tr>
<th>Proportion (p)</th>
<th>100</th>
<th>300</th>
<th>500</th>
<th>1000</th>
<th>5000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>0.03</td>
<td>0.017</td>
<td>0.009</td>
<td>0.004</td>
<td>0.002</td>
<td>0.001</td>
</tr>
<tr>
<td>0.8</td>
<td>0.04</td>
<td>0.023</td>
<td>0.015</td>
<td>0.008</td>
<td>0.004</td>
<td>0.002</td>
</tr>
<tr>
<td>0.7</td>
<td>0.05</td>
<td>0.029</td>
<td>0.019</td>
<td>0.011</td>
<td>0.005</td>
<td>0.003</td>
</tr>
<tr>
<td>0.6</td>
<td>0.04</td>
<td>0.025</td>
<td>0.017</td>
<td>0.011</td>
<td>0.005</td>
<td>0.003</td>
</tr>
<tr>
<td>0.5</td>
<td>0.04</td>
<td>0.023</td>
<td>0.017</td>
<td>0.011</td>
<td>0.005</td>
<td>0.003</td>
</tr>
<tr>
<td>0.4</td>
<td>0.03</td>
<td>0.019</td>
<td>0.013</td>
<td>0.008</td>
<td>0.004</td>
<td>0.002</td>
</tr>
<tr>
<td>0.3</td>
<td>0.02</td>
<td>0.017</td>
<td>0.010</td>
<td>0.006</td>
<td>0.003</td>
<td>0.001</td>
</tr>
<tr>
<td>0.2</td>
<td>0.02</td>
<td>0.017</td>
<td>0.010</td>
<td>0.006</td>
<td>0.003</td>
<td>0.001</td>
</tr>
<tr>
<td>0.1</td>
<td>0.01</td>
<td>0.013</td>
<td>0.008</td>
<td>0.004</td>
<td>0.002</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Effect of changes in the proportion

As a proportion approaches .5 the error will be at its greatest value for a given sample size. Proportions close to 0 or 1 will have the lowest error.

The error above a proportion of .5 is a mirror reflection of the error below a proportion of .5.

Effect of changes in sample size

As sample size increases the error of the proportion will decrease for a given proportion. However, the reduction in error of the proportion as sample size increases is not constant. As an example, at a proportion of 0.9, increasing the sample size from 100 to 300 cut the standard error by about half (from 0.03 to 0.019). Increasing the sample size by another 200 only reduced the standard error by about one quarter (0.019 to 0.013).

Example:  Interval Estimation for Proportions

Problem: A random sample of 500 employed adults found that 23% had traveled to a foreign country. Based on these data, what is your estimate for the entire employed adult population?

\( n = 500, \ p = 0.23, \ q = 0.77 \)

Use alpha .05 (i.e., the critical value is 1.96)

Estimate Sampling Error

\[ \sqrt{\frac{pq}{n}} \]

\( CV = 1.96 \)

Compute Interval

\[ CI = p \pm CV(\sqrt{\frac{pq}{n}}) \]

Interpret

You are 95% confident that the actual proportion of all employed adults who have traveled to a foreign country is between 19.3% and 26.7%.

Software Output Example: Summary data from the previous example

Margin of Error for Proportions

<table>
<thead>
<tr>
<th>n</th>
<th>500</th>
<th>Proportion = 0.23</th>
<th>Standard Error = 0.0188</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>0.0311</td>
<td>0.1989</td>
<td>0.2611</td>
</tr>
<tr>
<td>95%</td>
<td>0.0369</td>
<td>0.1931</td>
<td>0.2669</td>
</tr>
<tr>
<td>99%</td>
<td>0.0486</td>
<td>0.1814</td>
<td>0.2786</td>
</tr>
</tbody>
</table>

Note: The above example uses StatCalc output.
Applied Stat: Rate

The number of occurrences of a trait divided by the number of possible occurrences per an established unit of time.

Rate Example: Create death rate per 1000 given there are 100 deaths a year in a population of 10,000.

\[
\text{Rate Formula} \quad \frac{\text{No. Per Year}}{\text{Population}} \times 1000
\]

\[
\frac{100}{10,000} \times 1000
\]

10 deaths per 1000 people

Related Glossary Terms
Applied Stat: Ratio
Applied Stat: Ratio

The number of cases in one category divided by the number of cases in another category.

Ratio = \( \frac{f_1}{f_2} \) or frequency in one group (larger group) divided by the frequency in another group.

Ratio Example: Your community has 1370 Protestants and 930 Catholics

\[
\frac{1370}{930} = 1.47 \text{ or for every one Catholic there are 1.47 Protestants in the given population.}
\]
Applied Stat: Regression-Logistic

Used for multiple regression modeling when the dependent variable is nominal level data. Binomial logistic is used when the dependent variable is a dichotomy (0,1). Multinomial logistic regression is used when there are more than two nominal categories.

Assumptions

See multiple regression for a more complete list of ordinary least squares (OLS) regression assumptions. In addition to the key characteristic that the dependent variable is discrete rather than continuous, the following assumptions for logistic regression differ from those for OLS regression:

1. Specification error - Linearity

Linearity is not required in logistic regression between the independent and dependent variables. It does require that the logits of the independent and dependent variables are linear. If not linear, the model may not find statistical significance when it actually exists (Type II error).

2. Large Sample Size

Unlike OLS regression, large sample size is needed for logistic regression. If there is difficulty in converging on a solution, a large number of iterations, or very large regression coefficients, there may be insufficient sample size.

Software Output Example: Evaluating the affect of age, education, sex, and race on whether or not a person voted in the presidential election.

--------------- Logistic Regression ---------------

Dependent Y: VOTE92

Cases with Y=0:  417  No, did not vote
Cases with Y=1:  1028  Yes, voted
Total Cases:  1445

Explanatory Model

VOTE92 = Constant + AGE + EDUC + FEMALE + WHITE + e

Model Fit Statistics

Initial -2 Log Likelihood 1734.531
Model -2 Log Likelihood 1594.154
Cox & Snell Rsq 0.111
Nagelkerke Rsq 0.125
Model Chi-square 194.380 df 6 pc 0.0000 [0]

Coefficients

Variable b Coeff. SE Wald p < OR
Constant -4.1620 0.4184 98.9755 0.0000
AGE 0.0335 0.0040 69.1386 0.0000 1.033
EDUC 0.2809 0.0244 132.4730 0.0000 1.324
FEMALE -0.1249 0.1269 0.9686 0.3250 0.883
WHITE 0.1000 0.1671 0.3581 0.5495 1.105

Estimated Model

VOTE92 = -4.1620 + .0335(AGE) + .2809(EDUC) + -.1249(FEMALE) + .1000(WHITE) + e

Classification Table (.50 cutpoint):

<table>
<thead>
<tr>
<th>Observed Y</th>
<th>Predicted Y</th>
<th>1</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100</td>
<td>316</td>
<td>24.70</td>
</tr>
<tr>
<td>1</td>
<td>44</td>
<td>964</td>
<td>83.75</td>
</tr>
<tr>
<td>Total</td>
<td>144</td>
<td>1274</td>
<td>72.44</td>
</tr>
</tbody>
</table>

Interpretation

[a] This ranges from 0 to less than 1 and indicates the explanatory value of the model. The closer the value is to 1 the greater the explanatory value. Analogous to R2 in OLS regression.

[b] Indicates explanatory value of the model. Ranges from 0 to 1 and adjusts for the number of independent variables. Analogous to Adjusted R2 in OLS regression. In this model, the explanatory value is relatively low.

[c] Tests the combined significance of the model in explaining the dependent (response) variable. The significance test is based on the Chi-Square sampling distribution. Analogous to the F-test in OLS regression. The independent variables in this model have a statistically significant combined effect on voting behavior.

[d] A one unit increase in age (1 year) results in an average increase of .0335 in the log odds of vote equaling 1 (voted in election).

[e] The test statistic for the null hypothesis that the b coefficient is equal to 0 (no effect). It is calculated by (b/SE)^2.

[f] Based on the Chi-Square sampling distribution, there is a statistically significant relationship between age and voting behavior. The probability that one will vote increases with age, holding education, sex, and race constant.

[g] OR/Exp(OR) Ratio. The odds of a change in the dependent variable (vote) given a one unit change in age is 1.03. Note: if the OR is >1, odds increase; if OR <1, odds decrease; if OR =1, odds are unchanged by this variable. In this example, the odds of voting increases with a one unit increase in age.

[h] Of those who did not vote, the model correctly predicted non-voting in 24.7% of the non-voters included in the sample data.

[i] Of those who did vote, the model correctly predicted voting in 83.7% of the voters included in the sample data.

[j] Of those who did or did not vote, the model correctly predicted 72.4% of the voting decisions.

Related Glossary Terms

Drag related terms here
An extension of the simple (bivariate) regression model that allows us to incorporate more than one independent variable into our explanation of \( Y \). It helps clarify the relationship between each independent variable and \( Y \) by holding constant the effects of the other independent variables.

**Assumptions**

1. No measurement error
2. No specification error
3. Mean of errors equals zero
4. Error term is normally distributed
5. Homoskedasticity
6. No autocorrelation
7. No Multicolinearity

**Software Output Example:** A multiple regression example regressing individual income on education in years, work experience in months, and a dummy variable representing sex of an individual (0=Female, 1=Male).

