CSU **Hayward**

Statistics *Department*

Introduction to Minitab Demonstrations

What You Need to Get Started

This introduction is mainly for students who are working on their own and who have no previous experience with Minitab or with the network in the CSUH School of Science Computer Lab. If the instructor of your class has given you an orientation session or if you have used Minitab in the Science computer lab before, you may be able to skip parts of this introduction. In that case, just remember that useful information is here in case you need help.

Your location, computer equipment, and personal preferences will determine how you use these demonstrations. Even though you may get some benefit just from browsing through the notes for each part, they are **meant to be used interactively on a computer that is running Minitab statistical software.** You should try each procedure on the computer as you read about it. That way you will begin get the feel of hands-on, interactive data exploration. In order to use these notes as intended, you will need:

- **Parts 1-6 of these notes from the CSU Hayward Statistics website.**
  - You may wish to make a printed copy of these notes using your web browser so that you can write comments on the paper pages as you go along. (The notes are protected by copyright but making printouts for any non-commercial educational purpose is hereby authorized.)
  - If you have a system with enough speed and memory to run programs in two windows at once, you may wish to read the notes in one window on your screen and to work with Minitab in another. (Maximize the web window to read about a procedure; then maximize the Minitab window to try the procedure for yourself, then go back to the web window again, and so on.)
- **Access to the Data.** The datasets have been prepared in "Minitab worksheet" format so that they can be loaded instantly for use with Minitab.
  - At present, the worksheets are available only on the Cal State Hayward campus from the School of Science network. They are available on server drive `I:` (public files) in the path `I:\ COURSWRK\ STAT\ BTRUMBO\ MINDAT.` You need the Minitab worksheet `MINDEMO.MTW` (DOS or Windows, Minitab Release 7 or higher), or `MINDEMO.MTP` ("portable" format, for all versions of Minitab with sufficient capacity).
  - For the benefit of those using machines with limited memory, student versions of Minitab, or Minitab releases limited to 50 columns of data, the

    data have been broken out into separate worksheets for each part: `MINDEMO1.MTW, MINDEMO2.MTW,` and so on. Column numbers remain the same as in the combined worksheet (except for Part 6). Portable versions with extensions `.MTP` are also available.

    - An ASCII text printout of the data is provided. With some editing, each dataset could be extracted from this file for use in almost any statistical software package.
    - Eventually, we may be able to provide the "portable" version of the worksheet over the internet. If so, these instructions will be changed accordingly.

- **Minitab software.**
    - *Minitab available on campus.* Some version of Minitab is available in almost every computer lab on campus. The School of Science Computer Lab uses Release 11 for Windows. You are free to use Minitab in campus computer labs wherever it has been installed, but you may not copy Minitab software from university computers for use off campus. (Lab administrators have made this very difficult to do, the installation you would copy probably won't work on a different computer, and it is seriously illegal.)
    - *Purchasing Minitab software.* The Pioneer Bookstore on the Cal State Hayward campus has very good prices on Minitab software due to a special arrangement with Minitab, Inc. These are the *full versions* of Minitab used in businesses and by professional statisticians -- often at prices below those you will see elsewhere for stripped down student versions. We do not recommend the student versions.
    - *DOS Releases.* These notes were originally prepared using Minitab Release 7 for DOS, but at this elementary level almost everything should work -- even in older versions of Minitab. Read the Instructions for DOS users.
    - *Windows Releases.* Windows menu selections are shown for each procedure you are asked to do. Windows releases have the capability to make high-resolution graphics versions of many of the displays we show in these notes. We have purposely shown "character graphics" for the benefit of users who lack the hardware, software, bandwidth, or patience required to download graphics files over the web. Character graphics are pictures composed entirely of text symbols. Read the Instructions for Windows users.
    - *Macintosh releases.* Macintosh menu selections may differ somewhat from those shown here for Windows. You will need to use the portable version `MINDEMO.MTP` of the worksheet). We have not tested these notes on Macintosh releases of Minitab; you are pretty much on your own.

Part 1 -- IQ Scores

Setup

In this demonstration you will use the Minitab worksheet `MINDEMO.MTW.` You need to retrieve that worksheet from disk so that it is ready for use within Minitab. This worksheet contains the data for Parts 1-6 of these notes. (Where memory limitations are a concern, use `MINDEMO1.MTW` here and retrieve other numbered worksheets for other parts.) In this demonstration you will use the following columns, the contents of which will be explained as we go along:

- c2 Origl IQ
- c3 Final IQ

The Data

The data for this demonstration are IQ scores of 250 high school students in the San Francisco Bay Area, collected for a master's thesis in Educational Psychology at CSU Hayward.

Exploration of the Data

Dotplots. The dotplot is one of the simplest graphical devices. Each observation is represented by a dot appropriately placed along a horizontal axis. If several observations have the same (or nearly the same) value, they are stacked vertically.

In Minitab you might make a dotplot in either of two ways:

First, you may type the command `DOTPlot,` followed by the column identifier (here c2 or 'Origl IQ'). Minitab does not distinguish between capital and small letters in commands. We capitalize the first four letters here to signify that they are the only ones required. (If a command name has more than four letters you need to type only the first four letters, but you may type the entire command name if you like.)

