Descriptive Statistics

Graeme Warren[1]
Leavey School of Business

I think all of us certainly believed the statistics which said that probably 88% chance of mission success and maybe 96% chance of survival. And we were willing to take those odds - Alan Shepard.

Learning Objectives

This note deals with descriptive statistics. The associated readings in (Diez et al, 2015) are sections 1.1, 1.2 and 1.6 - 1.8. A mind map of the content of this note (which connects closely with the learning objectives for this material) appears in Figure 1.

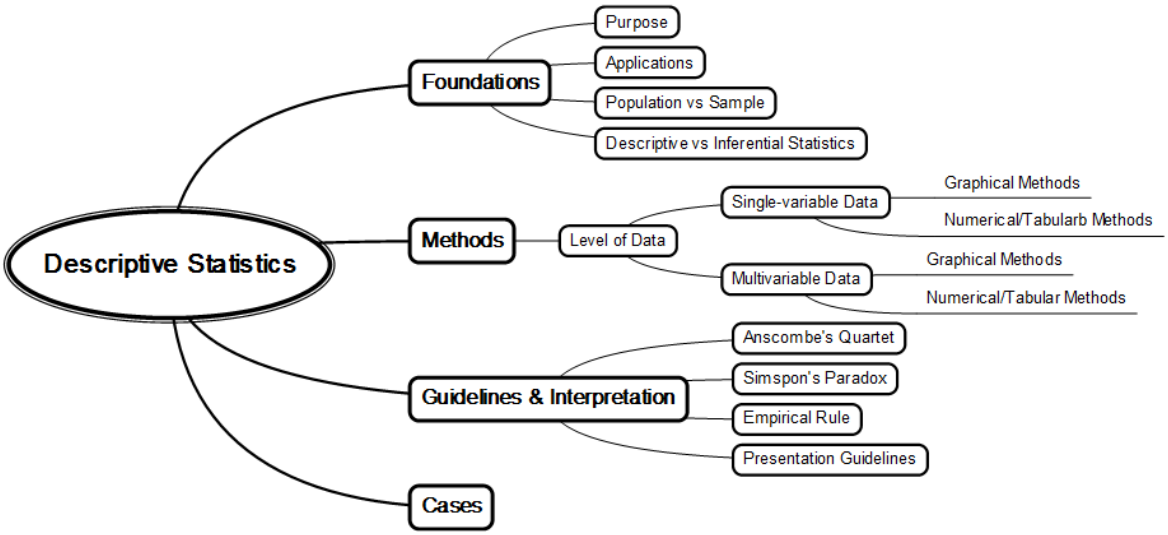

Figure 1. Mind map of note content.

After completing this module you should be able to:

1. Explain:
   (a) The concepts of statistics and descriptive statistics, why they are studied in this program, and some of their applications.
   (b) The difference between a sample and population, and the difference between descriptive statistics and inferential statistics.

   (c) The problem of induction.

2. Identify the level of data and interpret appropriate numerical, tabular and graphical methods of descriptive statistics including:
   (a) Single-variable graphical tools: bar charts, histograms, and box plots.
   (b) Single-variable numerical and tabular techniques:
      i. Frequency and relative frequency tables.
      ii. Measure of location: mean, median and mode.
      iii. Measures of relative standing: percentiles and quartiles.
      iv. Measures of dispersion: range, interquartile range, standard deviation, and variance.
      v. Miscellaneous measures: skewness and kurtosis.
   (c) Multiple-variable graphical tools: scatter plots and line charts.
   (d) Multiple-variable numerical and tabular techniques: contingency tables (cross tabs), correlation, covariance and the coefficient of determination.
3. Apply guidelines for presentation and interpretation of descriptive statistics including:
   (a) Implications of Anscombe's quartet.
   (b) Use of the empirical rule to interpret standard deviation, and
   (c) Implications of Simpson's paradox.
4. Use Microsoft Excel to produce numerical, tabular and graphical descriptive statistics.

Foundations

Statistics, Descriptive Statistics & Applications

Statistics is a way to make sense of data. More specifically, it is the science of collecting, cleaning, organizing, storing, analyzing and presenting data to enable better decision making.

Got it, but why would anyone want to *study* statistics? There are at least two reasonable responses to this question:

1. *Better* business decisions are *data-informed*, and statistics is the science underpinning the move from *data* to *informed!*
2. Knowledge of statistics is needed to be conversant with best practices in the business world.

This document deals with descriptive statistics. Descriptive statistics, a.k.a. summary statistics, are numerical, tabular and graphical methods for summarizing and presenting data in meaningful ways. Numeric descriptive statistics may describe characteristics of a data set such central tendency (using measures such as the mean, medium and mode), variability (using such measures as the standard deviation, range and interquartile range) and other features of the data. Graphical descriptive statistics are chart- and graph-based visualizations that help us to understand how data is distributed.

Common applications of descriptive statistics include:

- Dashboards - used to present key performance indicators (KPIs). Have a look at some examples.
- Exploratory data analysis. For an example, see the Google public data explorer.

Check your knowledge by reviewing answers to the following questions:

- What is statistics?
- What are descriptive statistics?
- Why study statistics?
- Identify applications of descriptive statistics.

Descriptive Statistics versus Inferential Statistics

It is worthwhile to establish an early contrast between descriptive statistics and inferential statistics to further understand the aims of descriptive statistics. To do so we will need a few definitions. A *population* is the entire collection of objects, individuals or data items of interest. It is measured in a *census*. A *sample* is subset of objects, individuals or data items drawn from a population of interest. Figure 2 illustrates this distinction. The process of obtaining a sample is *sampling*. While sampling is less costly and can be executed quicker than performing a census, it immediately raises the question of how representative a sample is of the population.

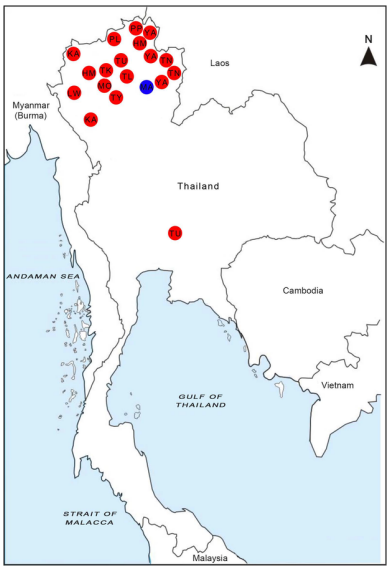

Figure 2. Geographical distribution of Thailand population samples. Blue dots on the map indicated sampling locations of Mlabri (MA). Red dots on the map indicated sampling locations of other populations. Shuhua Xu, Daoroong Kangwanpong, Mark Seielstad, Metawee Srikummool, Jatupol Kampuansai, Li Jin and The HUGO Pan-Asian SNP Consortium. CC-BY 2.0 license.

Descriptive statistics are used to organize, analyze and present whatever data is at hand and do not assume that the data comes from either a sample or population. The

techniques of inferential statistics, on the other hand, seek to draw conclusions about a population from sample data. Examples of inferential statistics include point and interval estimates, significance testing and regression analysis.

The reliability and validity of inferences can be improved, but cannot, in general, be perfected, by obtaining samples that are highly representative of their underlying population by careful sample collection and processing protocols. Much of the methodological structure in inferential statistics is aimed at improving the reliability and validity of inferences and addressing a fundamental problem that is widely attributed to the 18th-century Scottish philosopher David Hume. In his famous text, *A Treatise of Human Nature*, Hume raises the problem of induction (a.k.a the black swan problem), which alleges that one cannot generalize about the properties of a set of objects based upon observations of particular instances of the set. For example, absent knowledge that a few specimens of blue lobster (see Figure 3), which are estimated to have an occurrence rate of about one in two million, have been observed, one might argue that all American lobsters are non-blue. The consequence of this is that rare observations - the black swans and blue lobsters - are unlikely to be represented in a sample and this may influence our inferences. The problem of induction means that

Inferences about populations based upon sample data cannot be guaranteed.

We will concede that inference is always subject to potential for sampling error, and proceed on the assumption that the techniques and practices of statistics represent a pragmatic foundation for statistical inference.

Figure 3. Blue lobster, Shedd Aquarium, Chicago, 03/24/2017. Sadly, an agent of the author reported that the lobster was "no longer resident" (sob) as of September 2017.

Check your knowledge by reviewing answers to the following questions:

- What is the difference between a sample and a population?
- What is the problem of induction?
- What is the black swan problem?
- What is[2] the relationship between sampling, statistical inference and the problem of

[2]The use of samples may produce inferences that are incorrect. The problem of induction shows us that low-probability (and possibly previously unknown) outcomes may invalidate inferences.

induction?

Methods

This section starts by describing the different levels (types) of data. This is followed by a discussion of numerical, tabular and graphical descriptive statistics for single-variable and multiple-variable data available for each level of data.

The most important graphical tools of descriptive statistics are the bar chart (both simple and grouped/stacked), histogram, box plot, scatter plot, and line chart. These tools allow us to get a sense of how variables in the data, both individually and in combination, behave. These tools provide significant visualization capability in almost all situations. Nevertheless, a variety of new static, interactive and animated graphical tools and dashboards such as heat maps, choropleth maps and bubble charts are relevant and powerful tools in specific contexts and applications.