---

**Explanatory Model**

\[
\text{INCOME} = \text{Constant} + \text{EDUC} + \text{WORKEXP} + \text{SEX} + e
\]

**Model Fit Statistics**

<table>
<thead>
<tr>
<th>Cases(n)</th>
<th>R</th>
<th>R Square</th>
<th>Adj. R sq</th>
<th>SE Est.</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>0.865</td>
<td>0.748</td>
<td>0.721 [a]</td>
<td>2346.734 [b]</td>
</tr>
</tbody>
</table>

**ANOVA Statistics**

<table>
<thead>
<tr>
<th>Sum of Sqs</th>
<th>df</th>
<th>Mean Sg</th>
<th>F</th>
<th>p &lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>419605192.680</td>
<td>3</td>
<td>139868397.560</td>
<td>27.709</td>
</tr>
<tr>
<td>Residual</td>
<td>141339012.195</td>
<td>28</td>
<td>5047821.864</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>560944204.875</td>
<td>31</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Coefficients**

<table>
<thead>
<tr>
<th>Variable</th>
<th>b Coeff.</th>
<th>std. error</th>
<th>BetaWgt</th>
<th>t</th>
<th>p &lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-338.105</td>
<td>2027.916</td>
<td>-0.167</td>
<td>0.8688</td>
<td></td>
</tr>
<tr>
<td>EDUC</td>
<td>870.146</td>
<td>108.506</td>
<td>0.892</td>
<td>8.019</td>
<td>0.0000 [d]</td>
</tr>
<tr>
<td>WORKEXP</td>
<td>193.701</td>
<td>82.889</td>
<td>0.262</td>
<td>2.337</td>
<td>0.0268 [e]</td>
</tr>
<tr>
<td>SEX</td>
<td>2821.218</td>
<td>803.622</td>
<td>0.336</td>
<td>3.511</td>
<td>0.0015 [f]</td>
</tr>
</tbody>
</table>

**Estimated Model**

\[
\text{INCOME} = -338.105 + 870.146(\text{EDUC}) + 193.702(\text{WORKEXP}) + 2821.218(\text{SEX})
\]

**Interpretation**

[a] Adjusted \( R^2 \). Education, experience, and a person’s sex explain 72% of the variation in individual income.

[b] Standard error of the estimate. Used for creating confidence intervals for predicted \( Y \).

[c] The regression model is statistically significant.

[d] The impact of a one-unit change in education on income. One additional year of education will on average result in an increase in income of \$870 holding experience and sex constant.

[e] Standardized regression partial slope coefficients. Used only in multiple regression models (more than one independent variable), beta-weights in the same regression model can be compared to one another to evaluate relative effect on the dependent variable. Beta-weights will indicate the direction of the relationship (positive or negative). Regardless of whether negative or positive, larger beta-weight absolute values indicate a stronger relationship. In this model, education has the greatest effect on income followed in order by sex and experience.

[f] One additional month of experience will on average result in an increase in income of \$194 holding education and sex constant.

[g] Males will on average earn an income of \$2,821 more than women holding education and experience constant.

Related Glossary Terms

- Applied Stat: Regression-OLS Simple, Bias, Multicolinearity, Multiple-correlation coefficient
- Partial slope, Principle of least squares, Slope (b), Standardized partial slopes
Simple (bivariate) regression involves evaluating the linear relationship between one independent variable (X) and one dependent variable (Y) by fitting a line to a scatter of data points. If there is a relationship between X and Y, knowing the value of X will help explain and predict variations in Y.

**Assumptions**

See multiple regression for a complete list of assumptions. Basic theoretical assumptions for simple regression are the following (note that in practice it is difficult to find a perfect regression model that fully satisfies all the assumptions):

1. Both the independent (X) and the dependent (Y) variables are interval or ratio data.
2. There is a linear relationship between X and Y (confirm with scattergram).
3. Errors in prediction Y are normally distributed.
4. Errors in prediction Y are all independent of each other.
5. The distribution of the errors in prediction of the value of Y is constant regardless of the value of X.

**Formula**

*Equation for a straight line*

\[
\hat{Y} = a + bx
\]

- \(\hat{Y}\) = predicted score
- \(b\) = slope of a regression line (regression coefficient)
- \(x\) = individual score from the X distribution
- \(a\) = Y intercept (regression constant)

\[
a = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \left(\frac{\sum x}{n}\right)^2}
\]

\[
b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \left(\frac{\sum x}{n}\right)^2}
\]

- \(x\) = mean of variable X
- \(\bar{Y}\) = mean of variable Y

**Software Output Example:** Regressing the effect of education on individual income.

**Model Fit Statistics**

<table>
<thead>
<tr>
<th>Cases(n)</th>
<th>R</th>
<th>R Square</th>
<th>Adj Rsq</th>
<th>SE Est.</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>0.751</td>
<td>0.564</td>
<td>0.550</td>
<td>2854.601</td>
</tr>
</tbody>
</table>

**ANOVA Statistics**

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Sq</th>
<th>df</th>
<th>Mean Sq</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>316481758.400</td>
<td>1</td>
<td>316481758.400</td>
<td>38.838</td>
<td>0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>244462446.475</td>
<td>30</td>
<td>8148748.216</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>560944204.875</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Coefficients**

<table>
<thead>
<tr>
<th>Variable</th>
<th>b Coeff.</th>
<th>std. error</th>
<th>BetaWgt</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>5077.512</td>
<td>1497.830</td>
<td>3.390</td>
<td>3.390</td>
<td>0.0020</td>
</tr>
<tr>
<td>EDUC</td>
<td>732.400</td>
<td>117.522</td>
<td>0.751</td>
<td>6.232</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

**Estimated Model**

\[ \text{INCOME} = 5077.513 + 732.400(\text{EDUC}) \]

**Interpretation**

[a] The correlation coefficient.
[b] Education explains 56% of the variation in income.
[c] Standard error of the estimate. Used for creating confidence intervals for predicted Y.
[d] The regression model is statistically significant.
[e] The Y intercept and mean value of income if education is equal to 0.
[f] The impact of a one-unit change in education on income. Interpretation: One additional year of education will on average result in an increase in income of $732.
[g] Standard error of the slope coefficient. Used for creating confidence intervals for b and for calculating a t-statistic for significance testing.
[h] Standardized regression partial slope coefficients. Used only in multiple regression models (more than one independent variable), beta-weights in the same regression model can be compared to one another to evaluate relative effect on the dependent variable. Beta-weights will indicate the direction of the relationship (positive or negative). Regardless of whether negative or positive, larger beta-weight absolute values indicate a stronger relationship.
[i] The t-test statistic. Calculated by dividing the b coefficient by the standard error of the slope.
[j] Education has a statistically significant positive association with income.
Association

Changes in one variable are accompanied by changes in another variable. Finding associations is one of the major objectives of statistical analysis. The question answered by association is whether knowledge of one set of data allows us to infer or predict the characteristics of another set of data. Descriptive associations include those relationships identified between two sets of observations (e.g., sex and income) in nature. No conclusions about causality can be made based solely on data that shows statistical association between variables.

Related Glossary Terms
Autocorrelation

The error terms are correlated across observations in a multiple regression model. Violation of this assumption is likely to be a problem with time-series data where the value of one observation is not completely independent of another observation. (Example: A simple two-year time series of the same individuals is likely to find that a person's income in year 2 is correlated with their income in the prior year.) If there is autocorrelation, the parameter estimates for partial slopes and the intercept are not biased but the standard errors are biased and hence significance tests may not be valid.

Diagnosis

Suspect autocorrelation with any time series/longitudinal data

Use the Durbin-Watson (d) statistic.

\[
\begin{align*}
d=2 & \quad \text{no correlation between error terms} \\
d=0 & \quad \text{perfect positive correlation between error terms} \\
d=4 & \quad \text{perfect negative correlation between error terms}
\end{align*}
\]

Correction:

Use generalized least squares (GLS) or weighted least squares (WLS) models.

Related Glossary Terms

Regression model
Bar chart

A graphic display used for discrete variables. Each category is represented by a bar of equal width. The length of each bar represents the number or percentage of observations in that category.

Example of a Horizontal Column Chart

![Bar chart example]

Related Glossary Terms

Column Chart, Frequency polygon, Pie chart
Beta-weight

Standardized regression partial slope coefficients. Beta-weights indicate the amount of change in the standardized scores of Y (dependent variable) for a one unit change in the standardized scores of each independent variable, controlling for the effects of all independent variables included in the regression model. Beta-weights in the same regression model can be compared to one another to evaluate relative effect on the dependent variable. Beta-weights will indicate the direction of the relationship (positive or negative). Regardless of whether negative or positive, larger beta-weight absolute values indicate a stronger relationship.

Partial slope coefficients are not directly comparable to one another. This is because each variable is often based on a different metric. Beta weights (standardized slope coefficients) are used to compare the relative importance of regression coefficients in a model. They are not useful for comparisons among models from other samples.

Formula:

\[ \beta_i = b_i \left( \frac{s_{Y_i}}{s_y} \right) \]

or partial slope * (std dev. of Xi ÷ std dev. of Y)

Example:

Education partial slope coefficient \((b_i) = 1878.2\)

Education standard deviation \(s_{x_i} = 2.88\)

Begin salary standard deviation \(s_y = 7870.64\)

\[ \beta_i = b_i \left( \frac{s_{x_i}}{s_y} \right) \]

\[ \beta_i = 1878.2 \left( \frac{2.88}{7870.64} \right) \]

\[ \beta_i = .687 \]

This is interpreted as the average standard deviation change in Y (salary) associated with one standard deviation change in X (education), when all other independent variables are held constant. Or a one standard deviation change in education results in an average change of .687 standard deviations in income.

