

Criteria for Selecting Data Sets for an Introductory Applied Statistics Course

John D. McKenzie, Jr.
Babson College
Babson Park, MA 02457-0310
mckenzie@babson.edu

Beyond the Formula Conference
Rochester, NY
24 July 2003

Overview

1. Introduction
2. Three Common Criteria
3. Other Possible Criteria
4. Advantages and Disadvantages of Selecting Multi-Criteria Data Sets

Abstract

The first criterion for deciding upon a data set is that it illustrates a specific technique or concept. Two other common criteria are whether it will be interesting to students and if it comes from the "real world". But many other criteria should be considered when one selects the data sets for a first course in order to prepare students for their future statistical work. With today's software there should be data sets that include alphanumeric qualitative (categorical) variables as well as numeric qualitative variables. In addition to cross-sectional data sets, there should be time-series data sets. Some data sets should illustrate stacked data, while others should illustrate unstacked data. There should be data sets that are much larger than the typical sets included with most textbooks. And, some data sets should include missing data (which is rarely presented in textbooks but always seen in practice) and possibly include data that must be cleansed (which is 80% of the work of many statistical projects). This session will present examples of data sets that illustrate each of these and other criteria. Members of the audience will be asked to put forward their favorite data sets and to explain why they decided to use them.

Three Common Criteria

1. Ability to Illustrate Specific Technique (Method) or Concept
2. Of Interest to Students
3. "Real-World"

Illustration of a Concept

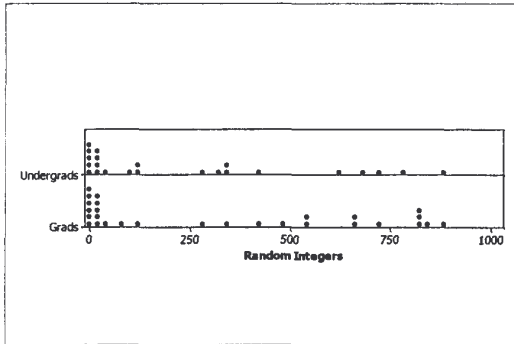
Humans
are
awful random number generators!

Results for: RandomIntegers.MTW

MTB > info

Information on the Worksheet

| Column | Count | Missing | Name |
|--------|-------|---------|----------|
| C1 | 24 | 0 | UGender |
| C2 | 24 | 1 | URandomI |
| C3 | 28 | 0 | GGender |
| C4 | 28 | 0 | GRandomI |



A Wonderful Artificial Data Set

Current worksheet: FA.MTW

MTB > info

Information on the Worksheet

| Column | Count | Name |
|--------|-------|------|
| C1 | 11 | X |
| C2 | 11 | Y1 |
| C3 | 11 | Y2 |
| C4 | 11 | Y3 |
| C5 | 11 | X4 |
| C6 | 11 | Y4 |

MTB > Print 'X'-'Y4'.

Data Display

| Row | X | Y1 | Y2 | Y3 | X4 | Y4 |
|-----|----|-------|------|-------|----|-------|
| 1 | 10 | 8.04 | 9.14 | 7.46 | 8 | 6.58 |
| 2 | 8 | 6.95 | 8.14 | 6.77 | 8 | 5.76 |
| 3 | 13 | 7.58 | 8.74 | 12.74 | 8 | 7.71 |
| 4 | 9 | 8.81 | 8.77 | 7.11 | 8 | 8.84 |
| 5 | 11 | 8.33 | 9.26 | 7.81 | 8 | 8.47 |
| 6 | 14 | 9.96 | 8.10 | 8.84 | 8 | 7.94 |
| 7 | 6 | 7.24 | 6.13 | 6.08 | 8 | 5.25 |
| 8 | 4 | 4.26 | 3.10 | 5.39 | 19 | 12.50 |
| 9 | 12 | 10.84 | 9.13 | 8.15 | 8 | 5.56 |
| 10 | 7 | 4.82 | 7.26 | 6.42 | 8 | 7.91 |
| 11 | 5 | 5.68 | 4.74 | 5.73 | 8 | 6.89 |

QTM1310

Fall Semester, 2002

Preferences for Course Data

Please check the three data types you would like to analyze in *Probability and Statistics* this semester.

- Airline Industry (e.g., arrival delays)
- Automotive Industry (e.g., mileage for sports cars)
- Babson (e.g., demographics of current students)
- Education Industry (e.g., tuition)
- Financial Industry (e.g., stock prices)
- Food Industry (e.g., pizza sales)
- Health (e.g., nutritional statistics)
- International Business (e.g., country statistics)
- Marketing (e.g., effectiveness of advertisements)
- Motion Picture Industry (e.g., ratings of recent films)
- Recording Industry (e.g., number of CDs sold)
- Sports Industry (e.g., salaries of professional athletes)
- Other (Please specify below.)

Singer, J. and Willet, J. (1990),
"Improving the Teaching of
Statistics: Putting the Data Back into
Data Analysis,"
The American Statistician,
44(3), 223-230.

