

1. The author Shere Hite undertook a study of women's attitudes toward love and sex by distributing 100,000 questionnaires through women's groups. Only 4.5% of the questionnaires were returned. Based on this sample of women, Hite wrote *Women and Love*, a best-selling book claiming that women are fed up with men. For example, 91% of the divorced women in the sample said that they had initiated the divorce and 70% of the married women said that they had committed adultery. Explain briefly why Hite's sampling method is nearly certain to produce a strong bias. Explicitly state what the population and the sample are. Hint: Think about the types of errors in surveys we considered in class.

SOLUTION: Presumably, Hite considers the population to be "all women," or perhaps "all American women." The sample, on the other hand, is "all women who returned Hite's questionnaire." In either case, there was a strong selection bias because the questionnaire was sent only to women's groups; women who did not belong to women's groups had no chance of appearing in the sample. There was also a potential non-response bias (perhaps only disgruntled women responded).

2. Consider this set of eight values:

250 206 235 211 261 208 174 214

Find the mean, median, and standard deviation for this set.

SOLUTION: The total of the values is 1,759, so the average (mean) is $\bar{x} = \frac{1,759}{8} = 219.875$

The values, sorted into order, look like this:

174 206 208 211 214 235 250 261

The median value occupies positions 5 and 6, so we report $\tilde{x} = x_{\text{median}} = \frac{211 + 214}{2} = 212.5$.

There are several strategies to compute the standard deviation $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}} =$

$$\sqrt{\frac{\sum_{i=1}^8 (x_i - \bar{x})^2}{7}} .$$

The tough part is the sum $\sum_{i=1}^8 (x_i - \bar{x})^2$. You can get directly the sum

$(x_1 - \bar{x})^2 + \dots + (x_8 - \bar{x})^2$, but since \bar{x} is given to six significant figures, this

procedure looks very awkward. Moreover, it is error-prone. The best strategy is to abandon this method.

You can use $\sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$. The hardest item here is $\sum x_i^2$, but it's still not bad. Find

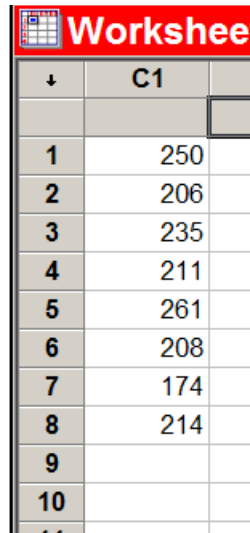
$$\sum x_i^2 = 250^2 + 206^2 + 235^2 + 211^2 + 261^2 + 208^2 + 174^2 + 214^2 = 392,139$$

Then

$$\begin{aligned} \sum (x_i - \bar{x})^2 &= 392,139 - \frac{(1,759)^2}{8} = 131,826 - \frac{3,094,081}{8} \\ &= 392,139 - 386,760.125 = 5,378.875 \end{aligned}$$

We are able then to assemble $s = \sqrt{\frac{5,378.875}{8-1}} \gg \sqrt{768.410714} \approx 27.72$.

You could bypass all this work and just enter the eight values into a data column in Minitab. Your sheet would look like this:



	C1
1	250
2	206
3	235
4	211
5	261
6	208
7	174
8	214
9	
10	

The command **Stat** \Rightarrow **Basic Statistics** \Rightarrow **Display Descriptive Statistics** would then give

Descriptive Statistics: C1

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
C1	8	0	219.88	9.80	27.72	174.00	206.50	212.50	246.25	261.00

3. Consider these values:

11 17 18 10 22 23 15 17 14 13 10 12 18 18 11 14

Find the mean, median, and mode for these.

SOLUTION: There are 16 values, and these 16 values have a total of 243, so the mean (or average) is $243 \div 16 = 15.1875$. If you sort the values into ascending order, you get

10 10 11 11 12 13 14 14 15 17 17 18 18 18 20 23

↑

The median of the 16 values occurs between the 8th and 9th largest, at the position indicated by the arrow. We give 14.5 as the median.

The value 18 occurs three times, more often than any other value, so we report 18 as the mode.