Related Glossary Terms

Regression model

Index
Bias

A misrepresentation of what is being measured. Bias can be the result of many factors to include measurement device or technique. Estimates of population parameters are sometimes referred to as either biased or unbiased. A statistic is biased if, in the long run, it consistently over or underestimates the parameter it is estimating.

Related Glossary Terms
Applied Stat: Regression-OLS Multiple, Interpretation bias, Regression model
Biased estimate

A sample statistic that tends to over or underestimate the true population value.

Related Glossary Terms
Regression model
Binomial variable

A variable that has only two values or attributes. Sex is a common example (male or female).

Related Glossary Terms
Drag related terms here
Bivariate normal distributions

This assumption specifies that two comparison groups must be normally distributed.

Related Glossary Terms

Drag related terms here
Bivariate table

A table that displays the joint frequency distributions of two variables. Generally referred to as a contingency table or crosstabulation. By convention, the independent variable is usually represented in the columns and the dependent variable is represented in the rows.

Example:

Crosstabulation: Manager (Rows) by Sex (Columns)

Column Variable Label: Respondent sex
Row Variable Label: In Management Position?

<table>
<thead>
<tr>
<th>Count</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row %</td>
<td>58.33</td>
<td>41.67</td>
<td></td>
</tr>
<tr>
<td>Col %</td>
<td>84.00</td>
<td>71.43</td>
<td>78.26</td>
</tr>
<tr>
<td>Total %</td>
<td>45.65</td>
<td>32.61</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No</th>
<th>21</th>
<th>15</th>
<th>36</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>58.33</td>
<td>41.67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>84.00</td>
<td>71.43</td>
<td>78.26</td>
</tr>
<tr>
<td></td>
<td>45.65</td>
<td>32.61</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Yes</th>
<th>4</th>
<th>6</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40.00</td>
<td>60.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16.00</td>
<td>28.57</td>
<td>21.74</td>
</tr>
<tr>
<td></td>
<td>8.70</td>
<td>13.04</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>25</th>
<th>21</th>
<th>46</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>54.35</td>
<td>45.65</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Related Glossary Terms

Causation

Causation represents a direct stimulus-response link between two variables. Since causation is a logical assertion and cannot be proven with statistics, statistical techniques are best used to explore (not prove) connections between independent and dependent variables. The following logical conclusions are a minimum requirement for concluding causation.

Association: Do the variables covary empirically? Strong associations are more likely to be causal than weak associations.

Precedence: Does the independent variable vary before the effect exhibited in the dependent variable?

Nonspuriousness: Can the empirical correlation between two variables be explained away by the influence of a third variable?

Plausibility: Is the expected outcome plausible and consistent with theory, prior knowledge, and other studies?

Related Glossary Terms

Association
Central limit theorem

As sample size increases, the sampling distribution of means approximates a normal distribution and is usually close to normal at a sample size of 30.

Software Output Example: Skewed income distribution for a population of 1,614 families.

Descriptive Statistics
Variable: REALINC FAMILY INCOME IN CONSTANT $

<table>
<thead>
<tr>
<th>Count</th>
<th>1614</th>
<th>Pop Var</th>
<th>297864464.1298</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum</td>
<td>37950010.0000</td>
<td>Sam Var</td>
<td>298051130.2576</td>
</tr>
<tr>
<td>Mean</td>
<td>23513.0173</td>
<td>Pop Std</td>
<td>17256.8083</td>
</tr>
<tr>
<td>Median</td>
<td>18375.0000</td>
<td>Sam Std</td>
<td>17264.1574</td>
</tr>
<tr>
<td>Min</td>
<td>245.0000</td>
<td>Std Error</td>
<td>429.7280</td>
</tr>
<tr>
<td>Max</td>
<td>68600.0000</td>
<td>CV%</td>
<td>7.4238</td>
</tr>
<tr>
<td>Range</td>
<td>68555.0000</td>
<td>95% CI (+/-)</td>
<td>842.8885</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.7979</td>
<td>t-test(mu=0)</td>
<td></td>
</tr>
<tr>
<td>Kurtosis</td>
<td>2.8317</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Software Output Example: Simulation of taking a random sample of 30 observations from the population 1,000 times and calculating a mean from each random sample.

Descriptive Statistics
Variable: Mean

<table>
<thead>
<tr>
<th>Count</th>
<th>1000</th>
<th>Pop Var</th>
<th>10491553.2947</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum</td>
<td>19247808.4165</td>
<td>Sam Var</td>
<td>10502055.3500</td>
</tr>
<tr>
<td>Mean</td>
<td>19247.8084</td>
<td>Pop Std</td>
<td>3239.0667</td>
</tr>
<tr>
<td>Median</td>
<td>19230.4583</td>
<td>Sam Std</td>
<td>3240.6875</td>
</tr>
<tr>
<td>Min</td>
<td>7231.5833</td>
<td>Std Error</td>
<td>102.4795</td>
</tr>
<tr>
<td>Max</td>
<td>30788.3333</td>
<td>CV%</td>
<td>16.8367</td>
</tr>
<tr>
<td>Range</td>
<td>23556.7500</td>
<td>95% CI (+/-)</td>
<td>201.0998</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.0753</td>
<td>t-test(mu=0)</td>
<td></td>
</tr>
<tr>
<td>Kurtosis</td>
<td>3.1711</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Related Glossary Terms
Independent random sample, Theoretical frequency distribution
Central tendency

A single summary value that suggests a representative measure of a set of observations. Measures of central tendency include mean, median, and mode.

Related Glossary Terms
Mean, Median, Mode
Codebook

A document that tells the location and meaning of variables and values in a data file.

Related Glossary Terms
Data Dictionary
Coefficient of determination

Represents the proportion of variation in the dependent variable (Y) explained by the independent variable (X). In the case of Pearson r, the coefficient of determination can be obtained by squaring the Pearson r coefficient. See also the coefficient of multiple determination.

Related Glossary Terms

Coefficient of multiple determination (Rsq)
Coefficient of multiple determination (Rsq)

A statistic that represents the proportion of total variation explained in the dependent variable by all independent variables included in the model. Multiply by 100 to produce a percentage of the explained variation.

Related Glossary Terms
Coefficient of determination
Coefficient of variation (CV)

The Coefficient of Variation (CV) is the ratio of the sample standard deviation to the sample mean: \( \frac{\text{sample standard deviation}}{\text{sample mean}} \times 100 \) to calculate CV\%. Used as a measure of relative variability, CV is not affected by the units of a variable. CV is useful for comparing the variation in two series of data which are measured in two different units. Examples include a comparison between variation in height and variation in weight or comparing different experiments involving the same units of measure but conducted by different persons.

Related Glossary Terms
Drag related terms here
Column Chart

A graphic display used for discrete variables. Each category is represented by a bar of equal width. The height of each bar represents the number or percentage of observations in that category.

Example of a Vertical Column Chart

![Example of a Vertical Column Chart](image)

Related Glossary Terms

Bar chart
Concept

An idea that represents a set of objects. Concepts are sometimes referred to as abstract concepts in that they are idealizations that may not be measured or observed directly.
Conceptual model

This is a construct system that connects our view of the world with data. It connects theory, hypotheses, and data with our vision of reality.

Related Glossary Terms

Theory
Confidence interval

An interval of values within which we can state with a degree of confidence that the true population parameter falls. Used in conjunction with the point estimate. Also known as the interval estimate.

Related Glossary Terms
Applied Stat: Mean-Confidence Interval, Applied Stat: Proportion-Confidence Interval, Confidence level, Confidence limits, Interval estimation, Margin of error, Point estimation
Confidence level

An alternative representation of alpha. It represents the probability that an interval estimate will contain the true population value.

Related Glossary Terms
Confidence interval, Constant
Confidence limits

The boundary values that contain the interval estimate. An interval estimate is an interval of values within which we can state with a degree of confidence that the true population parameter falls. Used in conjunction with the point estimate. Also known as the confidence interval and is often expressed as a margin of error.

Related Glossary Terms

Applied Stat: Mean-Confidence Interval, Applied Stat: Proportion-Confidence Interval, Confidence interval
Constant

When an object assumes only one value it is a constant. If an object is not a constant, it is a variable. We sometimes refer to constants as controls. As an example, a study of only female patients would have the constant female.

Related Glossary Terms
Confidence level
Construct validity

The degree to which one measure correlates with other measures of the same abstract concept.

Related Glossary Terms

Content validity, Criterion-related validity, External validity, Face validity, Internal validity, Predictive validity, Validity
Content validity

The extent to which the indicator reflects the full domain of interest.

Related Glossary Terms

Construct validity, Criterion-related validity, External validity, Face validity, Internal validity, Predictive validity, Validity
Contingency table

Used to represent the relationship between two variables. The variables are usually nominal or ordinal level and the relationship is described by counts and row and column percent. Also referred to as a crosstabulation or crosstab.

Related Glossary Terms
Continuous variable

A measure that can take on any value within a given interval or set of intervals. An infinite number of possible values such as distance in kilometer and loan interest rates.

Related Glossary Terms

Drag related terms here
Control variable

A variable that is held to one value to help clarify the relationship between other variables. As an example, sex may be controlled to investigate the relationship between education level and income (i.e., a separate analysis for males and females).