Seven Desired Characteristics

1. Authenticity
2. Background Information
3. Interest and Relevance
4. Substantive Learning
5. Availability of Multiple Analyses
6. The Importance of Raw Data
7. Case Identifiers

Raw and Summarized Data

Results for: EmployRaw.MTW
 MTW = jobs
 Information on the Worksheet

| Column | Count | Name |
|--------|-------|-----------|
| C1 | 148 | JobStatus |
| C2 | 148 | Ethnicity |

 MTW = Table 'JobStatus' 'Ethnicity'.
 Tabulated Statistics: JobStatus, Ethnicity

| Rows | JobStatus | Columns | Ethnicity | All | | |
|------|-----------|---------|-----------|-----|-----|-----|
| | Admin | 6 | 13 | 9 | 349 | 349 |
| | Faculty | 27 | 4 | 1 | 105 | 250 |
| | Service | 0 | 2 | 0 | 14 | 28 |
| | Staff | 7 | 9 | 1 | 34 | 64 |
| | All | 32 | 32 | 20 | 498 | 789 |

 Data Contents -->
 Count

Results for: EmploySum.MTW
 MTW = jobs
 Information on the Worksheet

| Column | Count | Name |
|--------|-------|-----------|
| C1 | 5 | JobStatus |
| C2 | 5 | Admin |
| C3 | 5 | Black |
| C4 | 5 | Hispanic |
| C5 | 5 | White |

 MTW = Print 'JobStatus'-'White'.
 Data Display

| Row | JobStatus | Admin | Black | Hispanic | White |
|-----|-----------|-------|-------|----------|-------|
| 1 | Admin | 6 | 13 | 9 | 240 |
| 2 | Faculty | 27 | 4 | 1 | 250 |
| 3 | Service | 0 | 2 | 0 | 14 |
| 4 | Staff | 7 | 9 | 1 | 34 |
| 5 | Staff | 7 | 9 | 1 | 34 |

Other Possible Criteria

1. Raw and Summarized Data
2. Alphabetic and Numerically Coded Qualitative (Categorical) Data
3. Cross-Sectional and Time Series Data
4. Date Data
5. Stacked and Unstacked Data
6. Paired Data
7. Size of Data Set
8. Missing Data
9. Dirty Data
10. Non-Significant Test Data

Alphabetic and Numerically Coded Qualitative (Categorical) Data

Results for: JEANS.MTW
 MTW = sally.cnt
 Tally for Discrete Variables: Spend

| Spend | Count |
|---------|-------|
| 0Degree | 47 |
| 1Degree | 42 |
| 2Degree | 74 |
| 3Degree | 46 |
| 4Degree | 45 |
| 5Degree | 34 |
| All | 334 |

 MTW = Name c14 = 'NumSpend'
 MTW = Code 1 '0Degree' | 2 | '1Degree' | 3 | '2Degree' | 4 | '3Degree' | 5 | '4Degree' | 6 | '5Degree'
 MTW = Table 'Spend' 'NumSpend'.
 Tabulated Statistics: Spend, NumSpend

| Rows | Spend | Columns | NumSpend | All | | | | |
|------|---------|---------|----------|-----|----|----|----|-----|
| | 0Degree | 1 | 2 | 3 | 4 | 5 | 6 | All |
| | 0Degree | 47 | 0 | 0 | 0 | 0 | 0 | 47 |
| | 1Degree | 0 | 42 | 0 | 0 | 0 | 0 | 42 |
| | 2Degree | 0 | 0 | 74 | 0 | 0 | 0 | 74 |
| | 3Degree | 0 | 0 | 0 | 46 | 0 | 0 | 46 |
| | 4Degree | 0 | 0 | 0 | 0 | 45 | 0 | 45 |
| | 5Degree | 0 | 0 | 0 | 0 | 0 | 34 | 34 |
| | All | 47 | 42 | 74 | 46 | 45 | 34 | 334 |

 Data Contents -->
 Count
 MTW = Print 'Spend' 'NumSpend'.
 Data Display

| Row | Spend | NumSpend |
|-----|---------|----------|
| 1 | 0Degree | 47 |
| 2 | 1Degree | 42 |
| 3 | 2Degree | 74 |
| 4 | 3Degree | 46 |
| 5 | 4Degree | 45 |
| 6 | 5Degree | 34 |
| 7 | 0Degree | 47 |

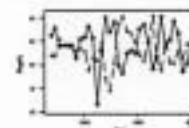
Cross-Sectional and Time Series Data

Most textbook data are cross-sectional but most data seen by consumers are time series.

Results for: Election2.MTW
 MTW = jobs
 Information on the Worksheet

| Column | Count | Name |
|--------|-------|------|
| C1 | 14 | Year |
| C2 | 14 | RepR |
| C3 | 14 | DemR |

 MTW = Plot 'RepR'-'Year' 'DemR'-'Year'.
 RDD = Symbol.
 RDB = FontSize.
 RDC = PostScale '+' Democratic 'R' and 'Republican 'R'.
 RDD = Overlay.
 RDB = Software.
 RDC = Annotation.
 Plot 'RepR'-'Year' 'DemR'-'Year'.