4. A set of $n = 80$ values has average 14,880.16. After all the work is completed, you discover that a value originally recorded as 12,148 should have been 11,248. If you replace the value 12,148 by 11,248, what will be the corrected average?

SOLUTION: The original total was apparently $80 \times 14,880.16 = 1,190,412.8$. If you remove the incorrect 12,148 and replace it by the correct 11,248, the total will change by $11,248 - 12,148 = -900$. The corrected total is then $1,190,412.8 - 900 = 1,189,512.8$, and the corrected average is $\frac{1,189,512.8}{80} = 14,868.91$.

There are other ways to think about this. For example, you could note that decreasing the total by 900 translates to decreasing the average by $\frac{900}{80} = 11.25$.

The new average must be $14,880.16 - 11.25 = 14,868.91$.

5. The file 97EMPLOY.CSV contains five columns. (This file is available on the course eeb site.) Data are numbers of employees for various US airlines in 1997; these are approximately the averages of the 12 monthly employee counts.

Variable names:

AIRLINE

FULLTIME

PARTTIME

TOTAL

FTEquiv

TYPE

Sum of FULLTIME + PARTTIME

FULLTIME + 0.5 * PARTTIME

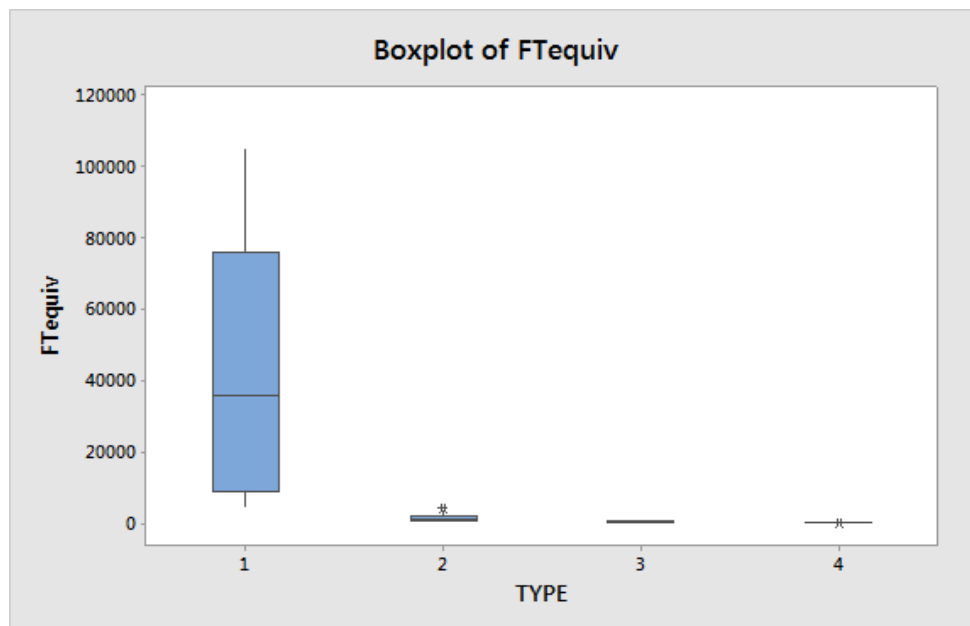
1=MAJOR, 2=NATIONAL, 3=LARGE REGIONAL,
4=MEDIUM REGIONAL

- a. The values of FTequiv vary substantially according to TYPE. Produce a display in which there are four side-by-side boxplots, so that the size differences among the TYPEs are displayed graphically. HINT: Use **Graph** \Rightarrow **Boxplot**. Then ask for **With Groups**; in the **Graph variables** box, name FTequiv (or C5) and in the **Categorical Variables** box name Type (or C6).
- b. The display created in part a will leave you with only one clear impression. We can see more if we utilize logarithms of FTequiv to create a new variable which is $\log_{10}(\text{FTequiv})$. Use **Calc** \Rightarrow **Calculator** \Rightarrow to make $C7 = \log_{10}(\text{FTequiv})$; the Minitab code for base-10 logarithms is LOGTEN. Now produce four side-by-side boxplots for the values of TYPE. Name this new variable as LogFTequiv.

You will find an unusual airline in the group TYPE=2. What is this airline?
HINT: Try **Editor** \Rightarrow **Brush** \Rightarrow on the graph. Please note that this calls for **Editor**, not **Edit**.