Related Glossary Terms
Partial tables, Statistical control
Covariation

Changes in one variable are accompanied by changes in another variable. Also known as association. This is one of the major objectives of statistical analysis. The question answered by association is whether knowledge of one set of data allows us to infer or predict the characteristics of another set of data. Descriptive associations include those relationships identified between two sets of observations (e.g., sex and income) in nature. Sometimes called correlational, no conclusions about causality can be made based solely on the data. Experimental relationships are experiment based which involves the ability of researchers to manipulate the levels of one variable to observe changes in another variable.

Related Glossary Terms

Applied Stat: Correlation, Association
Covary

Changes in one variable are accompanied by changes in another variable. Also known as association.

Related Glossary Terms
Applied Stat: Correlation, Association
Criterion-related validity

The degree to which one measure is similar to another measure. An example would be high school GPA and graduate rate in college.

Related Glossary Terms
Construct validity, Content validity, Predictive validity, Validity
Critical value

The point on the x-axis of a sampling distribution that is equal to alpha. It is interpreted as standard error. As an example, a critical value of 1.96 is interpreted as 1.96 standard errors above the mean of the sampling distribution.

Related Glossary Terms
Drag related terms here
Cross-sectional survey

Observations collected at one point in time.

Related Glossary Terms
Drag related terms here
Cumulative frequency

A cumulative frequency conveys how many observations are contained above or below a value of a variable. Generally described as a cumulative frequency distribution that reports the number of observations that occur up to a given value. Most statistical software provides this by default when you run a frequency procedure.

Related Glossary Terms

Applied Stat: Frequencies, Cumulative percentage, Deciles, Frequency distribution
Cumulative percentage

The percentage of cases within an interval and all preceding intervals. Generally displayed in a frequency distribution table.

Related Glossary Terms
Applied Stat: Frequencies, Cumulative frequency, Deciles
Data

Observations of our environment that are used to answer questions such as how much, how many, how long, how often, how fast, where, and what kind. These observations are often represented by numbers. The importance of numbers is that they have less ambiguity than words as to their meaning. Data are generally referred to in the plural.

Related Glossary Terms
Raw score, Secondary analysis
Data Dictionary

A document that tells the meaning of variables and values in a data file.

Software Output Example: Data from survey of U.S. Adults.

FILE DATA DICTIONARY

File Name: Workbook.dcs
File Size: 1,877 Bytes
Created: 2007-03-15 17:56:08
Created by: AcaStat Software

Variable Name: Idnum
Variable Label: Respondent Number
Value Labels: None
Missing Values: None

Variable Name: Age
Variable Label: Respondent age in years
Value Labels: None
Missing Values: None

Variable Name: Edu
Variable Label: Years of education
Value Labels: None
Missing Values: None

Variable Name: Sex
Variable Label: Respondent sex
Value Labels: 1 = Male
2 = Female
Missing Values: None

Variable Name: Manager
Variable Label: In Management Position?
Value Labels: 1 = No
2 = Yes
Missing Values: 9

Variable Name: JobSat
Variable Label: Satisfied with job
Value Labels: 0 = No
1 = Yes
Missing Values: None

Variable Name: Income
Variable Label: Respondent individual income
Value Labels: None
Missing Values: None

Variable Name: EduCat
Variable Label: Education Level
Value Labels: 1 = <12 yrs
2 = HS Grade
3 = College
Missing Values: None

Related Glossary Terms

Codebook
Data reduction

Summarizing data into generalized measures based on numerous values of raw data. Statistical analysis reduces large data sets that describe original (raw) data into measures that lose some detail but provide simplicity in interpretation.
Deciles

The points that divide a distribution of scores into 10ths.

Related Glossary Terms
Cumulative frequency, Cumulative percentage
Deduction

Expectations are based on general principles instead of observations.

Related Glossary Terms

Theory
Degrees of freedom

The number of observations in the data that are free to vary after the sample statistics have been calculated. Once the second to last calculation for a statistic has been completed, the final result is automatically determined without calculation. This last value is not free to vary, so generally one is subtracted from the count of observations (n-1).

Related Glossary Terms
Drag related terms here
Dependent variable

A measure not under the control of the researcher that reflects responses caused by variations in another measure (the independent variable). Also called the response variable.

Related Glossary Terms

Antecedent variable, Intervening variable, Response variable
Descriptive analysis

A direct and exhaustive measure of a complete group. A descriptive analysis can be a description of an entire population or a sample. It becomes an inferential analysis if you choose to use sample descriptors to describe a larger underlying population. To successfully describe the underlying population, you must base your measures on a random sample of that population.

Related Glossary Terms
Descriptive statistic
Descriptive statistic

A statistic that classifies and summarizes numerical data from a sample.

Related Glossary Terms
Descriptive analysis, Interval data
Dichotomous variable

This is a variable that is classified by two values. These variables are also referred to as binary variables or dummy variables. Dummy variables are qualitative variables that measure presence or absence of a characteristic. As an example, a dummy variable representing the characteristic "male" could be represented as 0=female (non-male) and 1=male.

Related Glossary Terms
Discrete variable, Dummy variable
Discrete variable

A variable that is limited to a finite number of values. Such as religion or number of parks in a city (you can't have 1.5 parks).

Related Glossary Terms

Dichotomous variable
Dispersion

The extent to which observations differ (vary) from one another. Also known as variation. Common summary measures of dispersion are range, variance, and standard deviation.

Related Glossary Terms

Variance
Dispersion measures

Statistics that indicate the variation in a distribution of scores. Examples include variance, standard deviation, range.
**Dummy variable**

A variable that is classified by two values. Dummy variables are qualitative variables that measure presence or absence of a characteristic. As an example, a dummy variable representing the characteristic "male" could be represented as 0=female (non-male) and 1=male. These variables are also referred to as binary variables or dichotomous variables.

---

**Related Glossary Terms**

Dichotomous variable
Ecological fallacy

Improperly drawing conclusions about individuals based entirely on the observation of groups.

Related Glossary Terms

Errors in human inquiry
Efficiency

The extent to which multiple sample outcomes are clustered around the mean of the sampling distribution. The efficiency of a statistic is the degree to which the statistic is stable from sample to sample. If one statistic has a smaller standard error than another, then the statistic with smaller standard error is the more efficient statistic.
Elaboration

A multivariate approach that expands the investigation of bivariate relationships by controlling for other variables. This approach produces a separate (partial) bivariate table (contingency or crosstabulation) for each value of a third variable.

Related Glossary Terms

Drag related terms here
Element

A term used in sampling to represent the unit of analysis. Examples include people, cities, states, etc.

Related Glossary Terms

Drag related terms here
Errors in human inquiry

*Inaccurate observation*: we tend to be casual observers. Science requires a more deliberate and conscious activity.

*Overgeneralization*: we use a few events to develop general patterns to simplify our understanding. Science uses large samples (numerous events) and replication.

*Selective observation*: we tend to pay attention to events/observations that support our understanding and exclude those that conflict. Science uses all observations and the peer review process discourages subjective opinion.

*Illogical reasoning*: individuals are comfortable with using exceptions to prove a pattern/rule. Science has rules for causation and peer review encourages honesty.

*Premature closure of Inquiry*: we are information satisficers who are comfortable basing patterns and explanations on erroneous observations. Science can suffer from the same ailment but one of the goals of science is to evaluate certain knowledge (obvious explanations of reality such as the world is flat, the universe revolves around the earth, "we have tried that before and it didn't work").

Related Glossary Terms

Ecological fallacy
Experimental design

Experimental research design uses a control group and applies a treatment to a second group. This design provides the strongest evidence of causation through extensive controls and random assignment to remove other differences between groups.

Related Glossary Terms

Quasi-experimental
External validity

The extent to which the association between the independent and dependent variable is accurate and unbiased in populations outside the study group.

Related Glossary Terms

Construct validity, Content validity, Face validity, Hawthorne effect, Quasi-experimental, Validity
Face validity

The extent to which an indicator appears to measure the abstract concept.

Related Glossary Terms
Construct validity, Content validity, External validity, Validity
Fisher's Exact Test

Examines the relationship in a 2x2 contingency tables. It is most commonly used in place of chi square when n is small (generally 30 or less). It should be used in place of chi square when the expected frequency of any cell is less than 1 or 20% of expected frequencies are less than 5.

Software Output Example: Data from survey of U.S. Adults.

Crosstabulation: Manager (Rows) by Sex (Columns)

Column Variable Label: Respondent sex
Row Variable Label: In Management Position?

<table>
<thead>
<tr>
<th>Count</th>
<th>Row %</th>
<th>Col %</th>
<th>Total %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Female</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>21</td>
<td>15</td>
<td>36</td>
</tr>
<tr>
<td>Yes</td>
<td>4</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>21</td>
<td>46</td>
</tr>
</tbody>
</table>

Chi-square

<table>
<thead>
<tr>
<th>Value</th>
<th>DF</th>
<th>p &lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>1.060</td>
<td>1</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>1.059</td>
<td>1</td>
</tr>
<tr>
<td>Yate's Correction</td>
<td>0.450</td>
<td>1</td>
</tr>
<tr>
<td>Fisher's Exact (2-tailed)</td>
<td>0.5017</td>
<td>[a]</td>
</tr>
</tbody>
</table>

Interpretation

[a] Not statistically significant.

Related Glossary Terms
Applied Stat: Counts-Chi-square
Formative evaluation

A method of judging the value of a program while the program activities are forming or happening. Formative evaluation focuses on the process (Bhola 1990).

Example: Collecting continuous feedback from participants in a program in order to revise the program as needed.