Levels of Data

The numerical, tabular and graphical techniques of descriptive tools that can be used in a specific context depend upon the type (level) of data involved. There are four levels of data:

1. Qualitative Data
   (a) Nominal-level data
   (b) Ordinal-level data
2. Quantitative Data
   (a) Interval-level data
   (b) Ratio-scale data

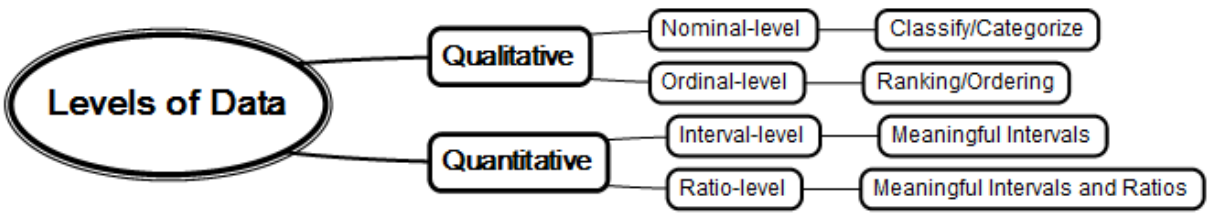

Figure 4. Mind map of levels of data.

The ideas in this section are summarized in Figure 4.

Nominal-level data, the most basic level of data, are used to categorize or classify. Nominal-level data are qualitative (non-quantitative). Examples of nominal-level variables include gender and nationality data. The only numerical descriptive statistics possible on

nominal data are those pertaining to counts. We can, for example, tabulate the frequency or relative frequency of each category - see Table 1 for an example of a frequency (number of speakers) and relative-frequency table (percentage of all speakers) of the top ten languages spoken[3] in the San Francisco Bay area.

Language	Frequency	Relative Frequency
English	2,485,786	64.2%
Spanish	540,255	14.0%
Chinese	278,285	7.1%
Tagalog	142,055	3.7%
Vietnamese	38,980	1.0%
Russian	29,770	0.8%
French	28,585	0.7%
Korean	25,830	0.7%
Persian	24,495	0.6%
German	24,065	0.6%

Table 1. *Frequency and relative-frequency table of nominal-level data of the top ten languages spoken in the San Francisco Bay area. The percentages do not total to 100% because there are more than 100 other languages spoken in the Bay area.*

Next we have ordinal-level data, which are ordered or ranked data. Ordinal-level data are also qualitative (non-quantitative). Examples include competition rankings and Likert-scale opinion ratings. See Table 2 for a frequency and relative-frequency table of Likert-scale opinion data. Note how the categories of an ordinal scale are ordered.

Opinion	Frequency	Relative Frequency
Excellent	23	36.9%
Very Good	29	45.3%
Good	9	14.1%
Poor	1	1.6%
Very Poor	2	3.1%

Table 2. *Frequency and relative-frequency table of ordinal-level Likert-scale opinion data.*

Interval-level data are quantitative (e.g., real-valued). Examples of interval-level data include credit scores, and scores on a variety of standardized tests such as the SAT and GMAT. Interval-level data have meaningful intervals. That is, the distance between data values is meaningful. For example, consider two credit scores $A = 800$ and $B = 700$. The difference (100) between A and B is meaningful. Ratios of interval-level data are *not* meaningful. It is *not* meaningful to say that a person with credit score B is $800/700 \approx 1.14$ times as creditworthy as a person with credit score A. Crucially, it makes no sense to analyze ratios of interval-level data using descriptive statistics because they are not meaningful.

[3] See, e.g., http://www.sfgate.com/bayarea/article/BAY-AREA-Report-112-languages-spoken-in-2692403.php

Finally, ratio-level data have all the properties of all the preceding levels in addition to having meaningful ratios and a meaningful zero. It is therefore reasonable to present descriptive statistics of ratios of ratio-level data. Ratio-level data are quantitative. Examples include scientific data such as length or density, economic data such as gross domestic product (GDP), and historical accounting data such as earnings per share (EPS). Consider the GDP per capita of two countries. Suppose country C, respectively D, has a GDP per capita of \$30,000, respectively \$60,000. It is meaningful to calculate the ratio of the two and say that the GDP per capita of D is twice that of C.

Check your knowledge by reviewing answers to the following questions:

- What are the four levels of data and their properties?
- What is the data level[4] of Transunion credit scores?
- What is the data level[5] of the price of a security?
- What is the data level[6] of the exchange(s) on which a security trades?
- What is the data level[7] of MCAT scores?
- What is the data level[8] of the analyst recommendation of a security on tipranks.com?

Single-Variable Descriptive Statistics

While dashboard and data visualizations are continually being innovated, the methods discussed below are tools suitable for presenting numerical, tabular and graphical descriptive statistics for single-variable data. The maxim

"first chart your data"

(to get an early sense of relationships in the data) is very important (see the discussion of Anscombe's Quartet later in this document) and widely practiced. The single-variable methods (depicted in a mind map in Figure 5) we consider are:

1. Single-variable graphical tools for nominal/ordinal data: bar charts.
2. Single-variable numerical and tabular tools for:
   (a) Nominal data: mode, counts, frequency tables and relative frequency tables.
   (b) Ordinal data: median, counts, frequency tables and relative frequency tables.
3. Single-variable graphical tools for interval/ratio data: histograms, and box plots.
4. Single-variable numerical techniques for interval/ratio data:

[4]Interval level.
[5]Ratio level.
[6]Nominal level.
[7]Interval level.
[8]The recommendations {Buy, Hold, Sell} are ordinal level.

(a) Measure of location: mean, median and mode.

(b) Measures of relative standing: percentiles and quartiles.

(c) Measures of dispersion: range, interquartile range, standard deviation, and variance.

(d) Miscellaneous measures: skewness and kurtosis.

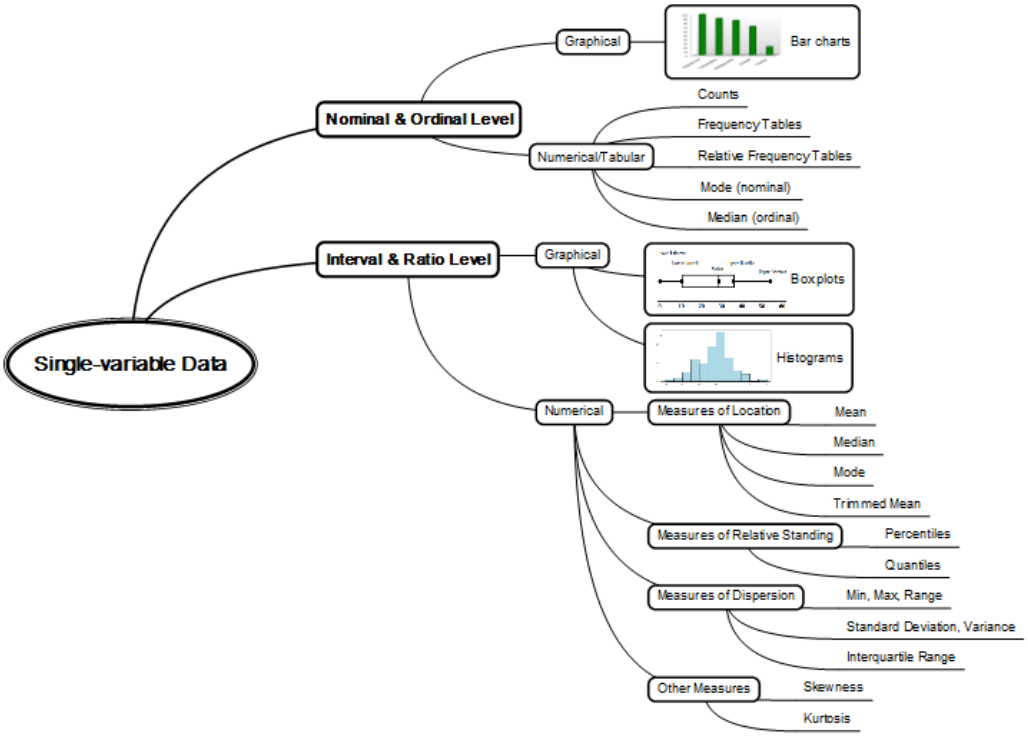

Figure 5. Mind map of single-variable descriptive statistics.

HOWTOs for producing frequency tables, contingency tables and bar charts in Microsoft Excel are available on lynda.com. HOWTOs are also available elsewhere (e.g., charts & graphs, pivot tables).

We will use data[9] about the celebrities who died in 2016 to illustrate concepts in this section.

Get the celebrity death data here!