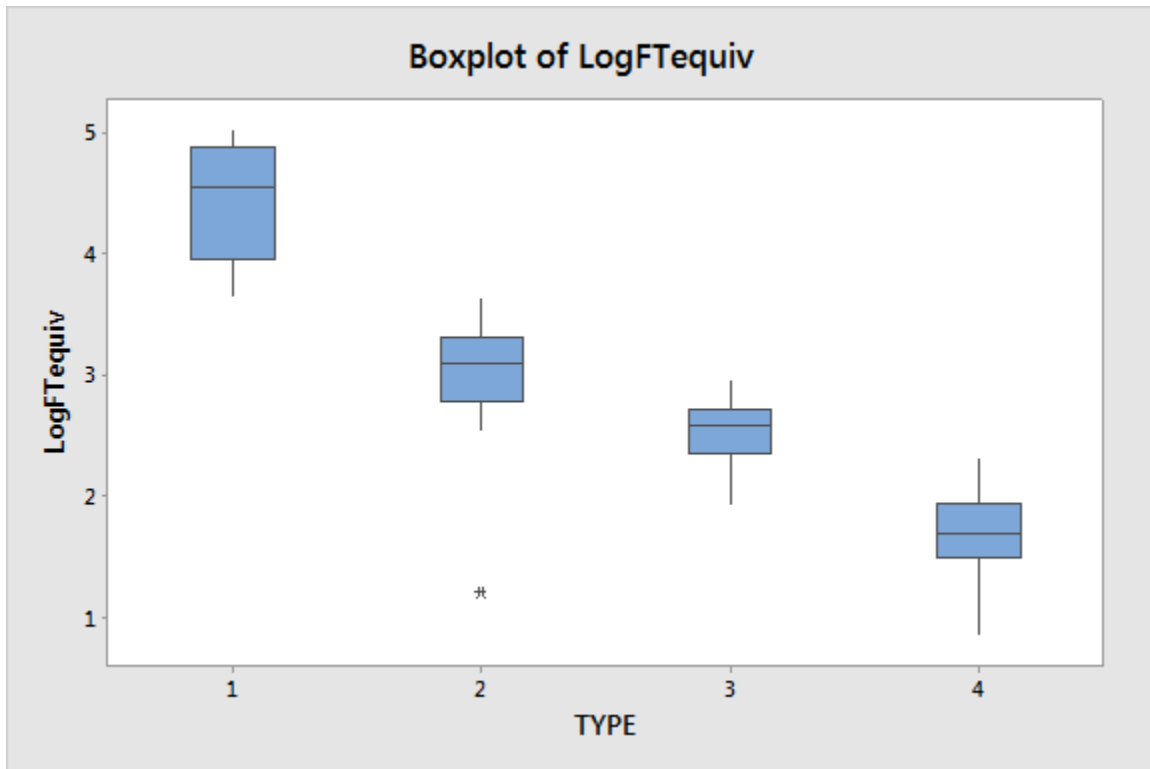
- c. Produce a graph showing PARTTIME on the vertical axis and FULLTIME on the horizontal axis. Identify any airline which seems to have an unusual mix of part-time and full-time employees.
- d. Part c might come out differently if you plot $\log(\text{PARTTIME})$ versus $\log(\text{FULLTIME})$. Continue to use base-10 logarithms. Do you still find the same unusual airline(s)?
- NOTE: A number of the airlines have no part-time employees. Use the transformation $\text{LOGTEN}(\text{PARTTIME}+0.5)$.

SOLUTION: For part a we do **Graph** \Rightarrow **Boxplot** \Rightarrow and put C6 (or TYPE) in the X column. The result will be this:



This certainly succeeds in showing us that the major airlines have many more employees. Unfortunately, there is not enough resolution in the picture to make judgments about the other three groups.

