Chapter 3 – Understanding and Comparing Distributions

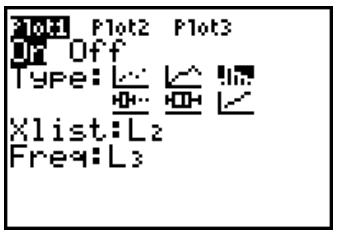
In this chapter, we will meet a new statistics plot based on numerical summaries, a plot to track the changes in a data set through time, and ways to use plots to compare two or more distributions.

BOXPLOTS

Box plots (sometimes called box-and-whisker plots) are another way of picturing a distribution. Unlike histograms, they are based on definite values and are not subjective. However, as no plot is perfect, they can hide some potential features such as bimodality. A good practice, since it is generally so easy, is to look at several plots. They all can show different features of the distribution.

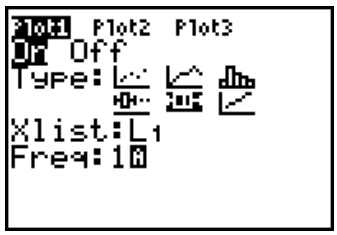
There are two types of boxplots – the original, which is based solely on the five-number summary (min, Q_1, median, Q_3, and max), and a "modified" boxplot, which has an objective criterion to identify outliers. Both types of plots divide the data into fourths – a "whisker" for the bottom and top quarters of the data, and a box for the middle half, with the median indicated inside the box. We always recommend using the modified boxplot, but your instructor may suggest otherwise.

As always, begin with data in a list. We will begin with the pulse rate data already examined in the last chapter. The data are in `L1`. Press 2nd Y= (`StatPlots`) then ENTER to get to the plot definitions screen for `Plot1`.

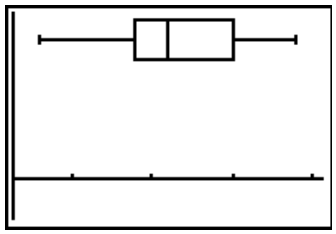


Notice there are two choices for boxplots - which is the boxplot where outliers are identified and which does not identify outliers. These are called `Box Plot` and `Mod Box Plot` on a TI-89. We will look at both to see the difference between the two.

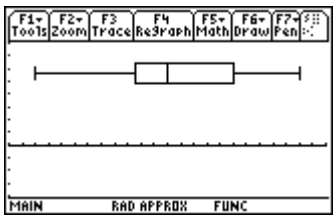
Move the cursor to highlight . Make sure `Xlist` is changed to `L1` (press 2nd 1); also make sure `Freq:` is set to 1 (press ALPHA 1 if necessary). Your plot definition screen should look like this.



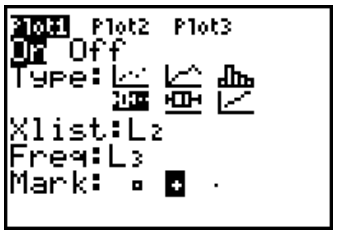
Press ZOOM 9 (F5 on an 89) to display the graph. The indications from the plot are that the distribution is (roughly) symmetric. Looking from the median line in the box to the two ends, the distances are relatively equal. However, since the median is somewhat to the left of the center of the box, one could call the distribution somewhat skewed. Pressing TRACE and using the right and left arrows will allow you to move around the graph, locating the median, quartiles, min, and max.



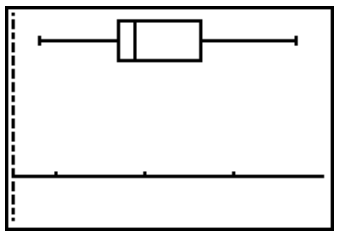
Return to the Plot definition screen and change the plot type to a boxplot identifying outliers (), or `Mod Box Plot` on an 89. Pressing ZOOM 9 will display the plot at right. We see here that none of the values are outliers by the $Q_3 + 1.5*IQR$ and $Q_1 - 1.5*IQR$ criteria.