A few sample records of the celebrity death data appear in Table 3. The human species appears to be over-represented in the data set! The meaning of the fields/variables

[9]The data has been adapted from a data set obtained from https://github.com/Antony74/celebrity-death-model

of the data set, namely `date of death, MONTH, DAY, name, age, bio, cause of death`, is mostly evident from the field names. `MONTH` (month of death) and `DAY` (weekday of death) are the month (January=1) and weekday (Sunday=1) of death that have been extracted from `date of death`. Notice that the data set contains many missing or uncertain entries. Specifically, the `cause of death` field is grossly incomplete. To repeat,

Date of Death	MONTH	DAY	Name	Age	Bio	Cause of Death
2016-01-01	1	6	Tony Lane	71	... art director	brain cancer
2016-01-01	1	6	Gilbert Kaplan	74	... conductor	cancer
2016-01-01	1	6	Brian Johns	79	... company	cancer
2016-01-01	1	6	Natasha Aguilar	45	... swimmer	... stroke

Table 3. *A few records of the 2016 celebrity death data set.*

selected analysis in Excel ("Graphical" and "Numerical" tabs) and the raw celebrity death data ("Raw Data") appears here.

Single-Variable Graphical Methods for Nominal- and Ordinal-Level Data. The most commonly used charting tool for single-variable nominal- and ordinal-level data is the bar chart. A variety bar chart types exist: horizontal/vertical and 2D/3D. Pie (and doughnut) charts are not recommended because they provide insufficient visualization power when the data produces thin pie slices (see, e.g., Figure 6). Bar charts of the `DAY` and `MONTH` data in the celebrity death data set appears in Figure 7.

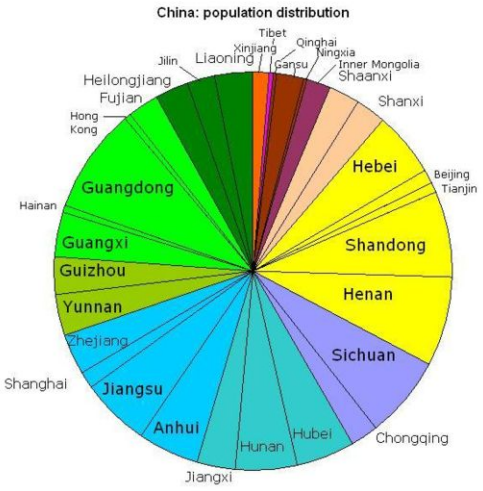

Figure 6. Pie chart showing population distribution across the People's Republic of China by province/municipality/autonomous region. Colors indicate broad regional subdivisions. Note the problem of thin slices. Macau does not have a population large enough to represent it on the chart. Public domain.

The same information could, as an alternative, be presented in frequency tables. From

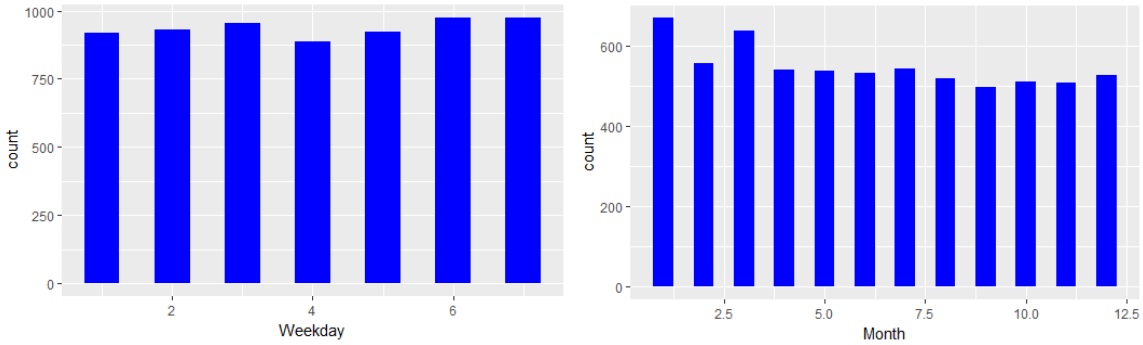

Figure 7. Bar charts of the `DAY` and `MONTH` fields in the 2016 celebrity data.

the charts it appears that more celebrities die in January or March but we cannot be sure whether these conclusions are significant without additional analysis.

Single-Variable Numeric Methods for Nominal- and Ordinal-Level Data. The most commonly used numerical and tabular tools for single-variable nominal- and ordinal level data are: counts, frequency tables, relative-frequency tables, the mode for nominal data, and the median for ordinal data. We have previously seen examples of some of these in Tables 1 and 2. The key difference between the nominal and ordinal levels of data is the requirement to preserve the ordering of the categories in the case of ordinal-level data.

The modes of the `DAY` and `MONTH` variables in the 2016 celebrity death data set are Saturday and Sunday (both 975 deaths) for the `DAY` variable and January (669 deaths) for the `MONTH` variable.

Single-Variable Graphical Methods for Interval- and Ratio-Level Data. All of the tools that can be used on nominal- and ordinal-level data can be applied to interval- or ratio-level data. The most common charting tools specific to single-variable interval- or ratio-level data are histograms and box plots. Whereas bar charts have categorical axis labels, histograms have a real-valued axis. Two histograms of the age of celebrities who died in 2016 are shown in Figure 8; the histogram on the left shows the frequency of age for 10-year buckets whereas the histogram on the right shows the age distribution for each age. Notice how the bucketing in the histogram washes out some of the details in the data.

Box plots vary significantly in convention. It is therefore common to note the meaning of the whiskers in a caption. It is possible that box plots will be supplanted by violin plots, which show the information in the box plot in addition distribution information. Unfortunately, violin plots are not available in Microsoft Excel at the time of writing.

Single-Variable Numeric Methods for Interval- and Ratio-Level Data. Here we consider numerical single-variable descriptive statistics for location (e.g., mean, median and mode), relative standing (e.g., percentiles and quantiles), dispersion (e.g., range and standard deviation) and miscellaneous distributional characteristics (e.g., skewness).

- **Measures of Location**. The most commonly used numerical statistics are those related to location, of which the mean, median and mode are dominant.

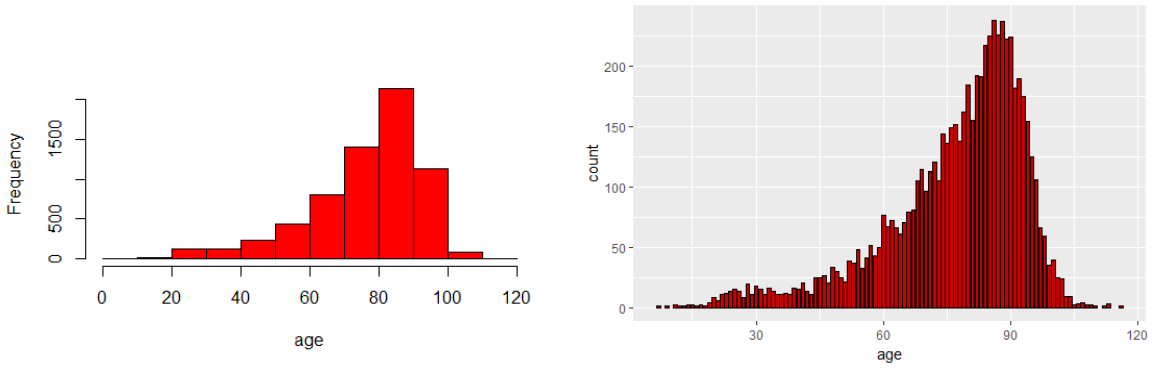

Figure 8. Histograms of the `age` of celebrities who died in 2016.

Hey diddle diddle, the Median's the middle; you add and divide for the Mean. The Mode is the one that appears the most, and the Range is the difference between.

Origin unknown

– The mean (a.k.a. the average) is an arithmetic average of the data items. If $x_1, x_2, \ldots, x_n$ are the n items in the data set, then the mean is defined as

$$\frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

For example, the mean of the data set $\{1, 2, 3, 4, 5\}$ is 3. This result can be obtained in Excel using the formula `=AVERAGE(1,2,3,4,5)`. Summary statistics can be completed using the Descriptive Statistics feature of the "Analysis Toolpak" add-in in Excel. Note that the Toolpak provides calculations for some statistics (such as the standard error) that we will not discuss here. Confirm that the mean of the age variable in the celebrity death data set is 77.03.

– The median of a data set is the datum separating the higher half of an ordered data set from the lower half. The median is therefore the "middle" value of a data set. For example, in the data set {1, 2, 3, 9, 12, 13, 129}, which has an odd number of data, the median is 9. If a data set has an even number of data (of say n items) then the median is the average of datum $n/2$ and datum $n/2 + 1$. For example, the median of the data set {1,3,9,24} is 6, the average of 3 and 9. These results can be corroborated in Excel using the formula `MEDIAN(1,3,9,24)`. Confirm that the median of the celebrity death age data is 81.

– The mode of a data set is the most commonly occurring value. A data set may have several modes (all of which would then have the same frequency). The Excel formula `MODE.SNGL(`*data range*`)` reports the mode of the age distribution depicted in Figure 8 to be 86.

– The trimmed mean is calculated as the average of a reduced data set obtained by dropping a certain number or percentage of observations at either end of the ordered data set). The Excel formula `TRIMMEAN(`*data range, percentage*`)`

accomplishes this calculation for the specified *percentage*. For example, the 10% (5% trimmed on either end of the data set) trimmed mean of the age variable in the celebrity deaths data set is 79.11.