Related Glossary Terms

Summative evaluation

Index
Frequency distribution

A count of how frequently each value of a variable occurs. Counts can be displayed in table or graphical form such as a histogram.

Related Glossary Terms

Applied Stat: Frequencies, Cumulative frequency, Frequency polygon, Histogram, Kurtosis, Normal distribution, Relative frequency
Frequency polygon

A graphical presentation of a frequency distribution. The y axis (vertical) represents the counts and the x axis (horizontal) represents the values of a variable. The difference between a polygon and histogram is that a line graph is used to represent each value's count rather than a bar.

Related Glossary Terms
Bar chart, Frequency distribution, Histogram, Normal distribution, Pie chart
**Gamma**

A measure of association used in bivariate tables of two ordinal level variables. Gamma is a proportional reduction in error technique that can range from -1 (negative association) to 1 (positive association). The closer the gamma coefficient is to an absolute value of one, the stronger the association.

**Related Glossary Terms**

Applied Stat: Counts-Chi-square, Association
Hawthorne effect

A commonly cited research error. A series of experiments in the 1920s at the Hawthorne Western Electric Company found that the increases in production researchers attributed to the manipulation of lighting in a control group was actually the result of the attention they gave to the study group, not the intervention (lighting).

Related Glossary Terms
External validity, Internal validity
Heteroskedasticity

Occurs when the variance of the error term does not have constant variation across multiple values of an independent variable.

Related Glossary Terms
Regression model
Histogram

A graphical presentation of frequency distribution. The y axis (vertical) represents the counts and the x axis (horizontal) represents the values of a variable. Unlike a column or bar chart, columns or bars in a histogram are continuous intervals and flow from one to the other. A histogram will not have a gap between the bars or columns except to differentiate one interval from another.

Related Glossary Terms
Frequency distribution, Frequency polygon, Normal distribution
Homogeneity of variance

The assumption of homogeneity of variance is that the variance within each of the populations is equal. This is an assumption of t-tests and analysis of variance (ANOVA).

Related Glossary Terms
Drag related terms here
Hомоседастичность

The variance of the Y scores in a correlation are uniform for the values of the X scores. In other words, the Y scores are equally spread above and below the regression line.

Related Glossary Terms

Regression model
Hypothesis

A formal statement that presents the expected association between an independent and dependent variable. A dependent variable is a variable that contains variations for which we seek an explanation. An independent variable is a variable that is thought to affect (cause) variations in the dependent variable.

Related Glossary Terms

Alternative hypothesis
Independent random sample

Random samples collected so that the selection of a particular case for one sample has no effect on the probability that any particular case will be selected for the other samples.

Related Glossary Terms

Central limit theorem
**Independent variable**

A measure that can take on different values which are subject to manipulation by the researcher.

---

**Related Glossary Terms**

Antecedent variable, Intervening variable
Induction

General principles are based on observations.
Inference

A procedure for stating the degree of confidence that the measures we have of our environment are accurate. Inferential measures are based on less than the entire population and represent an estimate of the true but unknown value of a population characteristic. Inference depends on the concept of random sampling.

Related Glossary Terms

Inferential statistic
Inferential statistic

Statistics that use characteristics of a random sample to include sampling error to predict the true values in the larger population. Any time one uses sample data to make estimates of a population one is employing an inferential statistic. Inferential statistics are treated as estimates of population parameters.

Related Glossary Terms
Inference
Instrumental variable

For regression modeling, an instrumental variable for a specific independent variable $X_i$ is another variable that a) is correlated with $X_i$, b) has no effect on the dependent variable other than indirectly through $X_i$, and c) is not correlated with the error term in the model. The basic premise is to use instrumental variables to purge the independent variable of random error by incorporating a two-stage regression analysis.
Interaction

Occurs when the statistical significance and/or direction of a bivariate relationship varies depending on a particular category of a controlling variable.

Related Glossary Terms

Drag related terms here
**Intercept (a)**

Represents the point where the regression line intercepts the Y axis and indicates the average value of Y when X is equal to zero (0).

---

**Related Glossary Terms**

Regression model
Interdependent

Changes in one variable are accompanied by changes in another variable. This is one of the major objectives of statistical analysis. The question answered by association is whether knowledge of one set of data allows us to infer or predict the characteristics of another set of data.

Related Glossary Terms

Drag related terms here
Internal validity

The extent to which accurate and unbiased association between the independent variable and dependent variable was obtained in the study group.

Related Glossary Terms
Construct validity, Content validity, Hawthorne effect
Interpretation bias

Errors in data collection that occur when knowledge of the results of one test affect the interpretation of a second test.

Related Glossary Terms
Bias
Interquartile range (Q)

The distance representing the range of scores between the second and third quartile of a distribution of scores.

Related Glossary Terms

Drag related terms here
Interval data

Objects classified by type or characteristic, with logical order and equal differences between levels of data. The meaning between each level of data is the same. Also known as scale data.

Related Glossary Terms
Descriptive statistic, Level of measurement, Nominal data, Ordinal data, Quantitative, Ratio data
Interval estimation

An interval of values within which we can state with a degree of confidence that the true population parameter falls. Used in conjunction with the point estimate. Also known as the confidence interval and is often expressed as a margin of error.

Related Glossary Terms

Confidence interval
Intervening variable

An independent variable that comes between another independent variable and a dependent variable. An intervening variable suggests the original link between the independent and dependent variables is primarily through a third variable.

Related Glossary Terms
Antecedent variable, Dependent variable, Independent variable
Kurtosis

The peakedness of a distribution. Leptokurtic is more peaked, Mesokurtic is a normal distribution, and Platykurtic is a flatter distribution.

Related Glossary Terms
Frequency distribution
Level of measurement

The mathematical characteristic of a variable and the major criterion for selecting statistical techniques. Variables can be measured at the nominal, ordinal, interval, ratio level.

Related Glossary Terms
Interval data, Nominal data, Ordinal data, Ratio data, Scale
Likert scale

A composite measure based on a collection of items that are measured on an ordinal scale such as strongly agree to strongly disagree survey questions. Developed by Rensis Likert. Some survey questions are described as Likert questions because they have a range of ordinal responses. A true Likert scale requires that a composite measure be created from two or more scaled questions.

Related Glossary Terms

Drag related terms here
Linear

The relationship between the independent variables and the dependent variable is conducive to being fit with a straight line.

Related Glossary Terms

Positive association
Longitudinal survey

A survey that collects data at different points in time on the same objects.
Margin of error

Used to express an interval of values within which we can state with a degree of confidence that the true population parameter falls. Used in conjunction with the point estimate. Also known as the confidence interval. The margin of error is commonly expressed as plus or minus a set value that represents half of the confidence interval. As an example, 50% with a margin of error +/- 3% indicates that the true population percentage is estimated to be between 47% and 53%. There is always a chance this interval is wrong. If the margin of error is based on a 95% confidence level, this means there is a 5% chance the true population value will not fall within the margin of error.

Related Glossary Terms

Confidence interval
Marginals

The row and column subtotals (counts) reported in a bivariate contingency table.

Related Glossary Terms
Drag related terms here
Mean

The arithmetic average of the scores in a sample distribution. Mathematically represented as the sum of a set of values divided by the number of values used. The mean has an advantage over the median and mode as a measure of central tendency. The mean includes information from every value and uses this information to present a central balancing point or measure of central tendency.

\[ \bar{X} = \frac{\sum X_i}{n} \]

Example:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Age</th>
<th>Also known as X</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>24</td>
<td></td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>Mode</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>Median</td>
</tr>
<tr>
<td></td>
<td>55</td>
<td></td>
</tr>
<tr>
<td></td>
<td>60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>65</td>
<td></td>
</tr>
</tbody>
</table>

\( n = 7 \) Number of scores (or cases)

\( 310 \) Sum of scores

\( 44.29 \) Mean

Related Glossary Terms

Central tendency, Median, Mode, Mu, Weighted mean
Mean absolute deviation

This summary measure of dispersion is computed by summing the absolute values of the deviations of each value from the mean and then dividing these values by the number of observations in the data. Also known as M.A.D.

Related Glossary Terms
Drag related terms here
Measurement scale

A reflection of how well a variable and/or concept can be measured. Generally categorized in order of precision as nominal, ordinal, interval, and ratio data.

Related Glossary Terms
Applied Stat: Counts-Lambda
Median

The point on a scale of measurement below and above which fifty percent of the observations fall. To obtain the median, the observations must be rank ordered by their values.

Example:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>42</td>
</tr>
<tr>
<td>Median</td>
<td></td>
</tr>
<tr>
<td></td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>65</td>
</tr>
</tbody>
</table>

Related Glossary Terms
Central tendency, Mean, Mode
Missing data

Observations that don't report data for a variable. As an example, a survey respondent may answer several survey questions but refuse to indicate their age. For the age variable, the data are "missing" for this particular case.

Related Glossary Terms
Drag related terms here
Mode

The most frequently occurring score in a distribution. Also referred to as the modal value. It is possible to have bimodal distributions (two values with equal counts) or multimodal distributions (three or more values with equal counts).

Example:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>65</td>
</tr>
</tbody>
</table>

Mode

Related Glossary Terms
Central tendency, Mean, Median
Mu

The arithmetic average of the scores in a population.

Related Glossary Terms

Mean

Index
Multicolinearity

Occurs when one of the independent variables has a substantial linear relationship with another independent variable in a multiple regression model. It occurs in any model and is more a matter of degree of colinearity rather than whether it exists or not.