Date Data

Results for: USAArrivals.MTW

MTB > info

Information on the Worksheet

| Column | Count | Name |
|--------|-------|------------|
| T C1 | 155 | Carrier |
| D C2 | 155 | Date |
| C3 | 155 | FlightNum |
| T C4 | 155 | Origin |
| D C5 | 155 | SchArrival |
| D C6 | 155 | ActArrival |
| C7 | 155 | SchElapsed |
| C8 | 155 | ActElapsed |
| C9 | 155 | ArrvDelay |

MTB > Print 'Carrier'-'ActArrival'.

Data Display

| Row | Carrier | Date | FlightNum | Origin | SchArrival | ActArrival |
|-----|---------|------------|-----------|--------|------------|------------|
| 1 | US | 11/26/2000 | 76 | PIT | 18:03 | 19:25 |
| 2 | US | 11/26/2000 | 98 | PIT | 9:22 | 9:40 |
| 3 | US | 11/26/2000 | 140 | PHL | 14:49 | 16:00 |
| 4 | US | 11/26/2000 | 142 | PHL | 19:59 | 22:30 |
| 5 | US | 11/26/2000 | 217 | CLT | 20:59 | 21:56 |
| 6 | US | 11/26/2000 | 225 | BUF | 18:03 | 0:00 |
| 7 | US | 11/26/2000 | 314 | PHL | 16:50 | 18:29 |
| 8 | US | 11/26/2000 | 392 | ROC | 18:47 | 20:46 |
| 9 | US | 11/26/2000 | 478 | PHL | 20:50 | 21:35 |
| 10 | US | 11/26/2000 | 502 | PHL | 8:42 | 12:12 |
| 11 | US | 11/26/2000 | 512 | CLT | 15:51 | 16:21 |
| 12 | US | 11/26/2000 | 515 | PHL | 13:43 | 14:04 |
| 13 | US | 11/26/2000 | 582 | PHL | 18:53 | 20:55 |
| 14 | US | 11/26/2000 | 720 | CLT | 17:10 | 19:36 |
| 15 | US | 11/26/2000 | 728 | BHI | 12:20 | 12:16 |

Stacked and Unstacked Data

It is more efficient
to store stacked data
but often easier
to explain a method with unstacked data.

Results for: Murders.MTW

MTB > info

Information on the Worksheet

| Column | Count | Name |
|--------|-------|------------|
| T C1 | 50 | State |
| C2 | 50 | MurderRate |
| T C3 | 50 | DPStatus |
| T C4 | 50 | Region |

Results for: Murder.MTW

MTB > info

Information on the Worksheet

| Column | Count | Name |
|--------|-------|-----------|
| T C1 | 9 | NState |
| C2 | 9 | NEMurder |
| T C3 | 9 | NEMStatus |
| T C4 | 12 | NState |
| C5 | 12 | NOMurder |
| T C6 | 12 | NOMStatus |
| T C7 | 16 | SSState |
| C8 | 16 | SMurder |
| T C9 | 16 | SMStatus |
| T C10 | 13 | WState |
| C11 | 13 | WMurder |
| T C12 | 13 | WStatus |

Paired Data

Results for: DrugMarkup.MTW

MTB > info

Information on the Worksheet

| Column | Count | Name |
|--------|-------|----------|
| T C1 | 10 | Location |
| C2 | 10 | Prozak |
| C3 | 10 | PGeneric |
| C4 | 10 | Xanax |
| C5 | 10 | XGeneric |

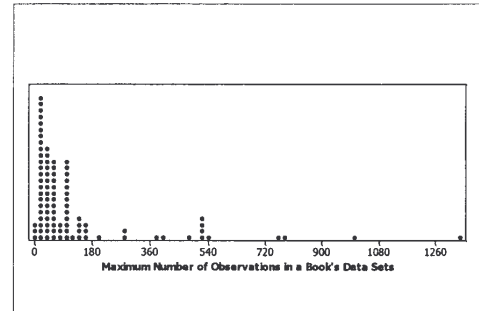
MTB > Print 'Location'-'XGeneric'.

Data Display

| Row | Location | Prozak | PGeneric | Xanax | XGeneric |
|-----|-------------|--------|----------|-------|----------|
| 1 | Wholesale | 77.82 | 4.83 | 31.98 | 0.93 |
| 2 | Chain1 | 111.00 | 25.29 | 52.49 | 12.29 |
| 3 | Chain2 | 98.00 | 71.59 | 47.09 | 10.00 |
| 4 | Chain3 | 103.00 | 81.00 | 51.59 | 10.00 |
| 5 | Warehouse | 85.00 | 9.63 | 37.06 | 7.11 |
| 6 | Internet | 86.00 | 46.00 | 33.76 | 7.99 |
| 7 | Mail | 91.51 | 75.80 | 47.68 | 10.70 |
| 8 | SuperMarket | 100.69 | 78.00 | 46.00 | 16.69 |
| 9 | MailPa2 | 102.95 | 92.95 | 50.56 | 19.75 |
| 10 | Chain4 | 101.00 | 75.49 | 48.99 | 11.79 |

Size of Data Set

The future is now!