Alternatively, in Windows versions of Minitab, you may select the menu path **`GRAPH > Character > Dotplot,`** and then c2 (Origl IQ). In these notes the menu path for Windows is shown at the beginning of each display, followed by the corresponding command.

```
GRAPH > Character > Dotplot
MTB > dotp 'Orig IQ'
```

```
                           .   :
                          ::   :  .
                          :: : : .:
                          :: : : ::
                         :::::.::::
                         :::::::::::
                      :  :::::::::::
                      :.::::::::::::: :
                      ::::::::::::::: :
                   :: :::::::::::::::::
               ..:.::::::::::::::::::::.:.    .              .
           -----+---------+---------+---------+---------+---------+-
Origl IQ
               60        90       120       150       180       210
```

From this dotplot of the data we see that most of the IQ scores are between 70 and 130, with a few outside this interval on both sides. However, the striking thing is the extreme IQ score of almost 200. From what we know about IQ scores this is probably an error.

Boxplots. The boxplot of a dataset is based on the "five-number summary" of the observations. From smallest to largest these five numbers are:

1. The minimum
2. The lower quartile (lower end of box)
3. The median (symbol within box)
4. The upper quartile (upper end of box), and
5. The maximum.

Notice that the "middle half" of the observations fall within the box of the boxplot.

An outlier is a value that falls relatively far away from the rest of the values in a dataset. A Minitab boxplot signals probable outliers with the symbol `O` (and possible ones with `*`).

```
GRAPH > [Character >] Boxplot
MTB > boxp 'Orig IQ'

                          --------
               -----------I  +   I--------    *              O
                          --------
          ------+---------+---------+---------+---------+---------+-
Origl IQ
               60        90       120       150       180       210
```

Note: The menu path **GRAPH > Boxplot, without X-variable** gives a pixel-graphic boxplot which runs vertically instead of horizontally, but gives the same information as the one shown here.

The boxplot explicitly highlights the extreme value, and labels it as a probable outlier. The symbol `*` indicates a "possible" outlier -- here not an error, just a very bright student.

Numerical Descriptive Statistics. Minitab makes it easy to compute a number of numerical descriptive statistics for a dataset.

```
STAT > Basic > Descriptive
MTB > desc 'Origl IQ'
```

```
                 N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
 Origl IQ      250   100.32   100.00   100.21    16.52     1.04

               MIN      MAX       Q1       Q3
 Origl IQ    58.00   196.00    90.00   112.00
```

The crucial information here is the maximum value MAX = 196. This is the exact numerical value of the outlier seen in the dotplot and the boxplot above.

> Notes on other descriptive statistics shown above: Check your textbook for the definitions.
>
> - The sample size = 250. (Minitab uses *N* here, but most texts use *n* for sample size and *N* for population size.)
> - The sample MEAN = 100.32. (Most texts use x-bar or y-bar for the sample mean.)
> - The sample MEDIAN = 100.00
> - The sample *standard deviation,* STDEV = 16.52
> - TRMEAN stands for the *trimmed mean* of the sample, computed by ignoring the highest 5% and lowest 5% of the data and averaging the middle 90%; this quantity is not as sensitive to erratic extreme values as is the mean.
> - Q_1 and Q_3 are the lower and upper *quartiles of the sample.*
> - SEMEAN is the (estimated) *standard error of the mean,* equal to the sample standard deviation divided by the square root of the sample size; this quantity is used in statistical inference.

In the actual situation upon which these data are based, the researcher rechecked the original list of IQ scores and found that the value 196 resulted from a data input error; the correct value is 96. The data in c3 (Final IQ) are identical to those in c2 except that this error has been corrected. Now we repeat our work, using the corrected data.

GRAPH > Character > Dotplot

```
MTB > dotp c3

                               .::   . :  .
                               ::: :.: :  :.:
                             : ::::::: : ::::
                          .  : :::::::::::::::
                         :: ::::::::::::::::::.. :
                     ..  ::::::::::::::::::::::: :.
               . .:..::.:::::::::::::::::::::::::::  :.      .
           -----+---------+---------+---------+---------+---------+-
Final IQ
               60        80       100       120       140       160
```

Here is a comparison of the numerical descriptive statistics for the incorrect and corrected IQ data. (Note that descriptive statistics can be computed for more than one column at a time.)

STAT > Basic > Descriptive, select both columns

```
MTB > desc 'Origl IQ' 'Final IQ'

                 N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
 Origl IQ      250   100.32   100.00   100.21    16.52     1.04
 Final IQ      250   99.920  100.000  100.076   15.367    0.972

               MIN      MAX       Q1       Q3
 Origl IQ    58.00   196.00    90.00   112.00
 Final IQ   58.000  150.000   90.000  112.000
```

The incorrect observation changed the mean by .4 of an IQ point (giving 100.3 compared with a correct mean of 99.9), the trimmed mean by about .1 of an IQ point, and the median not at all.

Comments

Unlike "textbook" examples, real data almost always contain some errors. In beginning to study a dataset it is well to use a number of graphical and numerical devices to screen the data for unreasonable and inconsistent values.