Related Glossary Terms
Applied Stat: Regression-OLS Multiple
Multiple correlation coefficient

A statistic that indicates strength of association between a dependent variable and two or more independent variables included in a regression model. Computed by taking the square root of the coefficient of multiple determination (R squared).

Related Glossary Terms
Applied Stat: Regression-OLS Multiple
**Multiplication rule of probability**

The probability of the joint occurrence of two or more outcomes across two or more events. The probability of the joint occurrence of outcome A and outcome B will be equal to the product of the probabilities of A and B.

**Example:**

Out of a deck of 52 playing cards, what is the probability of picking an Ace and a Jack from two draws (events)? Assume a well shuffled deck and mutually exclusive event possibilities (i.e., you cannot pull a Jack and an Ace at the same time).

*Multiplication Rule:* When two events, A and B, are mutually exclusive, the probability that A and B will occur is the product of the probability of each event. (The example assumes a card is not returned to the deck after selection.)

\[ P(A \text{ and } B) = P(A) \times P(B) \]

or

\[ P(\text{Ace and Jack}) = \frac{4}{52} \times \frac{4}{51} = \frac{16}{2652} \]

There is about a 0.6% chance of pulling an Ace and a Jack with two draws from a deck of playing cards.

**Related Glossary Terms**

Addition rule of probability
Mutually exclusive

No two outcomes can both occur in any given trial. As an example, males and females are mutually exclusive groups in the characteristic sex. The random selection of one individual from a population cannot (normally) produce a person who is both male and female.

Related Glossary Terms

Drag related terms here
N

Represents the number of cases (observations) in a population. Lower case "n" represents the number of cases (observations) in a sample.
Naturalistic study

A non-experimental study where relationships are simply observed as they occur or are exhibited in a natural environment (i.e., without investigator control). Common examples are social and attitudinal surveys.

Related Glossary Terms

Drag related terms here
Negative association

A relationship where two variables vary in opposite directions. As values in one variable increase, values in the other variable decrease.

Related Glossary Terms

Association
Nominal data

Objects classified by type or characteristic. There is no specific order. Examples of nominal variables include religion, sex, political affiliation, medical treatment. Nominal data are also known as nominal scale data.

Related Glossary Terms
Interval data, Level of measurement, Ordinal data, Qualitative, Ratio data
Non-orthogonal

Changes in one variable are accompanied by changes in another variable. This is one of the major objectives of statistical analysis. The question answered by association is whether knowledge of one set of data allows us to infer or predict the characteristics of another set of data.

Related Glossary Terms

Association
**Nonparametric**

A distribution free test where the assumption of a normal sampling distribution is not needed. Chi-square is considered to be a nonparametric test as are tests of medians.

---

**Related Glossary Terms**

Applied Stat: Counts-Chi-square
Normal curve

A theoretical distribution of scores that is symmetrical and unimodal (commonly called bell shaped). A standard normal curve has a mean of 0 and a standard deviation of 1.

Properties of a normal distribution. This is a theoretical construct; no population or sample will be perfectly normal.

1. Forms a symmetric bell-shaped curve.

2. 50% of the scores lie above and 50% below the midpoint of the distribution.

3. The curve is asymptotic to the x axis (i.e., never touches the x axis).

4. The mean, median, and mode are located at the midpoint of the x axis.

Related Glossary Terms

Normal distribution

Index
Normal distribution

A frequency distribution of scores that is symmetric about the mean, median, and mode. Distributions that are not symmetric are called skewed. A normal distribution is also sometimes referred to as a bell-shaped distribution.

Related Glossary Terms
Frequency distribution, Frequency polygon, Histogram, Normal curve, Skewness
Null hypothesis

This is a statement that asserts there is no difference between two population parameters. A statistically significant finding rejects the null hypothesis and supports the alternate hypothesis.

**Example:**

Null Hypothesis: *There is no statistically significant difference* between the historical proportion of clients reporting poor service and the current proportion of clients reporting poor service.

Alternate Hypothesis: *There is a statistically significant difference* between the historical proportion of clients reporting poor service and the current proportion of clients reporting poor service.

The four possible outcomes in hypothesis testing

<table>
<thead>
<tr>
<th>Actual Population Comparison</th>
<th>Null Hyp. True (there is no difference)</th>
<th>Null Hyp. False (there is a difference)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DECISION</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rejected Null Hyp</td>
<td>Type I error (alpha)</td>
<td>Correct Decision</td>
</tr>
<tr>
<td>Did not Reject Null</td>
<td>Correct Decision</td>
<td>Type II Error</td>
</tr>
</tbody>
</table>

(Alpha = probability of making a Type I error)

**Related Glossary Terms**

Alternative hypothesis

Index
Odds ratio

The odds ratio is used to determine if an event is the same for two groups. The odds of an outcome represents the probability that the outcome does occur divided by the probability that the outcome does not occur. An odds ratio of 1 means the event is equally likely in both groups. An odds ratio greater than one suggests the event is more likely in one group. A ratio less than one suggests the event is less likely in one group as compared to another.

Related Glossary Terms

Relative risk
One-tailed test

A hypothesis test used when the direction of the difference can be predicted with reasonable confidence or the focus of the test is only one tail of the sampling distribution.

Related Glossary Terms

Two-tailed test
Operationalize

Involves determining how an abstract concept will be measured. As an example, temperature is an abstract concept that could be measured by degree Celsius or simple statements such as cold, warm, hot.
Ordinal data

Objects classified by type or characteristic with some logical order. Also referred to as ordinal scale data. Examples of ordinal variables include military rank, letter grade, class standing.

Related Glossary Terms

Interval data, Level of measurement, Nominal data, Qualitative, Ratio data
P-value

The probability of a Type I error. Type I error is rejecting a true null hypothesis. Commonly interpreted as the probability of being wrong when concluding there is statistical significance.

Related Glossary Terms

Alpha, Paradigm, Statistical inference, Statistical significance, Type I error
Paradigm

The larger frame of understanding shared by the profession and/or scientific community; part of the core set of assumptions from which we may base our inquiry.

Related Glossary Terms

P-value, Theory
Parameter

The measure of a population characteristic. This requires that observations must be collected for every member of a population.

Related Glossary Terms

Population, Sample, Statistic
Partial correlation

A multivariate technique that produces a correlation coefficient that represents a bivariate relationship when controlling for other variables.
Partial slope

Reported in multiple regression, a partial slope represents the relationship (slope) between one independent variable and the dependent variable while controlling for all other independent variables included in the regression model.

Related Glossary Terms

Applied Stat: Regression-OLS Multiple
Partial tables

Bivariate tables resulting from controlling by a third variable. There will be one bivariate table for each value (category) in the control variable.

Related Glossary Terms

Contingency table, Control variable
Pearson's C

Pearson's Contingency Coefficient (C) is interpreted as a measure of the relative (strength) of an association between two variables. The coefficient will always be less than 1 and varies according to the number of rows and columns.

Software Output Example: Data from survey of U.S. Adults.

Crosstabulation: GUNLAW (Rows) by SEX (Columns)

Column Variable Label: Respondent's Sex
Row Variable Label: Gun permits

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favor</td>
<td>314</td>
<td>497</td>
<td>811</td>
</tr>
<tr>
<td>Oppose</td>
<td>111</td>
<td>62</td>
<td>173</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Row %</th>
<th>Col %</th>
<th>Total %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favor</td>
<td>38.72</td>
<td>61.28</td>
<td>73.88</td>
</tr>
<tr>
<td>Oppose</td>
<td>64.16</td>
<td>35.84</td>
<td>26.12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favor</td>
<td>314</td>
<td>497</td>
<td>811</td>
</tr>
<tr>
<td>Oppose</td>
<td>111</td>
<td>62</td>
<td>173</td>
</tr>
</tbody>
</table>

Chi-square Value DF p <

<table>
<thead>
<tr>
<th></th>
<th>Favor</th>
<th>Oppose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>37.622</td>
<td>37.417</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>37.417</td>
<td>36.592</td>
</tr>
</tbody>
</table>

Measures of Association

<table>
<thead>
<tr>
<th></th>
<th>Favor</th>
<th>Oppose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cramer's V</td>
<td>.196</td>
<td>.082</td>
</tr>
<tr>
<td>Pearson C</td>
<td>.192</td>
<td>.115</td>
</tr>
</tbody>
</table>

Note: 00.00% of the cells have an expected frequency <5

Interpretation

[a] Statistically significant (most common measure used for significance).
[b] When sample sizes are small, the continuous chi-square value tends to be too large. The Yates continuity correction adjusts for this bias in 2x2 contingency tables. Regardless of sample size, it is a preferred measure for chi-square tests on 2x2 tables.
[d] A symmetric lambda is used when identification of independent and dependent variables is not useful.
[e] Knowing a person's sex can reduce prediction error by 11.5%.

Related Glossary Terms

Applied Stat: Counts-Chi-square, Association
Percentile

The point below which a specific percentage of cases fall.

Related Glossary Terms

Drag related terms here
**Phi**

A chi-square based measure of association for nominal level variables displayed in a 2 by 2 contingency table (2 rows and 2 columns). Phi can range from 0 to 1.

---

**Related Glossary Terms**

Applied Stat: Counts-Chi-square, Association
Pie chart

Commonly used for nominal data. It is a circular figure with slices that represent the relative frequency (proportion) of a value of a variable.

Example:

![Pie chart](image)

- Education Level
- College: 44.00%
- HS Grade: 42.00%
- ≤12 yrs: 14.00%

Related Glossary Terms
Bar chart, Frequency polygon
Point estimation

A statistic based on a sample of observations that represents our best possible estimate of the true population parameter. A point estimate does not provide us with any information on how close our estimate is to the parameter.