BOXPLOTS WITH TABULATED DATA



Reconsider the data on heights of members of a choir. According to the histogram, this was somewhat right-skewed. What will its boxplot look like? With heights in L2 and frequencies in L3, the plot definition screen looks like that at right.



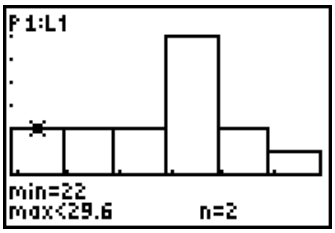
After pressing ZOOM 9 the graph at right will be displayed. Notice the median is not in the middle of the box; the right half of the data is longer than the left half (the data is right skewed) even though the two whiskers are relatively equal in length. Also, since we defined the plot to identify any possible outliers, none are flagged, so these data have no outliers.

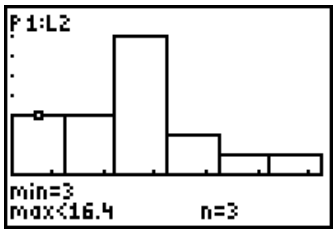
HISTOGRAMS TO COMPARE DISTRIBUTIONS

We'd like to compare the distributions of two historically great baseball hitters: Babe Ruth and Mark McGwire. We have the following information on the numbers of home runs hit each year by Babe Ruth for 1920 through 1934 and for McGwire from 1986 to 2001.

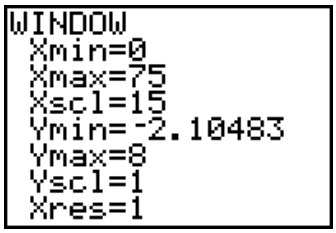
Ruth:		52	59	35	41	46	25	47	60	54	46	49
	46	41	34	22								
McGwire:		3	49	32	33	39	22	42	9	9	39	52
	34	24	70	65	32	29						

ZOOM 9 histograms as described in the previous chapter for both distributions are below.

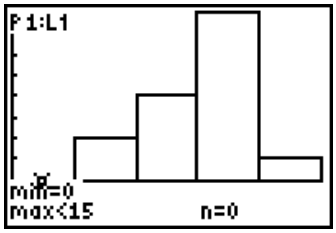


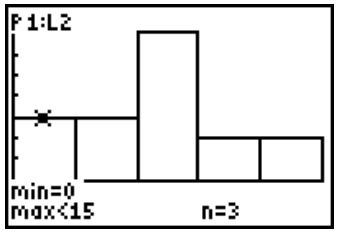


From the plots we can see both distributions are unimodal; Ruth's is skewed left and McGwire's is skewed right. But that's about all we can tell, because the graphs don't use the same scaling. People's eyes want to make a visual comparison, and since the graphs don't use the same values, this is impossible.



Let's change the scaling. Press WINDOW. Both the smallest and largest values occur in McGwire's distribution: 3 and 70. We also need a reasonable number of bars. It seems reasonable to set Xmin to 0, Xmax to 75 and use a bar width (Xscl) of 15. The settings used are at right, and the rescaled plots are below.



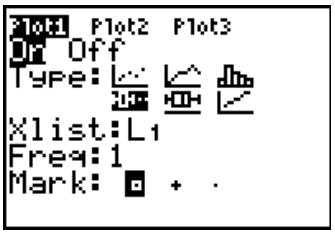


From these graphs it is easy to see that uth was the more prolific hitter. While McGwire had four years in which he hit more than 45 home runs, Ruth had 9. McGwire also had three seasons with fewer than 15 homers (due either to injury, strike, or his first, partial season). On this scaling, we still see Ruth's distribution as left-skewed, but McGwire's is fairly symmetric (even around the peak). Ruth's distribution has a smaller range (spread from low to high), and a higher center (visually about 60) than McGwire, whose center is between 45 and 60.