Using a computer with statistical software such as Minitab, we find it easy to take such a critical look at a dataset before we try to draw conclusions from it -- even if the sample size is fairly large, as in the present case. Consider for a moment how much work would be required to duplicate the work shown in this demonstration if we had to do it using pencil, graph paper, and a hand calculator.

Part 2 -- Sunflower Seedlings

Setup and Data

In this demonstration you will use columns c12, c13, and c14 of the Minitab worksheet `MINDEMO.MTW` (or `MINDEMO2.MTW`).

A standard botany lab experiment at Cal State Hayward is to follow the growth of sunflower seedlings grown under various conditions. As part of this experiment, 30 sunflower seedlings were grown in soil that contains no nitrogen nutrients. In this demonstration we will look at data on the heights of these plants measured at the end of the second week (c12), third week (c13), and fourth week (c14).

Exploration of the Data

Dotplots. In order to follow the progress of these seedlings grown in nitrogen-deprived soil, we compare dotplots of the heights for each of the three weeks.

For ease of interpretation it is important that these three dotplots be drawn on the same scale. If we use the `DOTPlot` command, we can accomplish this by using the subcommand `SAME`. (Notice that the command line ends with a semicolon, and that the subcommand is on a separate line, which ends with a period. Also notice that this time we use a range of column numbers instead of column names to identify the columns; either method works.) With menus, we select the three variables and the option to put them on the same scale.

GRAPH > Character > Dotplot, same scale.

```
MTB > dotp c12-c14;
SUBC> same.

                       .
                       :
                       :
                       :
             .    . :  : .  .
             : .  : :  : :  : :
           -------+---------+---------+---------+---------+---------+
No N Wk2
                            :
                            :      .
                            : .  . :  . .
                  . :    .  : :  : :  : :                      .
           -------+---------+---------+---------+---------+---------+
No N Wk3
                                             .
                                   .  : :    :    .
                       .    . .  : :  : :  : :  . :    :    .
           -------+---------+---------+---------+---------+---------+
No N Wk4
                 8.00     12.00     16.00     20.00     24.00     28.00
```

We may not know how tall the nitrogen-deprived seedlings could be expected to grow in three weeks. However, the value 26 in Week 3 not only looks suspicious in its own right, it is inconsistent with the data for Week 4. (Did the largest seedling shrink in size during the fourth week?)

We print out the data in these three columns, abridged here to save space. The row with the questionable observation is indicated with an arrow (edited in by hand).

MANIP > Display Data

```
MTB > print c12-c14

  ROW  No N Wk2  No N Wk3  No N Wk4

    1        11        15        18
    2        10        14        17
         ...       ...       ...
    8         6         8        12
    9        10        12        15
   10         8        13        17
   11        13        26        25     <---- 12 10 13 16 13 12 17 23
... ... ... 27 9 12 14 28 10 13 17 29 7 11 15 30 10 17 23
```

When checked, the original lab sheets showed that the 11th value in c13 should have been 20 instead of 26. There are two ways in which to correct this error:

1. Using the command `LET c13(11) = 20 .`
2. Going into the worksheet and changing the value directly.

Note that by either method only the data in the active version of the worksheet are changed. For a permanent record of the edited dataset, you must record it on disk. You could either use the command `SAVE` (followed by the path and name of the new file -- enclosed in single quotes) or select **`FILE > Save As`** from the Windows menu.

The data are now "cleaned up" and ready for statistical analysis. In this case it is unlikely that the error we corrected would have prevented us from making a useful analysis of the data, but it is not difficult to imagine cases in which one or more transcription errors could change the interpretation of a dataset.

We repeat the dotplots for the repaired data, and then look at numerical descriptive statistics.

`PLOT > Character > Dotplot, same scale`

```
MTB > dotp c12-c14;
SUBC> same.

                       .
                       :
                       :
                       :
           .     .  :  : .  .
           :  .  :  :  : :  :  :
           ---+---------+---------+---------+---------+---------+---No
N Wk2

                            :
                            :        .
                            :  .  .  :  .  .
                 .  :    .  :  :  :  :  :  :       .
           ---+---------+---------+---------+---------+---------+---No
N Wk3

                                                .
                                     .  :  :    :     .
                       .    .  .  :  :  :  : :  :  .  :     :    .
           ---+---------+---------+---------+---------+---------+---No
N Wk4
             7.00     10.50     14.00     17.50     21.00     24.50
```

`STAT > Basic > Descriptive`

```
MTB > desc c12-c14


                 N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
 No N Wk2       30    9.667   10.000    9.692    1.882    0.344
 No N Wk3       30   13.600   13.500   13.615    2.660    0.486
 No N Wk4       30   17.500   17.000   17.500    3.381    0.617
```

```
               MIN      MAX       Q1       Q3
 No N Wk2    6.000   13.000    8.750   11.000
 No N Wk3    8.000   20.000   12.000   15.250
 No N Wk4   10.000   25.000   15.000   19.250
```

Notice that, by almost any numerical criterion, the seedlings continued to grow in weeks 2 through 4 -- even though the soil has no nitrogen. Later, when the nutrients in the seeds themselves had been exhausted, these seedlings did very poorly compared to ones grown in properly fertilized soil.