Missing Data

Such data do not exist in texts
but always occur in practice.

Results for: UGradSurvey.MTW

MTB > info

Information on the Worksheet

| Column | Count | Missing | Name |
|--------|-------|---------|----------|
| T C1 | 24 | 0 | Gender |
| T C2 | 24 | 2 | HSGrade |
| C3 | 24 | 0 | IYear |
| C4 | 24 | 4 | ITimes |
| C5 | 24 | 4 | IMinutes |
| C6 | 24 | 0 | EMail |
| C7 | 24 | 0 | IM |
| C8 | 24 | 0 | Songs |
| C9 | 24 | 0 | Chat |
| C10 | 24 | 0 | Photos |
| C11 | 24 | 0 | Other |

MTB > Print 'Gender'-'IMinutes'.

Data Display

| Row | Gender | HSGrade | IYear | ITimes | IMinutes |
|-----|--------|---------|-------|--------|----------|
| 1 | Female | A | 1997 | 1 | 180.0 |
| 2 | Male | A- | 1996 | * | * |
| 3 | Male | B+ | 1994 | * | 120.0 |
| 4 | Male | B+ | 1996 | 1 | 60.0 |
| 5 | Female | B | 1994 | 10 | * |
| 6 | Male | A- | 1995 | 3 | 60.0 |
| 7 | Female | A- | 1996 | 1 | 5.0 |
| 8 | Female | B+ | 1996 | 2 | 210.0 |
| 9 | Male | A- | 1996 | 5 | 70.0 |
| 10 | Female | B+ | 1997 | 1 | 540.0 |
| 11 | Male | B | 1996 | 6 | 210.0 |
| 12 | Female | B+ | 1997 | 2 | 15.0 |
| 13 | Female | B | 1993 | * | 1440.0 |
| 14 | Female | B+ | 1996 | 6 | 200.0 |
| 15 | Male | B | 1995 | 1 | * |
| 16 | Male | B+ | 1998 | 10 | 60.0 |
| 17 | Female | B | 2000 | * | * |
| 18 | Male | A- | 1997 | 6 | 90.0 |
| 19 | Male | B+ | 1996 | 2 | 25.0 |
| 20 | Male | A- | 1993 | 4 | 120.0 |
| 21 | Male | B+ | 1995 | 3 | 720.0 |
| 22 | Male | B+/A- | 1995 | 2 | 22.5 |
| 23 | Male | A | 1992 | 2 | 35.0 |
| 24 | Female | A | 1997 | 8 | 120.0 |

Non-Significant Test Data

Another Example of Publication Bias?

Dirty Data

Ryan, Joiner, and Ryan's Classic

Results for: Restrnt.MTW

MTB > info

Information on the Worksheet

| Column | Count | Missing | Name |
|--------|-------|---------|----------|
| C1 | 279 | 0 | ID |
| C2 | 279 | 1 | Outlook |
| C3 | 279 | 25 | Sales |
| C4 | 279 | 55 | Newcap |
| C5 | 279 | 39 | Value |
| C6 | 279 | 42 | CostGood |
| C7 | 279 | 44 | Pages |
| C8 | 279 | 44 | Ads |
| C9 | 279 | 12 | TypeFood |
| C10 | 279 | 11 | Seats |
| C11 | 279 | 10 | Owner |
| C12 | 279 | 14 | FT-Emp1 |
| C13 | 279 | 13 | FT-Emp1 |
| C14 | 279 | 16 | Size |

Recent Discovery

Results for: TBIII2.MTW

MTB > info

Information on the Worksheet

| Column | Count | Name |
|--------|-------|-------|
| C1 | 104 | High |
| C2 | 104 | Low |
| C3 | 104 | Close |

MTB > Print 'High'-'Close'.

Data Display

| Row | High | Low | Close |
|-----|-------|-------|-------|
| 1 | 1.873 | 1.790 | 1.816 |
| 2 | 1.862 | 1.738 | 1.795 |
| 3 | 1.821 | 1.646 | 1.826 |
| 4 | 1.935 | 1.697 | 1.816 |
| 5 | 2.111 | 1.728 | 1.743 |
| 6 | 2.041 | 1.966 | 2.043 |

Selecting Multi-Criteria Data Sets

Disadvantages ☹

- Still Another Item to Cover
- Too Much for the Student
- Collection Time for the Instructor

Advantages ☺

- Better Preparation for Future Statistical Work!