Related Glossary Terms

Confidence interval
Pooled estimate

An estimate of the standard deviation of the sampling distribution of the difference in sample means based on the standard deviations from both samples. The pooled estimate is a standard error.

Related Glossary Terms

Drag related terms here
Population

Contains all members of a group. Also referred to as a universe. As an example, everyone living in the state of Indiana would be the population for data used to represent all persons living in Indiana. In comparison, a sample represents a subset of a population.

Related Glossary Terms

Parameter, Universe
Positive association

Two variables vary in the same direction. As the values of one variable increase the values of a second variable also increase. High values in one variable are associated with high values in the second variable.

Related Glossary Terms

Association, Linear
Post hoc test

A statistical technique used to determine which pairs of means from several possible pairings are significantly different. Associated with analysis of variance (ANOVA).

Related Glossary Terms

Applied Stat: Mean-Bonferroni post hoc
Power

The probability of correctly finding statistical significance. A common value for power is .80. To estimate power we need to know four factors: 1) size of the difference between parameters (effect size), 2) significance level and whether 1 or 2-tailed test, 3) standard deviation of each sampled population, and 4) sample size of each population.

Related Glossary Terms

Drag related terms here
Predictive validity

The ability of an indicator to correctly predict (or correlate with) an outcome (GRE and performance in a graduate program).

Related Glossary Terms
Construct validity, Content validity, Criterion-related validity, Validity
Principle of least squares

Used in simple and multiple regression techniques. Assumes a linear relationship between variables where a line can be fit to the data to represent the point at which the deviations of all observations from the regression line are at a minimum.

Related Glossary Terms
Probability distribution

The simple outcomes of a sample space and their associated probabilities comprise a probability distribution. Sample space is the set of probabilities for all possible mutually exclusive and exhaustive outcomes. Probability distributions are used in inferential statistics by providing critical values and their associated probabilities.

Related Glossary Terms

Drag related terms here
Proportional reduction in error (PRE)

An approach to measuring association that compares the number of errors made in predicting values in the dependent variable without considering the independent variable to the reduction in errors made in predicting values when the independent variable is considered. Values can range from 0 to 1 and in some uses can range from -1 to 1. Lambda and Gamma are two PRE measures of association.

Related Glossary Terms
Applied Stat: Counts-Chi-square, Association
Qualitative

Variables measured on a non-metric nominal or ordinal scale are commonly referred to as qualitative. Examples can include blood type, religion, social class.

Related Glossary Terms
Nominal data, Ordinal data
Quantitative

Variables measured on a metric interval or ratio scale are commonly referred to as quantitative. Examples include GPD, income, age, IQ.

Related Glossary Terms
Interval data, Ratio data
Quartiles

The points that represent a division of distribution into quarters.

Related Glossary Terms

Drag related terms here
Quasi-experimental

This study design does not have the ability to employ the controls employed in an experimental design. Although internal validity is less than the experimental design, external validity is generally better and statistical controls are used to compensate for extraneous variables.

Related Glossary Terms
Experimental design, External validity
Random error

Error that can randomly occur in the measurement and data collection process.

Related Glossary Terms
Random sample, Random sampling, Random variable
Random sample

A subgroup of a population or universe that is based on the principle that every member of the population has an equal chance of being included in the sample.

Related Glossary Terms
Random error, Random sampling, Random variable
Random sampling

Each and every element in a population has an equal opportunity of being selected.

Related Glossary Terms
Random error, Random sample, Random variable
**Random variable**

A variable whose values are not pre-determined. A measure where any particular value is based on chance through random sampling. A random variable requires that a researcher have no influence on a particular observation's value.

**Related Glossary Terms**

Random error, Random sample, Random sampling
Range

The simplest measure of variation. Range represents the difference between the lowest and highest observed value in a collection of data.

Related Glossary Terms

Drag related terms here
Ratio data

Objects classified by type or characteristic, with logical order and equal differences between levels, and having a true zero starting point. A ratio variable includes more information than nominal, ordinal, or interval scale data. Measures using this scale are considered to be quantitative variables. Examples include time, distance, income, age. Other examples include frequency counts or percentage scales.

Related Glossary Terms
Interval data, Level of measurement, Nominal data, Ordinal data, Quantitative
Raw score

The actual value of an observation in a data set.

Related Glossary Terms
Data
Regression line

A straight line that best displays the relationship between two variables. The equation for a straight line is used to fit the line to the data points. "Best" is defined as the regression line that minimizes the error exhibited between the observed data points and those predicted with the regression line.

Related Glossary Terms
Applied Stat: Regression-OLS Simple, Regression model
Regression model

A regression model represents the dependent and independent variables included in the analysis.

Related Glossary Terms
Autocorrelation, Beta-weight, Bias, Biased estimate, Heteroskedasticity, Homoscedasticity, Instrumental variable, Intercept (a), Principle of least squares, Regression line, Slope (b), Specification error, Standardized partial slopes
Relative frequency

Observed frequencies that have been converted into percentages. Percentages are based on the total number of observations. As an example, if 50 out of 100 subjects are female, the relative frequency would be 50% (50/100). Most statistical software provides this by default when you run a frequency procedure.

Related Glossary Terms

Related Glossary Terms
Relative risk

The ratio of two conditional probabilities. Relative risk is defined as the ratio of risk in the exposed group divided by the risk in the unexposed group. Relative risk equals 1 when an event is equally probable in both groups. A relative risk greater than 1 suggests the event is more likely in the one group as compared to another. A relative risk less than 1 suggest the event is less likely.

Related Glossary Terms
Odds ratio
Reliability

The extent to which a measure obtains similar results over repeat trials.

Related Glossary Terms

Validity

Index
Replication

Duplicating a study or experiment to confirm results or expose error.

Related Glossary Terms
Drag related terms here
Research question

Defines the purpose of the study by clearly identifying the relationship(s) the researcher intends to investigate.

Related Glossary Terms

Abstract concept
Response bias

Errors in data collection caused by differing patterns and completeness of data collection that are dominated by a specific subgroup within the sample.

Related Glossary Terms
Drag related terms here
Response variable

The measure not controlled in an experiment. Commonly known as the dependent variable.

Related Glossary Terms

Dependent variable
Sample

A subset of a population. In inferential statistics, sample data are used to describe (estimate) the population value.

Related Glossary Terms
Parameter, Sample distribution, Sampling distribution, Sampling frame
Sample distribution

A frequency distribution of data from one sample.

Related Glossary Terms
Sample, Sampling distribution, Sampling frame, Skewness
Sampling distribution

A probability distribution representing an infinite number of sample distributions for a given sample size drawn from the population.

Related Glossary Terms
Sample, Sample distribution, Sampling frame, Skewness
Sampling frame

A list of all possible members of a population under study. For a sample to be representative of a population, the sampling frame must include as best possible all members of the population.

Related Glossary Terms
Sample, Sample distribution, Sampling distribution
Scale

A scale is a numerical representation of the values of a variable. In statistics, we find four general types of scales (nominal, ordinal, interval, and ratio). Scale is also sometimes used to refer only to continuous variables (not ordinal or nominal).

Related Glossary Terms

Level of measurement
Secondary analysis

Research conducted on data collected by another researcher who probably collected the data for a purpose other than your primary interest.

Related Glossary Terms

Data
Skew

The extent to which a distribution is affected by a few extremely large or small scores. A few extremely large scores will result in a positively skewed distribution. A few extremely small scores will result in a negatively skewed distribution.

Related Glossary Terms
Skewness
Skewness

Skewness provides an indication of the how asymmetric the distribution is for a given sample. Knowing the mean and median will provide enough information to determine if there is a positive or negative skew. If the mean is greater than the median, there is a positive skew. If the mean is less than the median, there is a negative skew.

Related Glossary Terms

Normal distribution, Sample distribution, Sampling distribution, Skew
Slope (b)

Represents the average change in Y associated with a one unit change in X.

Related Glossary Terms
Spearman's rho

A measure of association for two ordinal level variables that have numerous rank ordered categories that are similar in form to a continuous variable. Spearman's rho can range from -1 (strong negative association) to 1 (strong positive association). Zero represents no association.

Related Glossary Terms
Association
Specification error

Specification error occurs when the functional form of a relationship does not match the statistical technique. Using simple regression as an example, if the relationship between X and Y cannot be modeled by a straight line, we may fail to find significance because our technique was not appropriate for the data and/or relationship we are attempting to model.

Related Glossary Terms

Applied Stat: Regression-OLS Simple, Regression model
Spurious association

Association between two variables that can be better explained by or depends greatly upon a third variable. Suggests there is not a relationship between the two variables. Instead, the two variables are caused or strongly related to a third (control) variable.