Comments

If we have access to the original data sheets for an experiment, it is sometimes possible to get rid of outliers by making corrections. But in some instances it will not be possible to explain an unusual observation. (What if the person who made the original measurement rather than the person who entered the data into the computer had made the error?) In those cases, hard choices must be made about whether to disregard questionable observations.

In general, great caution must me used in throwing out data that do not fit expected patterns. There is a story (perhaps true, perhaps not) that the "ozone hole" over the South Pole might have been discovered several years earlier than it was if a computer had not been programmed in such a way that it ignored the "obviously faulty" low ozone readings obtained.

Part 3 -- Quality Management

Setup and Data

This part continues to use the Minitab worksheet `MINDEMO.MTW` -- specifically columns c22 and c23. (Alternatively, you may retrieve `MINDEMO3.MTW.`)

The particular data shown here were collected in the early 1960s as part of the "quality control" program at a factory in Illinois where electromechanical devices were manufactured. However, the basic story is one that has been repeated many times in many settings -- and one that has been used by W. Edwards Deming to illustrate principles of quality management. (At the request of the company involved, the data were rescaled slightly before they were taken off-site.)

In order to function properly in the finished product of which they are a part, metal rods must be at least 1.000 cm in diameter. Of course, they must also not be *too* much larger than 1 cm, but here we focus on the crucial minimum diameter specification. A lot of 400 such rods was inspected with the results recorded in column c22 named 'Inspect'.

Exploration of the Data

Histograms. We begin by making a histogram of these data. Histograms made by many statistical packages and those usually published in printed articles and reports have the measurement scale along the horizontal axis with rectangular bars extending vertically. Before the histogram is drawn, the data are sorted into intervals, each of which forms the base of one of the bars.

> Versions of Minitab that run under Windows make histograms with a horizontal measurement scale and vertical bars, as described above: select **`GRAPH > Histogram`**. In these notes for the web, we prefer to use Minitab's character graphics because they do not require downloading graphics files.
>
> A character graphic histogram is plotted "sideways," with the measurement scale running vertically, and with rows of asterisks (*) instead of rectangular bars. Note that in our example each asterisk represents *up to* 2 observations; otherwise the histogram bars would run off the page. (In Minitab Release 7 and some later versions, one can also use the `GHIStogram` command to make a horizontal histogram on the graphics page.)

```
GRAPH > [Character >] Histogram
MTB > hist 'Inspect'

 Histogram of Inspect   N = 400
 Each * represents 2 obs.

 Midpoint   Count
    0.996       3  **
    0.997       8  ****
    0.998       0
    0.999       0
    1.000      93  ***********************************************
    1.001      63  ********************************
    1.002      72  ************************************
    1.003      68  **********************************
    1.004      46  ***********************
    1.005      27  **************
    1.006      13  *******
    1.007       5  ***
    1.008       2  *
```

The peak in the number of observations just at 1.000 cm and the absence of any observations at all just below at .999 and .998 is suspicious. It appears that the inspectors have fudged the results in order to pass rods that are just a bit too small. When questioned they readily admitted that they had not understood the importance of the 1.000 cm lower limit and that they had indeed recorded slightly undersized rods as being 1.000 cm in diameter in a misguided attempt to avoid throwing them out. The rods were subsequently reinspected (more honestly) with the results recorded in c23 named 'Reinsp':

```
GRAPH > [Character >] Histogram
MTB > hist 'Reinsp'

 Histogram of Reinsp   N = 400
 Each * represents 2 obs.

 Midpoint   Count
    0.996       3  **
    0.997       8  ****
    0.998      22  ***********
    0.999      30  ***************
    1.000      41  *********************
    1.001      63  ********************************
    1.002      72  ************************************
    1.003      68  **********************************
    1.004      46  ***********************
    1.005      27  **************
    1.006      13  *******
    1.007       5  ***
    1.008       2  *
```

Other Descriptive Methods. Notice that the numerical descriptive statistics are very much the same for the dishonest and honest inspection records; it is unlikely that our suspicions would have been aroused just by looking at the numerical summary of the original data in c22 ('Inspect').

```
STAT > Basic > Descriptive
MTB > desc 'Inspect' 'Reinsp'

                 N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
 Inspect       400   1.0021   1.0020   1.0020   0.0020   0.0001
 Reinsp        400   1.0019   1.0020   1.0019   0.0023   0.0001

               MIN      MAX       Q1       Q3
 Inspect    0.9960   1.0080   1.0000   1.0030
 Reinsp     0.9960   1.0080   1.0000   1.0030
```

Similarly, the boxplot of the dishonest data shows nothing that would have caused us to suspect their validity. Boxplots are good at highlighting extreme values but not at showing peculiarities in the central part of the sample distribution. Try making a boxplot of the honest data on your own; it will not not be much different from the boxplot of the dishonest data shown below. (It is a quirk of Minitab that one cannot draw two *boxplots* to the same scale in a single command as one can do for *dotplots.)*

GRAPH > [Character >] Boxplot

```
MTB > boxp 'Inspect'

                               -------------
               ----------------I       +   I----------------   *
                               -------------
           +---------+---------+---------+---------+---------+------
Inspect
      0.9950    0.9975    1.0000    1.0025    1.0050    1.0075
```

On the other hand, a dotplot of the original data gives much the same impression as the histogram, showing a peak adjacent to a gap. Actually, the dotplot is often a better bet to detect peculiarities in the "shape" of a sample. An unfortunate grouping of the data for the histogram might have put the peak and the gap into the same bar of the histogram, thus obscuring both.