Using Microsoft Excel for Applied Statistics Courses

John D. McKenzie, Jr.
Babson College
Babson Park, MA 02457-0310
mckenzie@babson.edu

Beyond the Formula Conference
Rochester, NY
25 July 2003

Abstract

In recent years Microsoft Excel's statistical functions and Analysis ToolPak have been increasingly used in the introductory applied statistics course. Yet, the American Statistical Association has recommended that it is "not sufficient for the teaching of statistics, let alone for research and consulting". This session will present an up-to-date overview of why one should or should not use this spreadsheet for K-16 applied statistics courses. It will expand upon the session at 2001 Joint Statistical Meetings on the use of Microsoft Excel for statistical analyses, organized by Jon Cryer and the speaker. At that session two of the speakers presented a large number of concerns about using this popular spreadsheet package in either the classroom or the workplace. Some of these concerns were computational; others were related to documentation and ease of use. Many users of Excel are often unaware of some of these concerns. Another panel member who is employed by Microsoft acknowledged that there were serious problems with its use for statistics. He hoped that the most severe problems would be eliminated in future, but not immediate, releases. This session will identify the strengths and weaknesses in using Excel for each component of the first course: descriptive statistics (graphical displays and summary measures), elementary probability, introductory inferential statistics, and common statistical analyses such as regression. It will also present alternatives to using this spreadsheet in the classroom, such as Excel add-ins, Internet freeware, and student versions of statistical software. The session is planned so that there will be ample opportunity for audience participation.

Overview

1. Excel's Statistical Capabilities
2. Strengths and Weaknesses
3. Some Examples of Specific Concerns
4. Past and Future Changes
5. Alternatives

Excel's Statistical Capabilities

Functions

Data Analysis Toolpak

Chart Wizard

Pivot Tables

Quiz

Indicate whether the following statements are true or false.

1. A data set cannot have more than one modal value.
2. In regression a normal probability plot is a scatter plot with y on the vertical axis and sample percentile on the horizontal axis.
3. Tied ranks are usually handled by assigning the lower rank to each of the observations.
4. Short-cut formulas are the best to use to calculate statistics.
5. The maximum of a set of missing data is 0.
6. Grey Matter International developed Excel's Analysis ToolPak.
7. A histogram is equivalent to a bar chart.
8. In practice one-sample mean problems are usually handled with the z distribution.
9. ANOVA tests the hypothesis that means from two or more samples are equal.
10. In linear regression it is possible to obtain a negative R^2 .

American Statistical Association Recommendation

Efficient computing tools are essential for statistical research, consulting, and teaching. Generic packages such as Excel are not sufficient even for the teaching of statistics, let alone for research and consulting.

Strengths

Core Routines

Spreadsheet Features

Familiarity

Cost

Weaknesses

Lack of Routines

Terrible Graphics

Computational Errors

Faulty Documentation

Poor Choice of Defaults

Not User-Friendly

Specific Concerns

for

Descriptive Statistics
(Displays and Summary Measures)

Elementary Probability

Introductory Inference

Common Analyses
(Linear Regression, Chi-Square Test)

Histogram

100 standardized values for data:

| | | | | | | |
|----------|----------|----------|----------|----------|----------|----------|
| -2.02181 | 0.24807 | -0.86078 | 0.87338 | 0.21472 | -0.03647 | -0.28403 |
| 1.15488 | 0.82424 | 0.40384 | -0.81804 | 0.07838 | 0.83488 | 0.03280 |
| -0.30394 | -0.17948 | -0.31014 | -0.34438 | -0.04800 | -0.01938 | 0.07488 |
| -0.12688 | -0.17948 | -0.00940 | -0.26837 | -0.21208 | 0.07948 | 0.17838 |
| 0.14780 | 0.27048 | 0.88018 | 0.40378 | -0.12367 | -0.07948 | -0.08012 |
| 0.27328 | 0.18348 | 0.83897 | 0.04888 | -0.51888 | 0.15808 | 0.48078 |
| 1.10438 | 2.44402 | 0.09842 | -1.88387 | -0.03818 | -0.72142 | 1.09044 |
| -0.48184 | 0.77052 | 0.08058 | 2.72488 | 0.42727 | 0.28322 | -0.27187 |
| -0.04470 | -0.51038 | 0.09048 | -1.08152 | -0.31418 | 0.11417 | 0.40388 |
| -1.48728 | -2.37638 | 1.08888 | -0.08932 | -0.40408 | 0.22217 | 0.48902 |
| 0.08448 | 0.40234 | -0.03778 | -0.82142 | 0.20440 | -0.07948 | -0.18487 |
| 1.02872 | 0.95788 | -1.23518 | 0.08088 | -0.08178 | 0.42802 | 0.37488 |
| 0.16238 | -0.28488 | -1.02378 | -0.78387 | -0.10872 | 0.01148 | 0.04188 |
| 0.11848 | 0.47917 | -0.18732 | -0.08428 | 0.48812 | -0.04481 | -1.14874 |
| 0.02272 | 0.01280 | | | | | |

Excel histogram (frequency table):

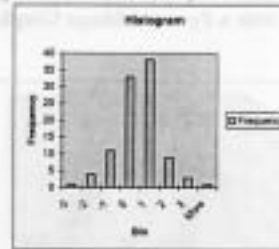
| Bin | Frequency |
|------|-----------|
| -3 | 1 |
| -2 | 4 |
| -1 | 11 |
| 0 | 33 |
| 1 | 38 |
| 2 | 9 |
| 3 | 3 |
| More | 1 |