Related Glossary Terms

Association
Standard deviation

A measure of dispersion that is represented in units of measurement consistent with the raw data. It is the square root of the variance. Standard deviation is also known as the root mean square (RMS), since it is the square root of the mean squared deviations.

\[ \sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} \]

Population Standard Deviation:

\[ S = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n - 1}} \]

Sample Standard Deviation:

Example:

<table>
<thead>
<tr>
<th>Variable (Xi)</th>
<th>Age</th>
<th>(Xi-mean)</th>
<th>(Xi-mean)^2 (Xi-mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>20.29</td>
<td>411.68</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>12.29</td>
<td>151.04</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>12.29</td>
<td>151.04</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>-2.29</td>
<td>5.24</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>10.71</td>
<td>114.70</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>15.71</td>
<td>246.80</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>20.71</td>
<td>428.90</td>
<td></td>
</tr>
</tbody>
</table>

\[ \text{n=} \quad 7 \quad 1509.43 \quad \text{sum of squares} \]
\[ \text{mean=} \quad 44.29 \quad 251.57 \quad \text{sample variance} \]
\[ \quad \quad \quad 15.86 \quad \text{sample standard deviation} \]

Related Glossary Terms
Standard error of the mean, Variance
Standard error of the mean

A measure reflecting the dispersion in a sampling distribution (not a sample distribution). Also known as the standard deviation of the sampling distribution.

\[ \frac{s}{\sqrt{n}} = \frac{s}{\sqrt{n}} \]

Note: assumes sample standard deviation was calculated using:

\[ s = \sqrt{\frac{\sum (x_i - \bar{x})}{n - 1}} \]

Related Glossary Terms

Standard deviation

Index
Standard score

Represents standard deviations from the mean. It is very useful in describing an object’s relative position in a data distribution. A positive standard score indicates an individual object’s score is above the mean of the group. Similarly, a negative standard score indicates the object’s value is less than the mean. The standard score value represents how many standard deviations a particular raw score is above or below the mean. Also called standardized values or z-scores.

\[ Z = \frac{X_i - \bar{X}}{S} \]

Example:  
(\textit{where mean}=44.29 and \textit{standard deviation}=15.86)

<table>
<thead>
<tr>
<th>Variable (Xi)</th>
<th>Age</th>
<th>(Xi-mean)</th>
<th>(Xi-mean)/s</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doug</td>
<td>24</td>
<td>-20.29</td>
<td>-20.29/15.86</td>
<td>-1.28</td>
</tr>
<tr>
<td>Jenny</td>
<td>32</td>
<td>-12.29</td>
<td>-12.29/15.86</td>
<td>-0.77</td>
</tr>
<tr>
<td>John</td>
<td>55</td>
<td>10.71</td>
<td>10.71/15.86</td>
<td>0.68</td>
</tr>
<tr>
<td>Beth</td>
<td>60</td>
<td>15.71</td>
<td>15.71/15.86</td>
<td>0.99</td>
</tr>
<tr>
<td>Ed</td>
<td>65</td>
<td>20.71</td>
<td>20.71/15.86</td>
<td>\textbf{1.31} [a]</td>
</tr>
</tbody>
</table>

Interpretation

[a] As an example of how to interpret z-scores, Ed is 1.31 standard deviations above the mean age for those represented in the sample. Another simple example is exam scores from two history classes with the same content but different instructors and different test formats. To adequately compare student A’s score from class A with Student B’s score from class B you need to adjust the scores by the variation (standard deviation) of scores in each class and the distance of each student’s score from the average (mean) for the class.

Related Glossary Terms
Test statistic, Z score

Index
Standardized partial slopes

Also known as beta-weights. Standardized regression partial slope coefficients or beta-weights indicate the amount of change in the standardized scores of Y (dependent variable) for a one unit change in the standardized scores of each independent variable, controlling for the effects of all independent variables included in the regression model. Beta-weights in the same regression can be compared to one another to evaluate relative effect on the dependent variable. Beta-weights will also indicate the direction of the relationship (positive or negative). Regardless of whether negative or positive, larger beta-weight absolute values indicate a stronger relationship.

Related Glossary Terms

Applied Stat: Regression-OLS Multiple, Regression model
Statistic

A measure of a sample characteristic.

Related Glossary Terms

Parameter
Statistical analysis

A field of inquiry that is concerned with the collection, organization, and interpretation of data according to well-defined procedures. It is the key to every scientific discipline that studies the behavioral and biological characteristics of human beings or our physical environment.

Related Glossary Terms

Drag related terms here
Statistical control

Holding the value of one variable constant in order to clarify associations among other variables.

Related Glossary Terms

Control variable
Statistical inference

Reaching a conclusion about a population on the basis of information obtained from a random sample drawn from that population. There are two such methods, statistical estimation and hypothesis testing.

Related Glossary Terms

P-value
Statistical significance

Interpreted as the probability of a Type I error. Test statistics that meet or exceed a critical value are interpreted as evidence that the differences exhibited in the sample statistics are not due to random sampling error and therefore are evidence supporting the conclusion there is a real difference in the populations from which the sample data were obtained.

Related Glossary Terms

P-value
Statistics

Mathematical techniques used for organizing and analyzing data.

Related Glossary Terms
Drag related terms here
Stochastic element

Unobserved random variable representing the error of a model.

Related Glossary Terms

Drag related terms here
Stratification

The grouping of units representing similar (homogeneous) segments of a population. Each group is referred to as a stratum.

Related Glossary Terms

Drag related terms here
Substantive Importance

The real world implications of statistical significance. A statistically significant finding may not be substantively important in practical application (i.e., does a statistically significant difference of one or two percent matter? -- it may but the context of the research is very important).

The substantive importance of a statistical result is a common problem when dealing with very large sample sizes where small differences in groups can be statistically significant because of the mathematics behind statistical testing (probability distributions and the affect of sample size on measures of variation).

Related Glossary Terms

Drag related terms here
Sum of squares

The sum of the squared deviations between each observed value and the group mean. Used to calculate variance and standard deviation.

Related Glossary Terms

Drag related terms here
Summative evaluation

A method of judging the value of a program at the end of the program activities. The focus is on the outcome.

Related Glossary Terms

Formative evaluation
Test statistic

The value computed from sample statistics that is converted to a standardized score for comparison to the critical value of a sampling distribution. If the test statistic equals or exceeds the critical value, statistical significance is assumed and the null hypothesis is rejected.

Related Glossary Terms

Standard score
Theoretical frequency distribution

A theoretical frequency distribution is identical in design to a frequency distribution but is based on the idea of an infinite number of observations. As the number of observations increases, a histogram of a theoretical frequency distribution begins to take on a smooth curve as compared to the normal bar representation. This smooth curve is considered a representation of a continuous distribution. In addition, the y axis represents the height of the curve instead of the count.

Related Glossary Terms
Central limit theorem

Index
Theory

One or more propositions that suggest why an event occurs that provides a framework for further analysis.

Related Glossary Terms

Conceptual model, Deduction, Paradigm
Two-tailed test

A type of hypothesis test used when the direction of the difference cannot be reasonably predicted a priori (before collecting or analyzing the data) or the focus is on both tails of the sampling distribution. Most statistical testing employs a two-tailed test. Statistical software typically reports p-values based on a two-tailed test. Divide the p-value by 2 to get the one-tailed p-value.

Related Glossary Terms

One-tailed test
Type I error

Rejecting a true null hypothesis. Commonly interpreted as the probability of being wrong when concluding there is statistical significance. Also referred to as p-value or significance.

Related Glossary Terms

Alpha, P-value, Type II error
Type II error

Retaining a false null hypothesis. Also referred to as Beta.

Related Glossary Terms
Alpha, Type I error
Unit of analysis

The object under study. This could be people, schools, cities, etc.
Univariate analysis

The study of one variable characteristic. Examples include test scores, voting record, salaries, education obtained, and medical test results.

Related Glossary Terms

Applied Stat: Bivariate Analysis-Counts
Universe

Contains all members of a group. Also referred to as a population. In comparison, a sample represents a subset of a population.

Related Glossary Terms
Population
Validity

The extent to which a measure accurately represents an abstract concept. There are various forms of validity (see links below).

Related Glossary Terms
Construct validity, Content validity, Criterion-related validity, External validity, Face validity, Predictive validity, Reliability
Variable

A characteristic that can form different values from one observation to another. If a characteristic is not a variable, it is a constant.

Related Glossary Terms

Drag related terms here
Variance

A measure of dispersion that involves summing the squared deviations of each value from the mean and dividing by the number of observations (n).

\[ \sigma^2 = \frac{\sum(X_i - \mu)^2}{N} \]

Population Variance:

\[ S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1} \]

Sample Variance:

Related Glossary Terms
Dispersion, Standard deviation
Weighted mean

A mean based on weighted observations. This weight can be some measure of importance, or in the case of obtaining a grand mean from several means with different sample sizes, the weight is the count for the subgroup producing the mean.

\[
\bar{X}_w = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2 + n_3 \bar{X}_3}{n_1 + n_2 + n_3}
\]

Example:

\[
\bar{X}_1 = 12, n = 10
\]
\[
\bar{X}_2 = 14, n = 15
\]
\[
\bar{X}_3 = 18, n = 40
\]

Wrong Method: \[
\frac{12 + 14 + 18}{3} = 14.7
\]

Correct Method: \[
\frac{10(12) + 15(14) + 40(18)}{10 + 15 + 40} = 16.2
\]

Related Glossary Terms

Mean

Index
**Weighting**

Used to adjust a sample that was created from an unequal representation of a population. Once weighted the sample is expected to be representative of the population. Weighting is used when over sampling is employed to collect enough subjects from small groups and is also used to compensate for disproportionate response rates.

---

**Related Glossary Terms**

Drag related terms here
Z score

A score that standardizes individual observation values in one distribution of scores that fit a normal distribution. A z score represents the distance of an individual score from the mean, the direction of the distance, and the variation of scores in the distribution.

Related Glossary Terms

Standard score
Zero-order correlation

A correlation coefficient for bivariate comparisons when no other variable is used as a control.

Related Glossary Terms

Applied Stat: Correlation