GRAPH > Character > Dotplot

```
MTB > dotp 'Inspect'

 Each dot represents 5 points

                              .
                              :
                              :       .
                              :   .   :   :
                              :   :   :   :
                              :   :   :   :   :
                              :   :   :   :   :
                              :   :   :   :   :   :
                              :   :   :   :   :   :   .
              .   :           :   :   :   :   :   :   :   .   .
          +---------+---------+---------+---------+---------+-------
Inspect
       0.9950    0.9975    1.0000    1.0025    1.0050    1.0075
```

Comments

The moral of this example -- and the two in preceding parts -- is that it is wise to look at each data set using a *variety* of graphical and numerical methods. No one method can be guaranteed to show up the anomalies that may be present. Computer analysis has a clear advantage in such a program of data exploration. Using a statistical computer package such as Minitab, one can quickly and easily use a variety of descriptive techniques to explore a dataset. Such a thorough analysis by hand would be quite tedious, and would probably seldom be done in practice.

Part 4 -- Sodium in Hot Dogs

Setup and Data

This part uses data from columns c32 and c33 of `MINDEMO.MTW.` (You may also use `MINDEMO4.MTW,` which contains only these two columns.)

In 1986 researchers at Consumers Union analyzed samples of 54 brands of hot dogs for fat and sodium content, and reported the results along with other information in the June 1986 issue of *Consumer Reports*. Sodium in hot dogs comes from salt and other preservatives. Guidelines vary, but there is general agreement that the typical American diet is much too high in sodium. Here we consider the sodium content (in mg/oz) for two general types of hot dogs:

- Red Meat: 36 brands containing either pork or beef, and
- Poultry: 17 brands for which the animal content consists entirely of chicken or turkey (or a combination).

MANIP > Display Data

```
MTB > print c32 c33

  ROW  RedMeat  Poultry

    1      248      269
    2      239      234
    3      213      248
    4      201      239
    5      241      242
    6      294      271
    7      216      224
    8      215      223
    9      240      264

   10      234      257
   11      193      213
   12      200      257
   13      241      298
   14      251      291
   15      323      294
   16      275      261
   17      211      273
   18      199

   19      199
   20      169
   21      229
   22      253
   23      237
   24      273
   25      248
   26      225
   27      242

   28      229
   29      254
   30      197
   31      203
   32      233
   33      256
   34      253
   35      268
   36      212
```

We derived the data given in the worksheet from information provided in *Consumer Reports* (on milligrams of sodium per hot dog and on weights of hot dogs in ounces). We eliminated one brand made of veal and having an exceptionally low sodium content.

Structure of the Data

Notice that the numbers of observations in the two columns are **not** equal (in technical language an *unbalanced* experimental design). Furthermore, in contrast to data we will see in Part 5, there is no connection between two observations that happen to be recorded in the same row of the worksheet. The two columns of data are independent of one another.

Exploration and Analysis of the Data

Descriptive Techniques. From the above listing of the data, it is difficult to tell whether sodium content is generally higher for one kind of hot dog than for the other. We begin by looking at dotplots for the two columns of data, drawn on the same scale.

```
GRAPH > Character > Dotplot, same scale
MTB > dotp 'RedMeat' 'Poultry';
SUBC> same.

                      .             :   .
            .       . ::. .::  .: :.:. :::   . ..     .         .
           -----+---------+---------+---------+---------+---------+-
RedMeat

                           .  ..  . .. .  :.. :.     ...
           -----+---------+---------+---------+---------+---------+-
Poultry
              180       210       240       270       300       330
```

While the samples with both the highest and the lowest concentrations of sodium are found among the meat hot dogs, there seems to be a clear tendency for poultry hot dogs to have higher concentrations of sodium than meat hot dogs do.

The data may be summarized numerically as follows:

`STAT > Basic > Descriptive`

```
MTB > desc 'RedMeat' 'Poultry'

                 N      MEAN    MEDIAN    TRMEAN     STDEV    SEMEAN
 RedMeat        36    233.72    235.50    232.34     30.98      5.16
 Poultry        17    256.35    257.00    256.47     25.25      6.12

               MIN       MAX        Q1        Q3
 RedMeat    169.00    323.00    211.25    252.50
 Poultry    213.00    298.00    236.50    272.00
```

The poultry hot dogs average about 256 mg of Sodium per ounce whereas the meat hot dogs average only about 234. Compare these means with what you see in the dotplots. (The sample mean can be viewed as the point at which the dotplot would balance if all dots have the same weight.)