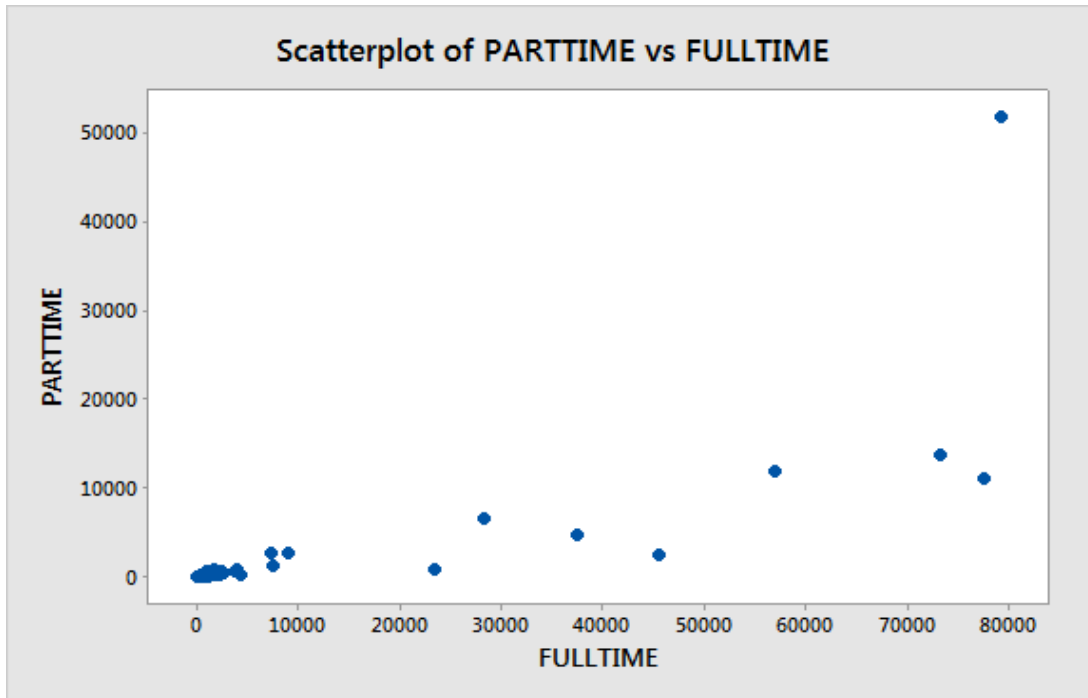
For part **b**, we use **Calc** \Rightarrow **Calculator** \Rightarrow to create a new variable (here called LogFTEquiv) which is the base-10 logarithm of FTEquiv. The box plot of these is the following:



Observe that the four TYPEs are cleanly separated. We see that the values in the TYPE=4 group (medium regional) spread out around $10^{1.7} \approx 50$, the values in the TYPE=3 group (large regional) are spread around $10^{2.6} \approx 400$, and the values in the TYPE=2 group (national) spread out around $10^3 \approx 1,000$.

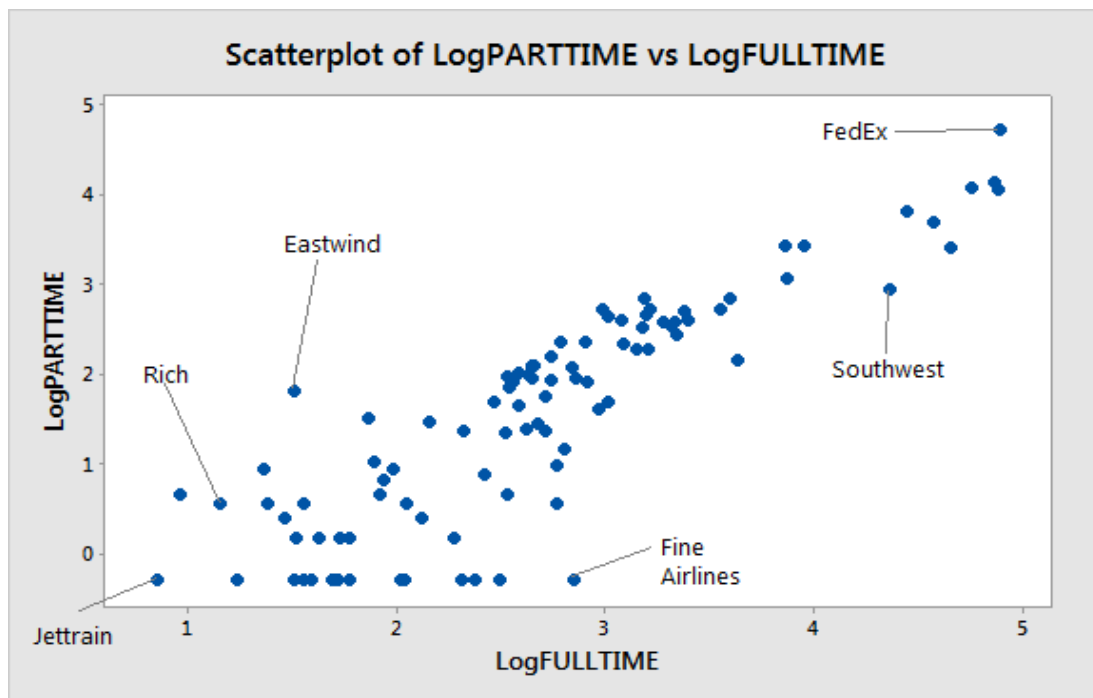
The low outlier in group TYPE=2 is point 36, Rich Airlines. Has anyone heard of this? Has anyone actually flown it?