To summarize, for the age distribution in the celebrity death data set, the mean is 77.03, the median 81, the mode 86, and the 10% trimmed mean is 79.11. Which measure of location do *you* prefer to describe the central tendency of the age distribution? One reasonable answer to this question is to prefer robust statistics - statistics that perform reliably for a variety of data distributions. The median and trimmed mean are more robust (less susceptible) than the mean or mode to communicate information about the central tendency of distributions that are non-symmetric and/or have a significant number of outliers.

Use robust statistics when the data is skewed or there are significant outliers.

- **Measures of Relative Standing**. The most commonly used numerical measures of relative standing are percentiles and quantiles.
  - A percentile (or centile) is the value below which a given percentage of observations in a group of observations fall. Percentiles may be calculated in Excel using the formula `PERCENTILE.INC(`*data range,k*`)` where $100k$ is the specified percentage.
  - Quartiles are a special case of percentiles. The first quantile is the 25th percentile, the median (a.k.a. the second quartile) is the 50th percentile, and the third quantile is the 75th percentile. Note that there are at least two widely-accepted methods for calculating percentiles - so do not be surprised when different software tools report different quartiles! Box plots usually graph the minimum, maximum, and first, second and third quartiles. Quartiles may be calculated in Excel using the formula `QUARTILE(`*variable,k*`)`. Confirm that the first and third quartiles of the celebrity death data set are, respectively, 69 and 89.
- **Measures of Dispersion**. Common numerical statistics for the dispersion of data include the range, standard deviation, variance, interquartile range and coefficient of variation.
  - The range (the difference between the maximum (largest datum) and minimum (smallest datum) is a relatively crude measure of the dispersion of a data set. For example, the minimum, maximum and range of the data set $\{1, 2, 3, 4, 5\}$ are, respectively, 1, 5 and 4. These results can be confirmed using the Excel formulas `MIN(`*variable*`)` and `MAX(`*variable*`)`. Notice that the age at death of celebrities who died in 2016 varies considerably - between a minimum age of 7 and a maximum age 116. The range is therefore 109 years.
  - The standard deviation is one of the most important measures of dispersion. It measures dispersion of data around the mean. The higher the standard deviation the greater the dispersion of data. There are two definitions that depend upon

whether the data is treated as a population or sample. The distinction is practically important only when the number of data records is small. For big data (i.e., large n), the two measures are, for most business purposes, numerically equivalent. Suppose that $x_1, x_2, \ldots, x_n$ are the n items in a data set.

The population standard deviation is the standard deviation *of* the data and is defined as

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}} = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_n - \mu)^2}{n}}$$

where μ is the population mean. Use the Microsoft Excel formula `STDEV.P(`*data range*`)` to compute the population standard deviation.

The sample standard deviation, which is calculated for data considered to be a sample, is an *estimate of the standard deviation of the associated population*. It is calculated as

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}}$$

where $\bar{x}$ is the sample mean. Use the Microsoft Excel formula `STDEV.S(`*data range*`)` (or simply `STDEV(`*data range*`)`) to estimate the population standard deviation using sample data.

- The variance is simply the square of the standard deviation. It can be calculated in Excel using `VAR.P(`*data range*`)` for the population variance and `VAR.S(`*data range*`)` for the sample variance. The sample variance is an estimate of the population variance, not the variance of the sample itself. Let's assume that the celebrity death data set is a population. Confirm that its standard deviation and variance are 16.45 and 270.4 respectively.
- The interquartile range (IQR) is the difference between the third and first quartiles of the data. There is no specific Microsoft Excel function for it. Confirm that the interquartile range of the age variable in the celebrity death data set is 20.
- The coefficient of variation is another measure of dispersion. It is simply the standard deviation divided by the mean. Confirm that the coefficient of variation for the age distribution of the celebrity death data set is 0.213.

- **Other Measures**. Two other measures of interest are skewness and kurtosis.
  - Skewness measures the extent to which a data distribution is asymmetric. Loosely speaking, if the distribution has a lengthy upper tail then the distribution is called positively (or right) skewed. If the lower tail is lengthy then the distribution is called negatively (or left) skewed. The histograms in Figure 8 are negatively skewed. A symmetric distribution has zero skewness and equal mean and median. Table 4 shows the interpretation of skewness values due to Bulmer[10]. Skewness can be computed in Excel using the `SKEW(`*data range*`)` formula. The age distribution of the celebrity death data set has a skewness of -1.22.

[10] Bulmer, M. G. (1979). *Principles of Statistics*. Dover.

Skewness	Interpretation
In the range [-0.5,0.5]	Approximately symmetric
In the range [-1,-0.5) or (0.5,1]	Moderately skewed
Less than -1 or greater than 1	Highly skewed

Table 4. *Bulmer's rule of thumb for skewness.*

– Kurtosis is a measure of the heaviness of the tails of the distribution compared to the tails of the normal distribution. The implementation of kurtosis depends upon the software package used. If kurtosis is greater than 0 (in Excel) or 3 (in R) then the distribution has more outliers (is heavier tailed) than a normal distribution and it is called *leptokurtic*. The kurtosis of a normal distribution is 0 (in Excel). If kurtosis is less than 0 (in Excel) or 3 (in R) then the distribution has fewer outliers (is lighter tailed) than a normal distribution and it is called *platykurtic*. These ideas are summarized in Table 5. The visual difference between platykurtic and leptokurtic distributions can be viewed here. Kurtosis can be calculated in Excel using the `KURT(`*data range*`)` formula. Confirm that the kurtosis of the age distribution in the celebrity data set is 4.49, indicating that its tails are heavy (that it has more outliers) compared to the normal distribution.

Excel Kurtosis	Interpretation
> 0	Leptokurtic - heavier tails than a normal distribution
< 0	Platykurtic - lighter tails than a normal distribution

Table 5. *Interpretation of kurtosis.*

The single-variable statistics discussed to date for the 2016 celebrity death set are summarized in Table 6.

Statistic	Value	Statistic	Value
Mean	77.03	Minimum	7
Median	81	Maximum	116
Mode	86	Range	109
10% Trimmed Mean	79.11	Standard Deviation	16.45
Geometric Mean	74.63	Variance	270.44
1st Quartile	69	Skewness	-1.22[a]
3rd Quartile	89	Kurtosis	1.49[b]
Interquartile Range	20	Coefficient of Variation	0.213

Table 6. *Single-variable statistics for the* `age` *variable in the 2016 celebrity death data set. Note that kurtosis was computed here using Excel.*

[a] Highly negatively skewed.
[b] Leptokurtic (heavier tails/more outliers than normal distribution).

1. What are[11] the recommended single-variable graphical tools?
2. What are[12] the most important measures of location?
3. What are[13] the most important measures of relative standing?
4. What are[14] the most important measures of dispersion?
5. What miscellaneous measures were[15] discussed?
6. What are[16] robust statistics?
7. How can[17] one interpret standard deviation?
8. How can one interpret skewness?
9. How can one interpret kurtosis?
10. What is[18] the difference between a bar chart and a histogram, and is the difference important?

Multiple-Variable Descriptive Statistics

We now consider numerical, tabular and graphical methods of descriptive statistics for multivariate settings, including:

1. Multiple-variable graphical tools for two qualitative (i.e., nominal- or ordinal-scale) variables: grouped or stacked bar charts.

2. Multiple-variable tabular tools for two qualitative variables: contingency tables (cross tabs).

3. Multiple-variable graphical tools for two quantitative (i.e., interval or ratio-scale) variables: scatter plots and line charts.

4. Multiple-variable numerical techniques for two quantitative variables: correlation, covariance and the coefficient of determination.

5. Multiple-variable graphical tools for a quantitative and qualitative variable: stacked histograms and side-by-side box plots.

6. Multiple-variable tabular tools for a quantitative and a qualitative variable: contingency tables (cross tabs) of numerical statistics.

These tools are overviewed in the mind map shown in Figure 9. HOWTOs for producing histograms, box plots and scatter plots in Microsoft Excel are available on `lynda.com`.

[11] Bar charts, histograms, and box plots.
[12] Mean, median and mode.
[13] Percentiles and quartiles.
[14] Range, interquartile range, standard deviation, and variance.
[15] Skewness and kurtosis.
[16] Statistics that are less sensitive to non-symmetric distributions and/or outliers. We discussed the median and trimmed mean as examples of robust statistics (as compared to the mean and mode).
[17] It is a measure of the spread of data.
[18] Bar charts are used to present nominal or ordinal data, histograms are used to present interval or ratio-scale data. The distinction is important because one has to appropriately label the horizontal axis in each case.