Statistical Inference

The question of interest here is whether the difference between the sample means we noticed in the dotplots (and verified by exact computation) indicates a real difference between the two types of hot dogs or whether it might just have resulted from sampling variation. If we were to take another sample, would we expect to see a higher mean for poultry hot dogs again?

One inferential procedure commonly used to decide such questions is called a "two-sample t-test." It gives the probability that such a large difference in sample means would be due to chance alone. This probability is the P-value given in the printout below.

`STAT > Basic > 2-Sample t, different columns (otherwise retain defaults)`

```
MTB > twos 'RedMeat' 'Poultry'

 TWOSAMPLE T FOR RedMeat VS Poultry
          N      MEAN     STDEV   SE MEAN
 RedMeat  36     233.7      31.0       5.2
 Poultry  17     256.4      25.2       6.1

 95 PCT CI FOR MU RedMeat - MU Poultry: (-38.9, -6.4)

 TTEST MU RedMeat = MU Poultry (VS NE): T= -2.83  P=0.0075  DF=  38
```

The P-value of 0.0075 indicates that there are fewer than 8 chances in 1000 that a difference in sample means of the size we found here would occur by chance. We are led to conclude that poultry hot dogs as a group tend to have higher concentrations of sodium.

Thus, the formal statistical procedure confirms what we see by eye from a comparison of the two dotplots. This is as it should be. *A mathematical result that contradicts what one sees intuitively in a properly made graphic display should be viewed with great skepticism.*

Comments

Because of the very small P-value, and because the data do not show any obvious flaws, we conclude that among hot dogs available in 1986, those made of poultry tended to have higher sodium concentrations than those made of red meat. Our statistical analyses have shown this difference to be real.

It is another matter whether this difference is of practical importance. This is something that statistical procedures cannot judge. Notice that there is enough dispersion in both groups that a customer who wants to eat poultry hot dogs can find a brand with lower sodium content than for most brands with red meat. On the other hand the customer could well select one of several brands of meat dogs with a higher sodium content than most poultry ones -- in fact the specimen showing the very highest sodium content in our dataset was a meat hot dog. However, the lowest levels of sodium in hot dogs are still so high that a single hot dog may contain more sodium that a person should consume in a day.

Perhaps even more important, someone who takes healthful eating as a really serious matter will probably not be shopping for hot dogs in the first place -- for reasons in addition to sodium content.

> As is often the case with real datasets, there are some questions here as to how the data were selected and how seriously we should take the results of the formal statistical analysis.
>
> For students who are beyond the first few weeks of an introductory statistics course, we discuss briefly some technical assumptions that one must make in performing the t-test just shown. We assumed that:
>
> - **The data in each group (meat and poultry) are a random sample of all individuals in the group.** It is unlikely that the people from Consumers Union made a careful random sample either of brands or of individual hot dogs. It is also hard to imagine that they had any reason deliberately to seek out poultry hot dogs with especially high sodium content to include in the study.
> - **The data are normally distributed.** This assumption is not as important here as it is in some instances because the sample sizes are moderately large and there is no evidence of serious skewness or outliers. Procedures which we shall not describe here can be used to test whether data are normal; they revealed no difficulty.

- **The data in the two groups are independent.** There is no reason to doubt this assumption from what we know of the data.

Note: The t-test we used does not assume that the variances of the two groups are equal. A pooled test, which does require this assumption, gives results similar to the ones obtained here. (To try it: use the subcommand `POOL` in command mode or select the pooled test in Windows.)

Part 5 -- Heart Attack Patients

Setup and Data

This part uses columns c42-c44 of the same Minitab worksheet, `MINDEMO.MTW,` used in the previous parts. (Alternatively, retrieve `MINDEMO5.MTW,` which contains just these three columns.)

The study examined here involves 28 heart-attack patients admitted to a large medical center. For each of the 28 patients, blood cholesterol levels were taken on the second and fourth days after the heart attack. These data are recorded in c42 ('2nd Day') and c43 ('4th Day'). The purpose of the study is to see whether cholesterol levels of heart-attack patients tend to change in the days immediately following the event. (These data are taken from a dataset provided along with Minitab software.)

Structure of the Data

As in the previous part, we have collected two columns of data here. However, the fundamental structure of these data differs from the structure in Part 4 -- these are "paired data." In order to make the point more clearly, we show the data for the first five patients:

```
MANIP > Display Data
MTB > print c42-c44

  ROW  2nd day  4th day  4th-2nd

    1      270      218      -52
    2      236      234       -2
    3      210      214        4
    4      142      116      -26
    5      280      200      -80
   ...     ...      ...      ...
  Printout abridged to save space.
```

The first patient had a cholesterol level of 270 on the second day and 218 on the fourth day. For later use, we record 218 - 270 = -52 in column c44, indicating that this patient's cholesterol level dropped by 52 points from day 2 to day 4. The values 270 and 218 are said to be "paired" because they are a pair of measurements of the same type on the *same patient.* If the observations 270 and 218 had not been paired in this way, it would have made no sense to compute their difference. (For paired data, the order of presentation is important: the paired structure of the data would be lost if the order of the data in one column were changed without making the corresponding change in order in the other column.)