Missing Data

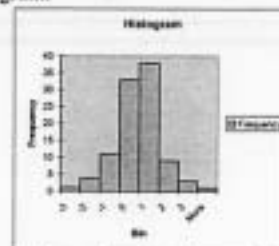
In statistical calculations in contrast to numerical calculations, a blank cell sometimes behaves as a zero and sometimes as a missing value. The cells (1, blank cell, 2) yield a sum of 3, an average of 1.5, and a standard deviation of 0.7070, while the cells (1, 0, 2) yield a sum of 3, an average of 1, and a standard deviation of 1.

Histogram

Excel histogram chart (bar chart):



Excel histogram:



The left-hand tick marks of each interval are really the right-hand tick marks.

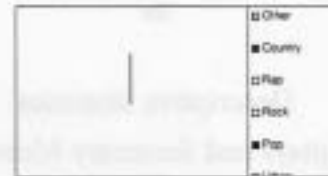
Pie Chart

A survey yields of 20 responses with both numerical codes and textual phrases.

| Music Preference | |
|------------------|--------------|
| Code | Category |
| 9 | Other |
| 2 | Country |
| 5 | Rap |
| 1 | Rock |
| 4 | Pop |
| 3 | Urban- |
| | Contemporary |
| 1 | Rock |
| 3 | Urban- |
| | Contemporary |
| 8 | Jazz |
| 1 | Rock |
| 9 | Other |
| 9 | Other |
| 3 | Urban- |
| | Contemporary |
| 1 | Rock |
| 5 | Rap |
| 3 | Urban- |
| | Contemporary |
| 3 | Urban- |
| | Contemporary |
| 5 | Rap |
| 1 | Rock |
| 9 | Other |

Pie Chart

Pie Chart for the Textual Data



There are no warning messages.

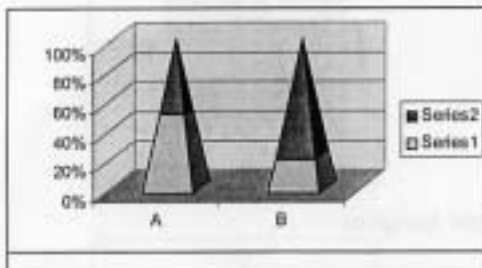
Pie Chart for Numerical Codes



The raw data were not converted into summary data. This is undocumented.

Pyramid Charts

Excel 100% Stacked Column with a Pyramid Shape Graph

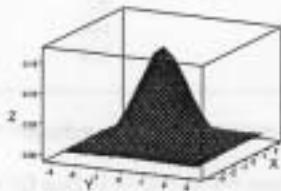


Default XY (Scatter)



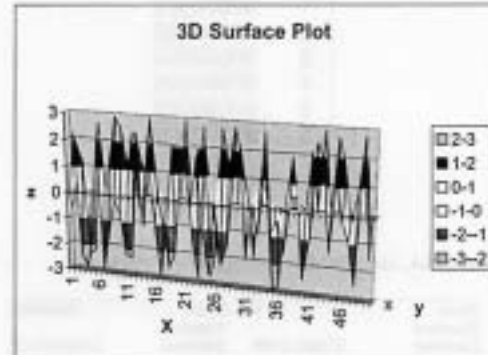
Bivariate Standardized Normal Distribution

Minitab 3D Representation of Bivariate Standardized Normal Distribution



Bivariate Standardized Normal Distribution

Excel 3D Representation of Bivariate Standardized Normal Distribution



Display of Digits

For example, consider the output from a regression involving the explanatory integers (1, 2, 3, 4, 5, 6, and 7) and corresponding response integers (18, 17, 22, 22, 24, 27, and 27), then Excel reports the Standard Error, t statistic (t Stat), P-value, and other values with an accuracy of nine digits.

Partial Excel Data Analysis ToolPak Regression Output

Presenting a False Sense of Accuracy

| Standard Error | t Stat | P-value |
|----------------|-------------|-------------|
| 1.066656037 | 14.46442985 | 2.84977E-05 |
| 0.238511541 | 7.337171166 | 0.000737717 |

Mode

Data ordered lowest to highest, mode = 183.
 Data ordered from largest to smallest, mode = 186.
 Data order in a random way, mode = 184.5

| | | |
|-------|-------|-------|
| 109 | 184.5 | 186 |
| 109.5 | 184.5 | 188.5 |
| 174 | 185 | 194.5 |
| 178.5 | 186 | 195 |
| 183 | 186 | 202.5 |
| 183 | 188.5 | 203 |
| 184.5 | 194.5 | 204 |
| 184.5 | 195 | 214 |
| 185 | 202.5 | 109 |
| 186 | 203 | 109.5 |
| 186 | 204 | 174 |
| 188.5 | 214 | 178.5 |
| 194.5 | 109 | 183 |
| 195 | 109.5 | 183 |
| 202.5 | 174 | 184.5 |
| 203 | 178.5 | 184.5 |
| 204 | 183 | 185 |
| 214 | 183 | 186 |
| Mode | Mode | Mode |
| 183 | 184.5 | 186 |

The first number is considered the modal value even if there are multiple modes.