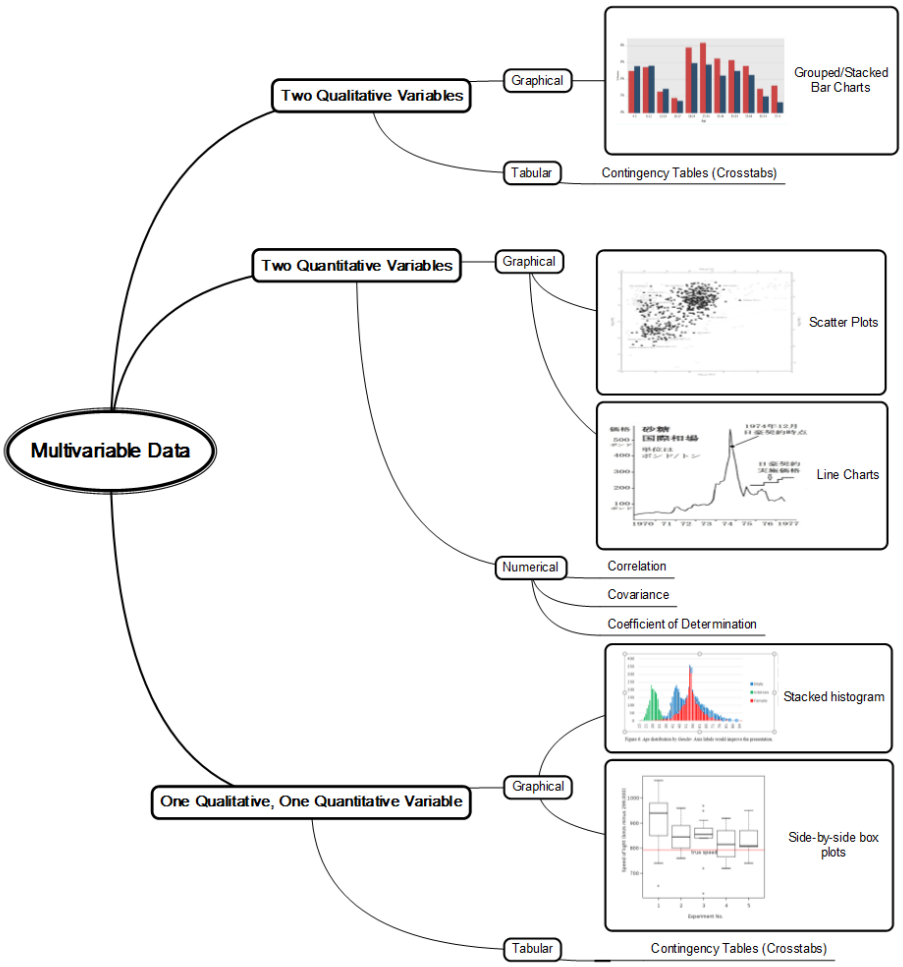

Figure 9. Mind map of multiple-variable descriptive statistics.

Multiple-Variable Graphical Methods for Presenting Two Qualitative Variables. Grouped or stacked bar charts are the most common charts for presenting multiple-variable nominal- and ordinal-level data. Heiberger and Robbins argue for the use of diverging stacked bar charts for Likert and other ranking data (such as semantic differential scale data). A sample diverging stacked bar chart appears in Figure 10.

Multiple-Variable Tabular Method for Presenting Two Qualitative Variables. The tabular tool for presenting multiple-variable nominal- and ordinal-level data is the contingency table (a.k.a. cross-classification table or crosstab). Contingency tables are essentially frequency tables with multiple variables. Table 7 shows contingency tables of information related to the busiest domestic and international route data[19] from Norman Y. Mineta San José International Airport (airport code SJC).

[19]Data obtained from https://en.wikipedia.org/wiki/San_Jos%C3%A9_International_Airport on December 21, 2016.

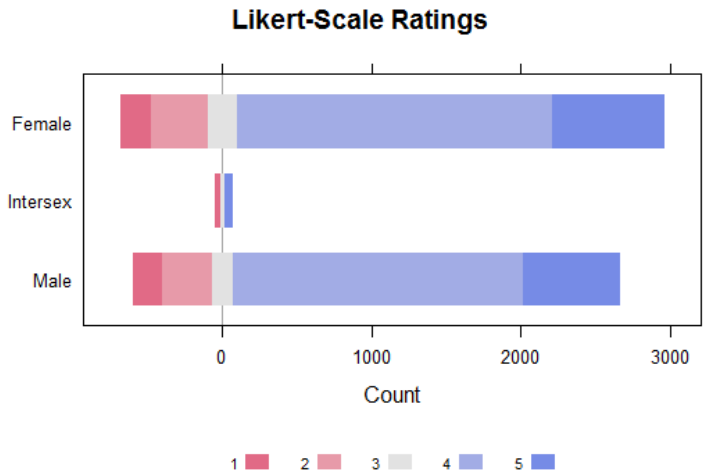

Figure 10. Diverging stacked bar chart of Likert-scale rating data data by gender.

Rank	Domestic Passengers	International Passengers
1	586K	104K
2	501K	95.3K
3	397K	87K
4	382K	24K

Rank	Domestic Route	International Route
1	Los Angeles	Tokyo
2	Seattle	San José del Cabo
3	Las Vegas	Guadalajara
4	Phoenix	Beijing

Table 7. *Contingency tables of 2015-2016 SJC flight data and associated routes.*

An extract of the information in Table 5 of Campbell, D., & Frei, F. (2010)[20] is shown in the contingency table presented in Table 8 of this document. It depicts the number of self-service transactions in various service channels for three research groups - the control, passive-adopter and active-adopter groups - in an experiment that aimed to determine (some years ago) whether consumer adoption of online banking reduces the total cost of serving the customer. The table shows the mean transaction cost and mean number of transactions for each group for each service channel before and after the study. We will explore data underlying this table in our forthcoming discussion of interval estimates.

Multiple-Variable Graphical Tools for Presenting Two Quantitative Variables. The most common charting tools specific to multiple-variable interval- or ratio-level data are scatter plots and line charts. Line charts are commonly used for time-series data (data that captures the time evolution of a variable of interest). Line charts are scatter plots with ordered data points that are connected by lines. Figures 20 and 21 are examples of, respectively, a line chart and a scatter plot.

Multiple-Variable Tabular Tool for Presenting Two Quantitative Variables. We now turn to recommendations for multiple-variable numerical statistics for interval- and

[20]Campbell, D., & Frei, F. (2010). Cost structure, customer profitability, and retention implications of self-service distribution channels: Evidence from customer behavior in an online banking channel. *Management Science*, 56(1), 4-24.

Channel	Cost/ Transaction	Control Before	Control After	Passive Before	Passive After	Active Before	Active After
Branch	1.34	1.24	1.34	1.51	1.73	1.75	2.28
Call Center	0.39	0.18	0.22	0.19	0.24	0.31	0.35
ATM	0.16	2.12	2.3	2.48	2.7	3.41	3.54
IVR	0.13	1.41	1.57	1.51	1.4	2.9	1.45
Online	0.035	0	0	0	1.39	0	9.91

Table 8. *Contingency table based upon information extracted from Table 5 of Campbell, D., & Frei, F. (2010). IVR = Interactive Voice Response.*

ratio-level data, considering three measures: covariance, correlation and the coefficient of determination.

- Covariance is a measure of how changes in two quantitative variables are related. It is a useful measure of their linear association. Assume that pairs of (x_i, y_i) data with $i = 1, \ldots, n$ of variables x and y are available.

  If the data is a population then the population covariance is defined as

  $$\sigma_{xy} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{n}$$

  where μ_x and μ_y are the population means of, respectively, x and y. Use the Excel formula `COVARIANCE.P`(*data range*) to compute the population covariance.

  The sample covariance, which is an *estimate* of the population covariance (and not the covariance of the sample), is defined as

  $$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

  where $\bar{x}$ and $\bar{y}$ are the sample means of, respectively, x and y. Use the Excel formula `COVARIANCE.S`(*data range*) to estimate the population covariance.

- Pearson's correlation coefficient, the most commonly encountered correlation measure, is a scaled variant of covariance and thus also a measure of the linear relationship between two variables. Pearson's correlation coefficient (which is computed in the same manner for samples and populations) is defined as

  $$\rho = \frac{s_{xy}}{s_x s_y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

  where s_x and s_y are, respectively, the sample standard deviations of x and y, and σ_x and σ_y are, respectively, the population standard deviations of x and y.

  Pearson's correlation coefficient is real-valued on the range [-1,1]. A correlation of +1 indicates that x increases as a constant multiple of increases in y (and vice versa). A correlation of -1 indicates that x increases as a constant multiple of decreases in y (and vice versa). Note that correlation does not imply causality. Pearson correlation indicates the strength of linear relationship between two variables.

Correlation coefficients may be calculated in Excel using the `CORREL(`*x-range*`,`*y-range*`)`.

Let us consider correlation-coefficient and sample-covariance calculations for three small data sets:

1. First, suppose that $x = \{1, 2, 3, 4, 5\}$ and $y_1 = \{2, 4, 6, 8, 10\}$. The data is presented in Figure 11. We get $s_{x,y_1} = 5$ and $\rho_{x,y_1} = 1$. x and y_1 are perfectly positively correlated.
2. Now suppose that $x = \{1, 2, 3, 4, 5\}$ and $y_2 = \{4, 1, 3, -1, 3\}$. The new data is presented in Figure 12. We get $s_{x,y_2} = -1$ and $\rho_{x,y_2} = 0.316$.
3. Finally, suppose that $x = \{1, 2, 3, 4, 5\}$ and $y_3 = \{10, 8, 6, 4, 2\}$. The new data is presented in Figure 13. We get $s_{x,y_3} = -5$ and $\rho_{x,y_2} = -1$. x and y_3 are perfectly negatively correlated.