For part **c**, we do **Graph** \Rightarrow **Scatterplot** \Rightarrow and name PARTTIME as Y, FULLTIME as X. The resulting picture will be this:



This suggests one very unusual point. Using **Editor** \Rightarrow **Brush** on the graph, this can be identified as point 7, Federal Express. This is interesting. We see that Federal Express has a very large number of part-time people. This is, of course, *not* a passenger airline.

d. Using **Calc** \Rightarrow **Calculator** \Rightarrow create the logarithms of these two variables. Now the plot will be this:



Here Federal Express is still the point at the upper right, however it takes a very subtle eye to decide that it is unusual. Another point noted as possibly unusual is point 73, Eastwind Airlines. This airline has twice as many part-time as full-time people!

Besides the graphical methods, you might simply look at the obvious ratio $\text{PARTTIME} \div \text{FULLTIME}$. Based on this, we would find

There are 14 airlines with no part-time people at all.

Beyond these, there are an additional 18 airlines with this ratio in the interval $(0, 0.05)$.

There are three airlines for which this ratio exceeds 0.50. These are Executive (0.55), Federal Express (0.65), and Eastwind (1.97).

6. Indicate (without computation) which sample in each set has the higher standard deviation.

Set 1, Sample A: 16, 16, 16, 16, 16

Set 1, Sample B: 15, 16, 16, 16, 16

Set 2, Sample A: 20, 25, 25, 25, 30

Set 2, Sample B: 15, 25, 25, 25, 35

Set 3, Sample A: 20, 20, 30, 40, 40

Set 3, Sample B: 20, 25, 30, 35, 40

SOLUTION: In Set 1, Sample B has the larger standard deviation. In Set 2, Sample B has the larger standard deviation. In Set 3, Sample A has the larger standard deviation.

7. Sincich, problem 2.66. The data file presents the data to you in two formats.

Separate columns format:

Column C1 gives the scores for the 33 children in the Honey group.

Column C2 gives the scores for the 35 children in the DM group.

C3 gives the scores for 37 children in the the Control group.

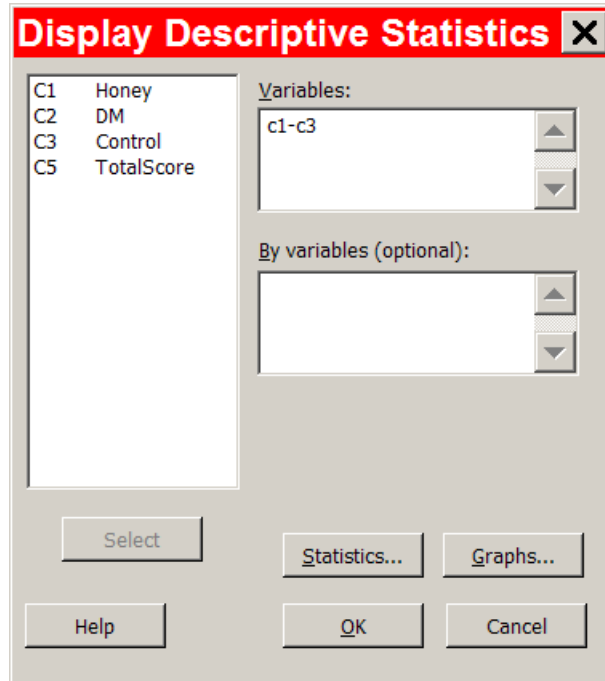
Single column plus identifier format:

Column C5 gives the scores for all $33 + 35 + 37 = 105$ children.