Variability of Data

When large constants are added to the nine integers {1, 2, 3, 4, 5, 6, 7, 8, 9} the variance and standard deviation change for large constants. Adding 90000000 yields

| | |
|---|----------|
| 1 | 90000001 |
| 2 | 90000002 |
| 3 | 90000003 |
| 4 | 90000004 |
| 5 | 90000005 |
| 6 | 90000006 |
| 7 | 90000007 |
| 8 | 90000008 |
| 9 | 90000009 |

The mean, standard deviation, and variance:

| | | | |
|--------------------|-------------|--------------------|-------------|
| Mean | 5 | Mean | 90000005 |
| Standard Deviation | 2.738612788 | Standard Deviation | 2.828427125 |
| Sample Variance | 7.5 | Sample Variance | 8 |

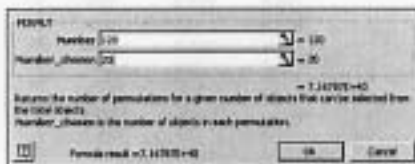
There are no warning messages.

Permutations and Combinations

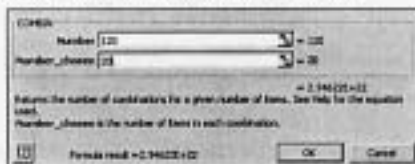
Find the number of ways one can choose 20 objects from 120 when order is important and when order is not important.

The correct answer for the number of permutations is found in statistical functions. The number of combinations is only found in Math & Trig functions.

The output from the permutation function:



The output from the combination function:



Quartiles

Find the upper and lower quartiles for a simple set of 11 numbers. Excel gives $Q1 = 35$ and $Q3 = 85$

| Position | Value |
|----------|-------|
| 1 | 10 |
| 2 | 20 |
| 3 | 30 |
| 4 | 40 |
| 5 | 50 |
| 6 | 60 |
| 7 | 70 |
| 8 | 80 |
| 9 | 90 |
| 10 | 100 |
| 11 | 110 |

A more conventional algorithm produces

$Q1$ is at position $= 1*(n+1)/4 = 3$. Thus $Q1 = 30$.
 $Q3$ is at position $= 3*(n+1)/4 = 9$. Thus $Q3 = 90$.

Probability Distributions

"TDIST is calculated as $TDIST = p(x < X)$, where X is a random variable that follows the t-distribution."

"NORMSDIST ... Returns the standard normal cumulative distribution function. The distribution has a mean of 0 (zero) and a standard deviation of one. Use this function in place of a table of standard normal curve areas."

Confidence “Interval”

Excel Output Producing Different Results
for Same Statistical Function

| Annotated CONFIDENCE Function Output | |
|--------------------------------------|--------------|
| Confidence Interval Function | 6.147563548 |
| Data Analysis ToolPak Output | |
| Mean | 33 |
| Standard Error | 3.136573806 |
| Median | 31 |
| Mode | #N/A |
| Standard Deviation | 12.14789811 |
| Sample Variance | 147.5714286 |
| Kurtosis | -1.013153254 |
| Skewness | 0.440527085 |
| Range | 39 |
| Minimum | 17 |
| Maximum | 56 |
| Sum | 495 |
| Count | 15 |
| Confidence Level(95.0%) | 6.727287728 |

Two Independent Sample t Test

From Excel Help Here is Excel Help on the t-Test:
Two-Sample Assuming Equal Variances Analyses.’

This analysis tool performs a two-sample Student’s t-test. This t-test form assumes that the means of both data sets are equal; it is referred to as a homoscedastic test. You can use t-tests to determine whether two sample means are equal.

Paired t-test

A Data set has two missing values.
There are three ways to do the paired t-test:
(1) include the blank, missing cells in his analysis;
(2) eliminate blank cells by shifting the data up in each column and;
(3) eliminate the students with incomplete data.

Results when blank cells are included

| t-Test: Paired Two Sample for Means | | |
|-------------------------------------|--------------|-------------|
| | Exam1 | Exam2 |
| Mean | 85.25806452 | 88.25806452 |
| Variance | 27.06451613 | 57.5311828 |
| Observations | 31 | 31 |
| Pearson Correlation | 0.405885548 | |
| Hypothesized Mean Difference | 0 | |
| df | 30 | |
| t Stat | -0.537530947 | |
| P(T<=t) one-tail | 0.297433071 | |
| t Critical one-tail | 1.697260359 | |
| P(T<=t) two-tail | 0.594866141 | |
| t Critical two-tail | 2.042270353 | |