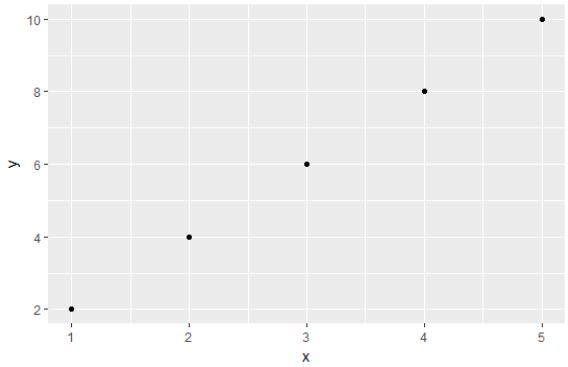

Figure 11. Graph of x and y_1.

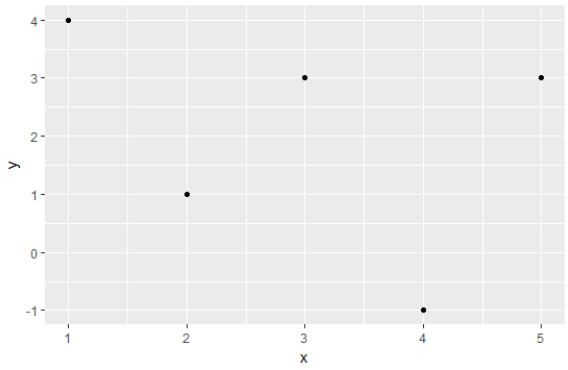

Figure 12. Graph of x and y_2.

- Other correlation measures such as Spearman's rank correlation coefficient are better able to accommodate non-linear relationships but will not be considered here. Measures such as multiple correlation and the coefficient of determination, often denoted

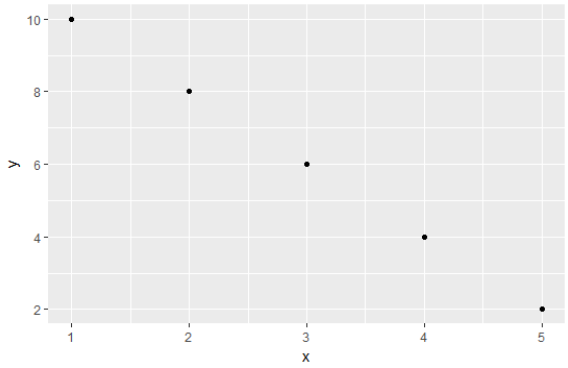

Figure 13. Graph of x and y_3.

by R^2, address the predictability of a variable by a linear function of multiple predictor variables. Further discussion of R^2 will be pursued later in the course. It can be calculated in Excel using the formula `RSQ(`*x-data range,y-data range*`)`.

Multiple-Variable Graphical Tools for Presenting a Quantitative and a Qualitative Variable. Options for presenting the interaction between a qualitative and a quantitative variable include stacked histograms and side-by-side box plots. See Figure 14 for a stacked histogram of the `age` variable of the 2016 celebrity death data set that is color coded by `DAY` of death. See Figure 19 for an example of side-by-side box plots in which the qualitative variable identifies the exchange traded fund (ETF) on the horizontal axis and the quantitative variable captures the daily fractional price change (quantitative variable) on the vertical axis.

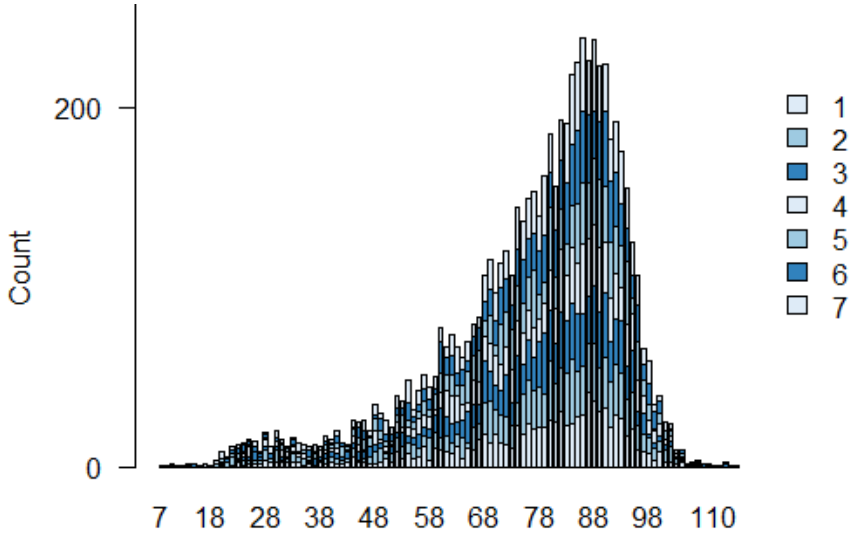

Figure 14. Stacked histogram of `age` color coded by `DAY` in the 2016 celebrity death data.

Multiple-Variable Tabular Tool for Presenting a Quantitative and a Qualitative Variable. A contingency table may be used to present the statistics of a quantitative variable for the different categories of a qualitative variable. For example, the crosstab shown in Table 9 identifies breaks out statistics for a ratio-scale age variable by gender, which is a qualitative variable.

	Gender		
Statistic	Female	Intersex	Male
Mean	48.6	34.4	52.38
Median	49	35	52
Standard Deviation	4.8	5.7	5.9
Minimum	15	10	21
Maximum	101	84	94

Table 9. *Crosstab of age statistics broken out by categories of a qualitative variable (gender).*

Check your knowledge by reviewing answers to the following questions:

1. What are[21] the recommended graphical tools for presenting the interaction between two qualitative variables?

2. What is[22] the recommended tabular tool for presenting the interaction between two qualitative variables?

3. What are[23] the recommended graphical tools for the interaction between two quantitative variables?

4. What are[24] the recommended numerical techniques for the interaction between two quantitative variables?

5. What are[25] the recommended graphical tools for presenting the interaction between a qualitative and a quantitative variable?

6. We didn't cover this, but can[26] you think of a way to present two quantitative and one qualitative variable?

7. What is[27] the recommended tabular tool for presenting the interaction between a qualitative and a quantitative variable?

[21] Grouped or stacked bar charts.
[22] Contingency tables.
[23] Scatter plots and line charts.
[24] Correlation, covariance and the coefficient of determination.
[25] Stacked histograms and side-by-side box plots.
[26] You can use a scatter plot that is color coded for the categories of the qualitative variable. This is called a scatter plot with groups. HOWTOs are available online.
[27] A contingency table (crosstab) showing statistics of the quantitative variable for different categories of the qualitative variable.

8. How can[28] the one interpret correlation?

9. How can[29] one interpret covariance?

Guidelines

In this section we start by considering interpretive issues and opportunities emerging from Anscombe's Quartet, the empirical rule and Simpsons Paradox. The section concludes by providing guidelines for the presentation of descriptive statistics.

Anscombe's Quartet

Recall our maxim - "first chart your data." The maxim is critical when we consider the following insight:

It may not be enough to know the mean or even the mean *and* variance!

To see why this is true, realize that data sets that are fundamentally different may have the same numerical descriptive statistics. Anscombe produced a set of a four (x, y) data sets ("Anscombe's quartet" - depicted in Figure 15), with virtually identical means, variances and correlation coefficients that illustrate the magnitude of the problem.

A recent paper (Matejka and Fitzmaurice, 2017), beautifully demonstrates the limitations of means, standard deviations and correlation coefficients in discerning differences between twelve data sets in the Datasaurus Dozen (see Figure 16). In addition, it shows that box plots are unable to discern the differences between different data sets. Violin plots may be suitable alternatives in this setting.

These examples highlight the limitations of only a single number such as "estimated" travel time in map apps. While a single number may suffice for many everyday travel needs, it fails to provide enough information needed to manage the risk associated with missing an important meeting or an expensive connection on public transportation.

Empirical Rule

The empirical rule is useful for interpreting standard deviation. Assuming the distribution is reasonably symmetric (approximately bell shaped or normally distributed), the empirical rule states that

- 68% of the observations lie with one standard deviation of the mean,
- 95% of the observations lie within two standard deviations of the mean, and
- 99.7% of the observations lie within three standard deviations of the mean.

See Figure 18 for a graphical depiction of the empirical rule.

[28] Pearson correlation is a statistic that measures the strength of the linear relationship between two quantitative variables. It measures the degree to which the two variables move in relation to each other but does not completely characterize the relationship between them. Pearson correlation is unitless and takes values in the interval $[-1, 1]$.