> All of the data in column c44 have been derived by computing such differences. For this example, the differences have already been computed and recorded in the Minitab worksheet for you. The command we used to make c44 was `LET c44 = c43 - c42.` (The same result could have been obtained from Windows menus: **`CALC > Calculator, store in c44, expression c43 - c42,`** either typing the minus sign or clicking the mouse on the minus sign on the "calculator.") Then we named the new column '4th-2nd'.
>
> You can test your understanding of this procedure by re-computing the differences and putting them into c45, naming your column of differences 'Diff', and then checking the worksheet to see that your 'Diff' is identical to our '4th-2nd'.

Exploration and Analysis of the Data

Descriptive Techniques. We could use parallel dotplots of the data for the second day and for the fourth day to try to understand whether cholesterol levels tend to change just after a heart attack. Such dotplots do show a slight difference between the patterns of 2nd-day levels and 4th-day levels. However, this is **not** an effective or proper way to look at paired data. The main difficulty is that we cannot see which dot in the first plot corresponds to which dot in the second. Because of the pairing, such comparisons are important.

INEFFECTIVE PROCEDURE
GRAPH > Character > Dotplot, same scale

```
MTB > dotp c42 c43;
SUBC> same.
                                                .
                                       .        :
                    .   .    .   .. .: :..   .:.: :.  . .       .
           -+---------+---------+---------+---------+---------+-----2nd
day
                                .   :   .
               .     .      ... :  .:  ::.: . ... ..          .
           -+---------+---------+---------+---------+---------+-----4th
day
         100.00    150.00    200.00    250.00    300.00    350.00
```

What we really want to show in an effective plot is the difference *each patient* shows in cholesterol levels between Day 2 and Day 4. Hence, it is best to plot the differences in c44.

```
GRAPH > Character > Dotplot
MTB > dotp c44

                                                 .
            .   .    .     .....   :  :..  : : . :. :    .   . .
           -------+---------+---------+---------+---------+---------
4th-2nd
               -90.00    -60.00    -30.00      0.00     30.00     60.00
```

This plot shows that, even though cholesterol levels increased for some patients (specifically, 9 of the 28) over the two day span of time, levels decreased for most of them (the other 19). In social science and medical data there are seldom absolutes. All we can conclude from this picture is that decreases in cholesterol seem to happen more often than increases.

Inferential Procedure. The average difference in our *sample* is a decrease of about 23 units. If our small sample is typical of the population of heart-attack patients, our best guess at the average *population* decrease is also about 23 units. How much different from 23 units might the population value be? A confidence interval procedure (based on the t-distribution and giving rise to a command called `TINTerval` in Minitab) says that we can have reasonable confidence that the average change in the population is a decrease of between 8 and 38 units (i.e., 23 plus or minus an error factor of 15).

Because the confidence interval does not include zero or any positive values, it is very likely that the actual population tendency would be for a decrease in cholesterol levels following a heart attack.

```
CALC > Basic > 1-Sample t, confidence interval
MTB > tint c44

              N      MEAN    STDEV  SE MEAN   95.0 PERCENT C.I.
 4th-2nd     28    -23.29    38.28     7.23  (  -38.13,   -8.44)
```

If you have studied one-sample t-tests, you should try the command `TTESt 0 c44` to test the null hypothesis that the population mean difference is zero. (In Windows select the same menu path as above, but test the mean, with alternative "not equal," instead of computing a confidence interval.) The very small P-value indicates that the null hypothesis should be rejected. Furthermore, the observed average change is a decrease in cholesterol levels. In plain English this means that the data show a meaningful decrease in cholesterol levels between Day 2 and Day 4.

Comments

Again we see that a formal inferential procedure can confirm what we see in a properly made graphic display. Notice, however, that a clear understanding of the structure of the dataset is required in order to do a reasonable analysis -- graphic or numerical. It was not enough to see that the data from this experiment were recorded in two columns representing experimental variables that need to be somehow "compared." An understanding of the paired structure of the data was crucial to the proper analysis of the data. Contrast the analysis of this data set with the analysis of the hot dog data in Part 4 which also involved a "comparison" of two columns of data.

Part 6 -- Education and Income

Setup and Data

This part uses the last two columns of the Minitab worksheet `MINDEMO.MTW` or `MINDEMO6.MTW.`

It consists of data for two demographic variables collected by the U.S. Bureau of the Census in 1970 and summarized for zip codes:

- Yrs Educ: Median years of education (for adults 25 years or older),
- HH Incom: Median household income.

Out of the approximately 32,000 zip codes in the U.S. we have data for a sample of only 500. (This particular sample is not random, but similar results to the ones we shall see here would be obtained from using a random sample of residential zip codes.) As we work with these data remember that we are not dealing with characteristics of individual people, but summaries for zip codes which may contain individuals with a wide variety of incomes and educational backgrounds.

One suspects -- at least hopes -- that there is a positive association between education and income. (Positive association means that an increase in one variable is associated with an increase in the other). One common way to measure the degree of association is the *coefficient of correlation,* often denoted by *r*. We use Minitab to find the correlation between income and education for the 500 zip codes in the sample.