Paired t-test

Results when the cells are shifted up:

| t-Test: Paired Two Sample for Means | | |
|-------------------------------------|--------------|-------------|
| | Exam1 | Exam2 |
| Mean | 85.25806452 | 88.25806452 |
| Variance | 27.06451613 | 57.5311828 |
| Observations | 31 | 31 |
| Pearson Correlation | 0.105538949 | |
| Hypothesized Mean Difference | 0 | |
| df | 30 | |
| t Stat | -1.912646539 | |
| P(T<=t) one-tail | 0.032691167 | |
| t Critical one-tail | 1.697260359 | |
| P(T<=t) two-tail | 0.065382334 | |
| t Critical two-tail | 2.042270353 | |

Results when using only complete data

| t-Test: Paired Two Sample for Means | | |
|-------------------------------------|-------------|------------|
| | Exam1 | Exam2 |
| Mean | 85.48275862 | 88.5862069 |
| Variance | 27.04433498 | 7 |
| Observations | 29 | 29 |
| Pearson Correlation | 0.405885548 | |
| Hypothesized Mean Difference | 0 | |
| Df | 28 | |
| t Stat | 2.315480216 | |
| P(T<=t) one-tail | 0.014065863 | |
| t Critical one-tail | 1.701130259 | |
| P(T<=t) two-tail | 0.028131727 | |
| t Critical two-tail | 2.048409442 | |

This last analysis gives the correct results.

Coefficient of Determination

Compute the value of the coefficient of determination (R-squared) for the following set of response and explanatory variables, that is due to Eakin (1996):

Response and Explanatory Values

| | |
|-----|------------|
| 1.1 | 10000000.1 |
| 1.9 | 10000000.2 |
| 3.1 | 10000000.3 |
| 3.9 | 10000000.4 |
| 4.9 | 10000000.5 |
| 6.1 | 10000000.6 |

Excel's RSQ function yields 2.092810987

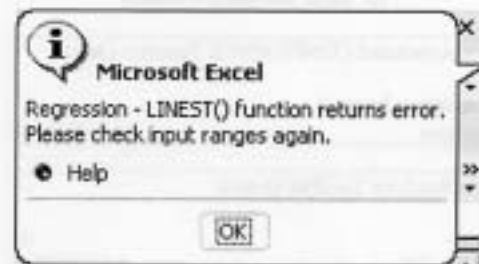
Excel's Regression Analysis Tool yields -0.816484141.

Values must be between 0 and 1.

The correct answer is 0.997.

Multiple Regression

Excel Error Message for Noncontiguous Data



Unreliable Algorithms

Partial Excel Data Analysis ToolPak
Output for Collinear Data

Regression Statistics

| | |
|-------------------|--------------|
| Multiple R | 65535 |
| R Square | -0.460636874 |
| Adjusted R Square | -3.65197E-09 |
| Standard Error | 0.000142846 |
| Observations | 9 |

Chi-square



Rank Correlation Coefficient

Compute the value of the rank correlation coefficient for

| Test 1 | Test 2 |
|--------|--------|
| 93 | 18.0 |
| 83 | 15.0 |
| 90 | 2.0 |
| 60 | 5.0 |
| 25 | 7.5 |
| 50 | 20.0 |
| 94 | 23.0 |
| 99 | 16.0 |
| 62 | 9.0 |
| 97 | 4.0 |
| 43 | 11.5 |
| 95 | 18.0 |
| 84 | 21.0 |
| 79 | 14.0 |
| 62 | 24.0 |
| 100 | 7.5 |
| 83 | 11.5 |
| 85 | 11.5 |
| 52 | 11.5 |
| 100 | 6.0 |
| 100 | 22.0 |
| 25 | 1.0 |
| 84 | 3.0 |
| 41 | 18.0 |

Excel yields 0.075593; the correct answer is 0.101.

Excel handles ranks as standings are reported in the world of sport. For example, if two teams are tied for second each receives a rank of 2.

Past and Future Fixes

Positive SS in Two-Way ANOVA

Improved Random Number Generation

Correlation(X,X)=1

Corrected Covariance Formula

And

Revised Functions in Excel 2003

Alternatives

Excel add-ins

Internet freeware

Student versions of statistical software

At a Minimum Cautions

USING MICROSOFT EXCEL FOR STATISTICS

As some researchers have noted (see reference 7), certain Microsoft Excel statistical capabilities contain flaws that can lead to invalid results, especially when data sets that are very large or that have unusual statistical properties are used. In this text, such flawed capabilities are either avoided or used with data that have been carefully chosen to minimize or eliminate the flaw. If you use Microsoft Excel (and the PHStat2 add-in) with any of the data sets of this text, you will always be able to draw the correct conclusion from the results produced by Microsoft Excel.

Because such an outcome cannot be guaranteed when using actual business data sets, the use of Microsoft Excel in this text does not constitute an endorsement by the authors of Excel's use in lieu of statistical packages in actual business settings. However, Microsoft Excel (with PHStat2) can be a convenient tool for becoming knowledgeable about statistics, even if you are required to use some other program at a later time.

Berenson, Levine, and Krehbiel (2004)