Column C6 gives the labels corresponding to the three groups.

You can do this problem through either of the two formats provided. The single column plus identifier format generalizes to other situations and is vastly more useful.

SOLUTION: For parts **a** to **c**, you can use the separate columns format by asking for **Stat** ⇒ **Basic Statistics** ⇒ **Display Descriptive Statistics** for columns C1, C2, C3. The panel will look like this:



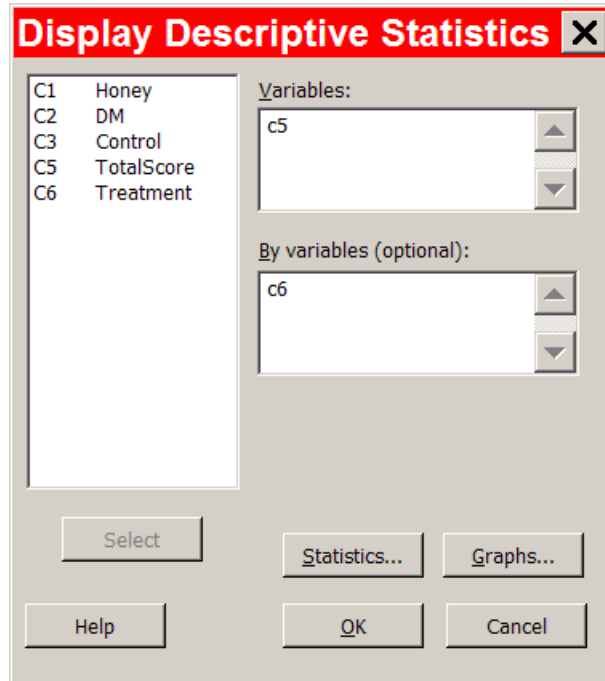
The output is the following:

Descriptive Statistics: Honey, DM, Control

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Honey	35	0	10.714	0.483	2.855	4.000	9.000	11.000	12.000	16.000
DM	33	0	8.333	0.567	3.256	3.000	6.000	9.000	11.500	15.000
Control	37	0	6.514	0.483	2.940	0.000	5.000	7.000	8.000	12.000

The requested standard deviations are 2.855 (Honey), 3.256 (DM), and 2.940 (Control). We would regard these as very close. You could then supply the answer to **d**, but the wisest response might be “too close to worry about the differences.”

If you used the single column plus identifier format, the panel would be this:



The output would be nearly identical to the previous, except for the labels:

Descriptive Statistics: TotalScore

Variable	Treatment	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
TotalScore	C	37	0	6.514	0.483	2.940	0.000	5.000	7.000	8.000	12.000
	DM	33	0	8.333	0.567	3.256	3.000	6.000	9.000	11.500	15.000
	H	35	0	10.714	0.483	2.855	4.000	9.000	11.000	12.000	16.000

8. Larcenous Larry has a problem. The data column SALES had the following summary:

Descriptive Statistics: SALES

Variable	N	N*	Mean	StDev	Sum	Minimum	Q1	Median	Q3	Maximum
SALES	28	0	4319	2693	120925	1674	1842	3062	6542	8947

Unfortunately, he had promised that SALES would average at least 5,000. To cover this up, he maliciously edited the values to this:

Descriptive Statistics: SALES

Variable	N	N*	Mean	StDev	Sum	Minimum	Q1	Median	Q3	Maximum
SALES	28	0	5019	2693	120925	1674	1842	3062	6542	8947

When an auditor looks at the latter display, Larry's misdeeds will be quickly detected. How?

SOLUTION: The auditor will most likely notice the disconnect between the average and the total. If Larry changed the mean to 5,019, he should also have changed the total to $28 \times 5,019 = 140,532$.