[29] Covariance is a statistic that measures the strength of the linear relationship between two quantitative

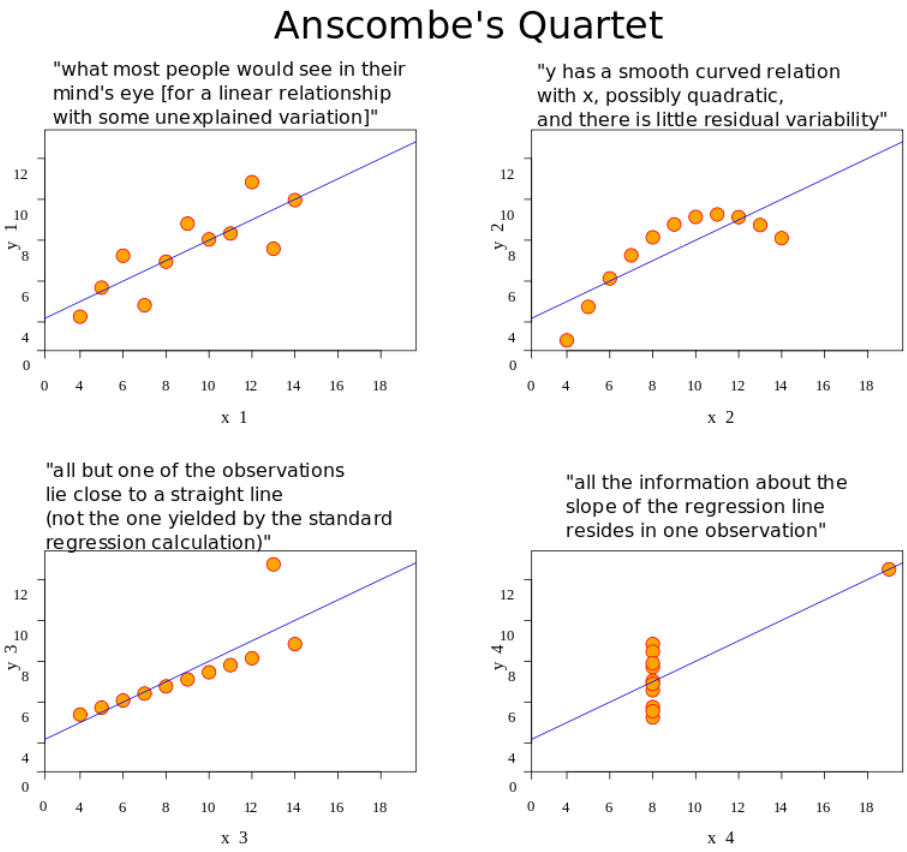

Figure 15. Anscombe's Quartet.

As an example of the use of the empirical rule, suppose that you learn that the mean and standard deviation of travel times of recent Lyft customers to your destination is, respectively, 60 minutes and 5 minutes. Assuming the distribution of those travel times is roughly symmetric, the empirical rule allows you to conclude that about 68% of the trips took between 55 and 65 minutes (60±5), 95% between 50 and 70 minutes (60±10), and almost all between 45 and 75 minutes (60±15). Leaving enough time for a 75 minute trip therefore almost guarantees that you will not be late.

Simpson's Paradox

Simpson's Paradox arises when different conclusions are reached when data is aggregated in various ways. The Wikipedia article on Simpson's paradox discusses a few well-known examples of the paradox, including an originally-supposed case of gender discrimination at UC Berkeley in the 1970s and a batting-average analysis of the MLB players Derek Jeter and David Justice in the 1990s. We are fortunate to have a much more recent

variables. It measures the degree to which the two variables co-vary in the sense of simultaneously differing from their mean values.

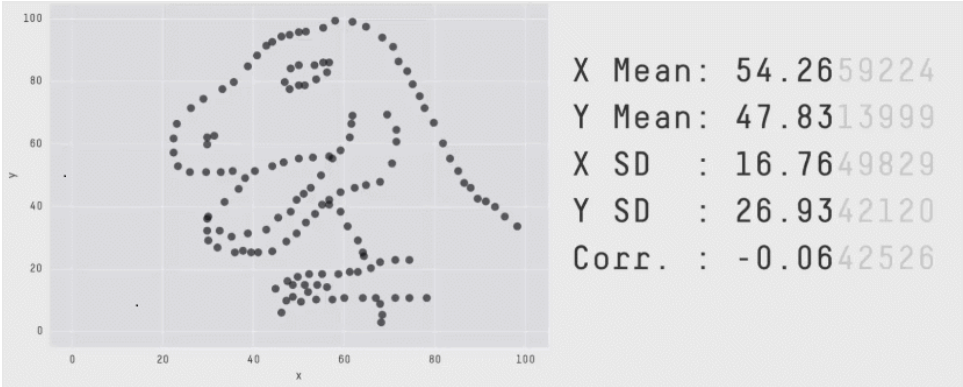

Figure 16. The Datasauarus Dozen - twelve data sets with the same means, standard deviations and correlation coefficients. See the animated version here.

Figure 17. Data sets showing that violin plots dominate box plots. See the animated version here.

example in form of the ubiquitous 2016 US Presidential Election which will be discussed in Case I!

Guidelines

In this section we start by considering general guidelines for compiling and presenting descriptive statistics. Recognize that these guidelines are very general/minimalistic and many applications may have situation-specific practices in place.

- Start by ensuring that the goals of the descriptive-statistics exercise are well articulated.
- The selection and collection of relevant data for measuring pertinent variables that can be accessed in a timely and cost-justifiable manner can proceed once the goals are well defined.
- The data has to be scraped, cleaned, munged, verified and prepared prior to use. This includes deletion of erroneous records, addressing the issue of missing fields, and calculation of derivative variables.
- Then **chart your data!** Start with single-variable graphs and charts and proceed to multiple-variable charts.

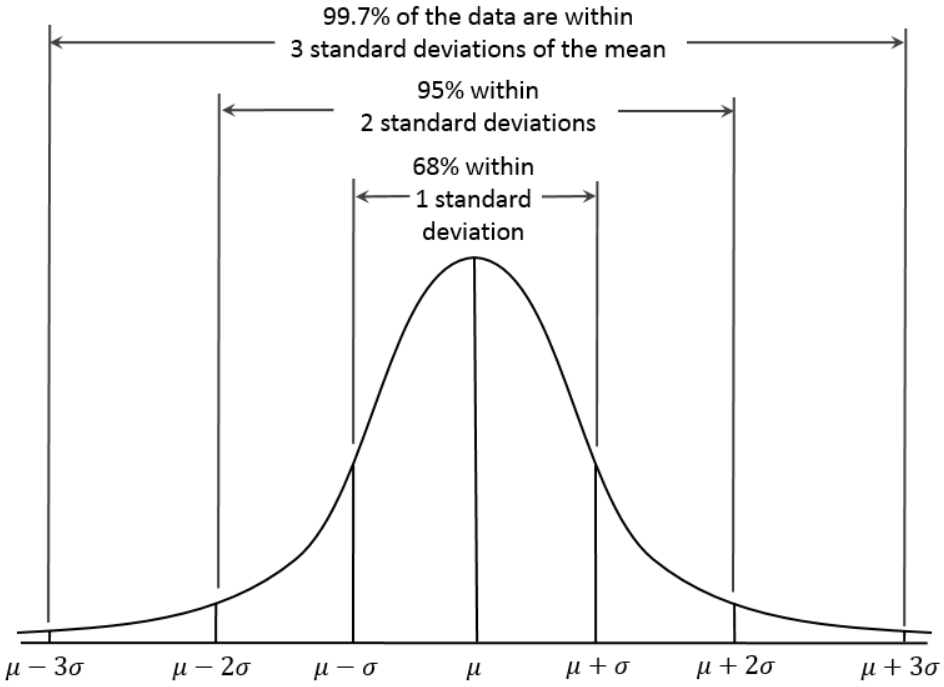

Figure 18. Empirical Rule. By Dan Kernler (Own work) [CC BY-SA 4.0 (http://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons.

- Then prepare single-variable numerical statistics, followed by multiple-variable numerical statistics. With regard to single-variable measures of location, prefer the mean for non-skewed distributions and the median for skewed distributions. With regard to single-variable measures of dispersion, prefer the standard deviation as a general guideline. For numerical measures of linear relationship, prefer the Pearson coefficient of correlation and the coefficient of determination.

- In selecting charts, tables and numerical measures, **keep your audience firmly in mind**. For example, the use of robust statistics such as the trimmed mean that are unfamiliar to the general public may well be appropriate and/or preferable for internal corporate purposes. Keep charts simple. Avoid pie charts,3D charts, cylinders, and cones. Surface charts, doughnut charts and radar charts may be unfamiliar to your audience. Avoid misleading graphs and charts. See http://www.statisticshowto.com/misleading-graphs/ for some poor examples.

- Pay attention to the potential of disabilities in your target audience. The use of blue and orange palettes may mitigate the effects of color blindness, but consult recent guidelines such as those at wearecolorblind.com. Recognize that this problem may be relatively common among audience members: red–green color blindness affects up to 8% of males and 0.5% of females of Northern European descent[30]. To test yourself, see this.

[30] https://en.wikipedia.org/wiki/Color_blindness

- As advertised previously, the techniques and methods of descriptive statistics appropriate in a given situation will depend upon the level of data involved. Higher-level data measures can be treated as lower level, but not vice versa. This means, for example, that we have the option of treating ratio-level data interval data as ordinal- or nominal-level, and ordinal-level data as nominal-level. Conversely, the use of ratio-level methods for nominal-level data would be an example of a serious mistake.

Check your knowledge by reviewing answers to the following questions:

1. What are[31] the implications of Anscombe's quartet?
2. How can[32] the empirical rule be applied to interpret standard deviation?
3. What are[33] the implications of Simpson's paradox.
4. What are the guidelines for presentation and interpretation of descriptive statistics?