STAT > Basic > Correlation

```
MTB > corr 'HH Incom' 'Yrs Educ'

 Correlation of HH Incom and Yrs Educ = 0.606
```

Based on this information we might be tempted to try to find the equation of a regression line that expresses income as a function of education. Minitab's computation of this line is shown below.

STAT > Regression > Regression, with one predictor

```
MTB > regr 'HH Incom' 1 'Yrs Educ'

 The regression equation is
 HH Incom = - 23833 + 4041 Yrs Educ

 Predictor       Coef       Stdev    t-ratio        p
 Constant      -23833        2882      -8.27    0.000
 Yrs Educ      4040.7       237.5      17.01    0.000

 s = 6523        R-sq = 36.7%     R-sq(adj) = 36.6%

 ...      ...      ...
```

> Notes: In a simple linear regression such as we have here, we attempt to express a "dependent" (or "response" or "predicted") variable in terms of one "independent" (or "explanatory" or "predictor") variable. Minitab's `REGRession` command can also be used for the situation in which there are several independent variables. For simple linear regression, the dependent variable is mentioned first and the single independent variable follows. Here the number "1" between variables tells Minitab that we will use only one independent variable.
>
> Minitab's "Analysis of Variance" table and a long list of "Unusual observations" are omitted here to save space. If your installation can display only one page of information at a time, press "y" as often as necessary to see all of this information, or "n" at any point to avoid further output.

In the kind of interpretation of the regression line often seen in the popular press, one might say that each year of education is worth a little more than $4000 dollars in increased annual household income. With this interpretation, the negative term in the regression equation would surely cause trouble if we tried to use the equation to predict incomes in zip codes where the median education is less than six years.

The R-square value of about 37% says that about 37% of the variability in income can be "explained" in terms of the amount of education. This quantity, the square of the correlation, is called the coefficient of determination. For a relationship in which one variable can be perfectly predicted as a linear function of another, the coefficient of determination is 100%. (In this demonstration, we will not examine the other numbers in the Minitab printout in detail.)

In fact, it is an abuse of computer technology to grind out the numbers for correlation and regression for these data. Both of these techniques rely on the *assumption* that the association between income and education can usefully be viewed as a *linear* one -- and the fact of the matter is that the true nature of the association is much more interesting than that. (The very long list of "unusual observations" generated by Minitab, but not reproduced above, is a strong indication that the assumption of a linear relationship is not appropriate here.)

A scatter plot of these two variables is shown below. Note that the variable on the vertical axis is mentioned first in the command. The Minitab `PLOT` command used here makes a rather crude "character graphics" image which is adequate for many purposes. An asterisk `*` indicates a single data point. Numbers `2-9` indicate the number of points that fall at the same plotting location. The `+` symbol indicates that more than 9 points fall at a location.

> Release 7 of Minitab allows more detailed plotting with the command `GPLOt` instead of `PLOT.` Windows menus can also generate a different kind of `PLOT` command in which the variables are separated by an asterisk and the resulting plot is more elegant than the one shown here. If you have access to one of these versions of Minitab, you should also look at the corresponding higher-resolution plot.

GRAPH > Character > Scatterplot or **GRAPH > Plot**

```
MTB > plot 'HH Incom' 'Yrs Educ'

          -
     75000+                                                        *
          -
 HH Incom -
          -                                                *      **
          -                                  *       *        2  *
     50000+                                       *              *
          -                                   *2    *     *
          -                                 34**2 *      *       *
          -                            * 3*+925 ** 2  *      2    *
          -                2         * **+7++*2*   *          *
     25000+      ** *   2**   *   2*3* 234++++36   *  *     *
          -         ***3*2 23 232*5422568++++65   **
          -    3  * * 2  2** 3324222****237 *
          -       *   **  *   *   *  **                             *
          -
         0+
            ----+---------+---------+---------+---------+---------+-Yrs
Educ
              8.0       9.6      11.2      12.8      14.4      16.0
```

We see from this plot that the true relationship between income and education is that values fall only in the triangle that lies below the diagonal running from lower left to upper right. There are **no** zip codes in the sample of 500 having both low education and high income. (There may be occasional low-education, high-income *individuals* lurking in the zip codes, but not enough of them to show up in zip code summary figures.)

On the other hand, high education is sometimes paired with low income. (Perhaps the one zip code with very high median years of education and very low median household income consists predominantly of graduate student housing.) Education can be viewed as providing a potential for high income, but not a guarantee.

The problem with correlation and regression methods here is not that the points fail to lie precisely on a line, nor that the coefficient of determination is only 37%. In the social and biological sciences an R-squared value of 37% sometimes indicates a meaningful association between two variables. A perfect fit to a line is not required. What is required is that no other relationship works substantially better.

This dataset shows that exploratory graphical analysis can reveal unexpected structure of practical significance. Information gained from looking at simple graphic displays can be very helpful in deciding what kinds of more formal analysis is appropriate.

Notes for Minitab Demonstrations by Bruce E. Trumbo,
Department of Statistics, CSU Hayward, Hayward CA, 94542,
Email: btrumbo@csuhayward.edu.
Comments and corrections welcome.