Cases

We will now consider some examples in two cases. I suggest using Excel to follow the analysis of the case data sets in service of Learning Objective 4.

Case I: 2016 US Presidential Election

Let's return to Simpson's Paradox. Consider the relative frequency information in Table 10 built from election data obtained here. If we factor the winner by the number of regions (Northeast, Midwest, South and West) then the result is a tie. Note that the regions and candidates variables are both nominal-level variables. If we factor the winner by the number of states won then Trump is victorious. The states variable is nominal-level. If we factor the winner by the total number of votes received (i.e., the so-called popular vote) then Clinton wins. Of course the electoral vote identified Trump as the actual winner. Johnson: remember him?

The point of this exercise is to emphasize that care must be taken in the selection of aggregation processes. Simpson's paradox is a special case of the so-called omitted-variable bias problem. Clearly, any model that fails to capture the electoral vote mechanism for the election data has the potential for producing an erroneous conclusion.

Case II: Financial Time Series Data

For our final case we consider market data available to investors trying to figure out where the markets were headed on Martin Luther King Jr Day, 2017, in the run up to Donald Trump's inauguration on January 29, 2017. Specifically, adjusted closing prices for every day

[31] Providing the mean may not be enough. Providing the mean and standard deviation may not be enough. Anscombe's example goes further than this, of course. It (and the Datasaurus Dozen) suggests the need to explore a graphical representation of the data.

[32] 68% of the observations are within one standard deviation of the mean, 95% within two standard deviations of the mean and 99% within three standard deviations of the mean.

[33] One has to be careful when one aggregates data. The conclusion that is drawn may depend upon the aggregation.

	Candidates			
Basis	Trump	Clinton	Johnson	Winner
No. of Regions	50%	50%	0%	Tie
No. of States & Washington DC	60.8%	39.2%	0%	Trump
Total no. of Votes	45.96%	48.05%	3.28%	Clinton
No. of Electoral Votes	56.88%	43.12%	0%	Trump

Table 10. *2016 US Presidential Election Results.*

from November 9, 2016 until January 13, 2017 data for four exchange-traded funds (ETFs) - tickers NYSEARCA:SPY, NYSEARCA:GLD, NYSEARCA:BND, and NYSEARCA:VDE - were obtained from `finance.yahoo.com`. The four ETFs are used to model prices in the stock (SPY), gold (GLD), bond (BND) and oil (VDE) sectors. This data was used to compute the fractional price change $((p_t - p_{t-1})/p_{t-1} = p_t/p_t - 1$ where p_t denotes the price on day t) for each day for each ETF. For example, if the price was \$100 yesterday and \$101 today then the fractional price change is $101/100 - 1 = .01$. The processed data (and Excel analysis) is available here.

A line chart of the time evolution of the change in each ETF appears in Figure 20. Box plots of the four ETF price changes appears in Figure 19. Observe from the box plots that pricing in the bond market appears to have been the least variable of the four markets. A sample scatter plot appears in Figure 21 illustrating the change in GLD versus the change in BND. We eschew the opportunity to present all six pairs of scatter plots and the single-variable statistics in the interest of space and focus only on presenting the correlation half-matrix - a measure of the linear relationship between the different ETF variables.

$$\begin{bmatrix} \rho_{GLD,SPY} & - & - \\ \rho_{BND,SPY} & \rho_{BND,GLD} & - \\ \rho_{VDE,SPY} & \rho_{VDE,GLD} & \rho_{VDE,BND} \end{bmatrix} = \begin{bmatrix} 0.074 & - & - \\ 0.032 & 0.702 & - \\ 0.372 & -0.038 & -0.144 \end{bmatrix}$$

Recommended Reading

Diez, D., Barr, C., Cetinkaya-Rundel, M. (2015) *OpenIntro Statistics* downloaded from `https://www.openintro.org/stat/textbook.php`

Campbell, D., & Frei, F. (2010). Cost structure, customer profitability, and retention implications of self-service distribution channels: Evidence from customer behavior in an online banking channel. *Management Science*, 56(1), 4-24.

Matejka, J., & Fitzmaurice, G. (2017, May). Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 1290-1294). ACM. Download available at `https://www.autodeskresearch.com/publications/samestats`.

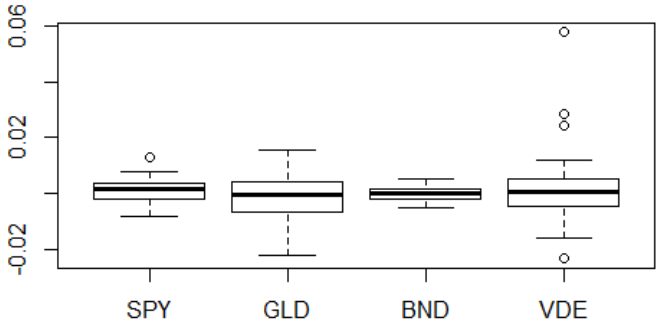

Figure 19. Box plot of daily fractional price change in the four ETFs - SPY, GLD, BND and VDE.

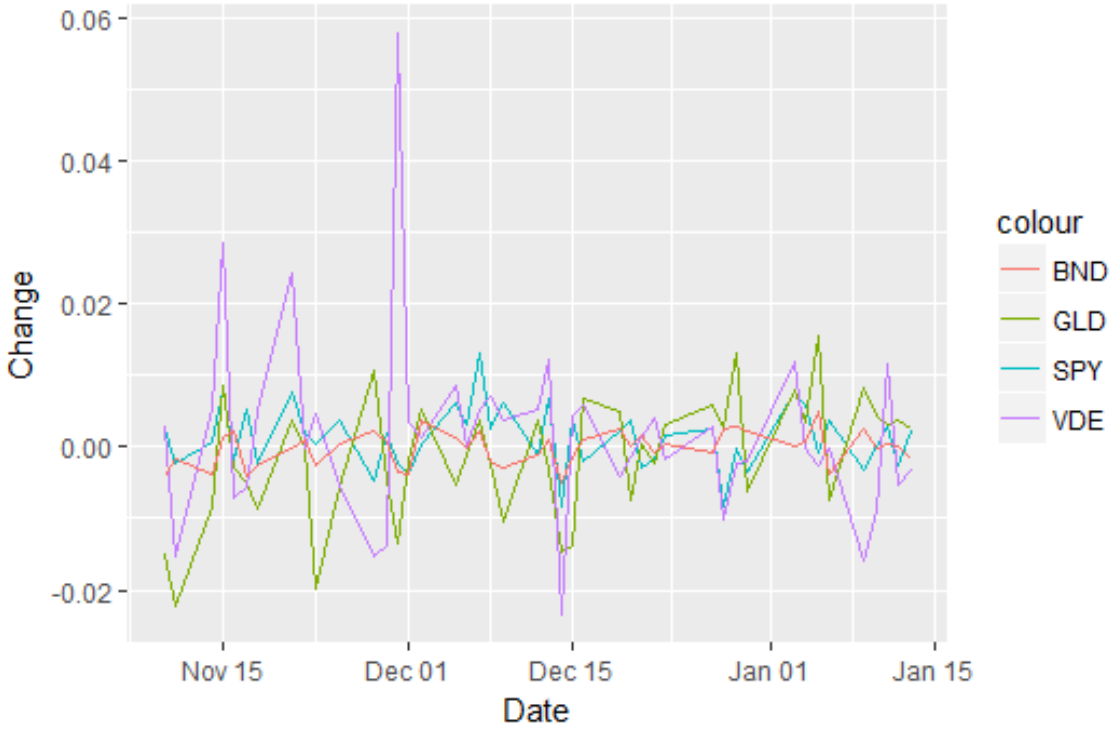

Figure 20. Line chart of daily fractional price change in the SPY, GLD, BND and VDE ETFs.

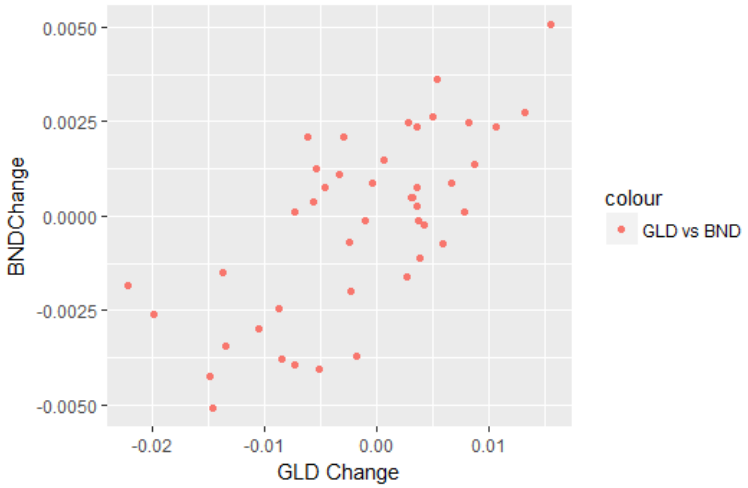

Figure 21. Scatter plot of daily fractional price change in GLD vs daily change in BND.

Wilkinson, L. (2005). *The Grammar of Graphics (Statistics and Computing)*. Springer-Verlag.