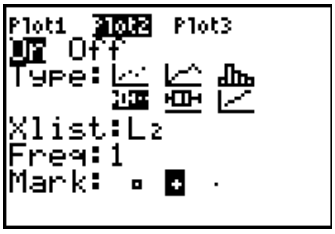
BOXPLOTS TO COMPARE DISTRIBUTIONS

Boxplots are very useful in comparing distributions. This is one of the few exceptions to the rule about only one plot being turned on at a time. Up to three boxplots can be displayed at once. Displaying boxplots side-by-side is a good way to make a visual comparison of distributions.

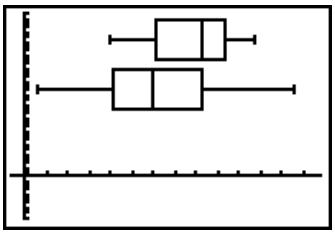
How would these distributions look displayed together as boxplots? The data have been entered into L1 (Ruth) and L2 (McGwire). We define Plot1 to use Ruth's home run values as at right.



Returning to the STAT PLOT menu, we will arrow down to Plot2, and define it to use McGwire's numbers as at right.



Pressing ZOOM 9 gives the display at right. The top plot is Ruth's home run distribution; the bottom is McGwire's. (Pressing TRACE will identify each plot; to move from one to the other, press the down and up arrows.) Neither distribution has outliers. Ruth's is much less variable than McGwire's and is skewed left. McGwire's distribution appears rather symmetric.

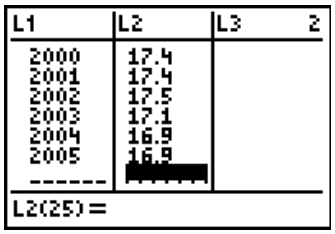


TIME PLOTS

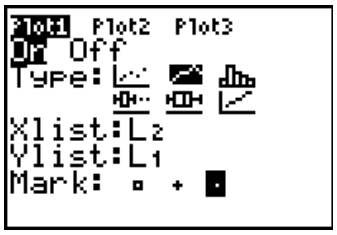
Many variables are often measured at different points in time. It's not enough just to picture the distribution in this case. Time is an important factor, and we will want to know what (if any) part it plays. To answer this question, we will do a time (series) plot of the data. By convention, these are connected scatter plots with time represented on the *x*-axis and the actual variable values on the *y*-axis. They are connected because this makes any pattern easier to see than if the data points were just shown by themselves.

Problem 46 of the text in Chapter 5 is concerned with drunk driving. It lists the number of deaths (in thousands) attributed to alcohol-related driving form 1982 to 2005. The data are reproduced on the next page for convenience.

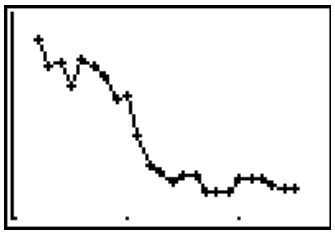
Year	Deaths	Year	Deaths
1982	26.2	1994	17.3
1982	24.6	1995	17.7
1984	24.8	1996	17.7
1985	23.2	1997	16.7
1986	25.0	1998	16.7
1987	24.1	1999	16.6
1988	23.8	2000	17.4
1989	22.4	2001	17.4
1990	22.6	2002	17.5
1991	20.2	2003	17.1
1992	18.3	2004	16.9
1993	17.9	2005	16.9



We could enter all the years manually, but it is easier to use the seq(command as described on page 10 of this manual. The data are in L2 and the years have been entered into L1.



Connected scatter plots are the second plot type on the plot definition screen. The Xlist is the time indices and Ylist is the number of drunk driving deaths. There is a choice of three options for marking the actual data points. You may pick whichever one you like; however, from past experience this author recommends the single pixel for this particular graph, as the others can make the plot look too cluttered. Once the plot has been defined, it can be displayed by pressing ZOOM 9.

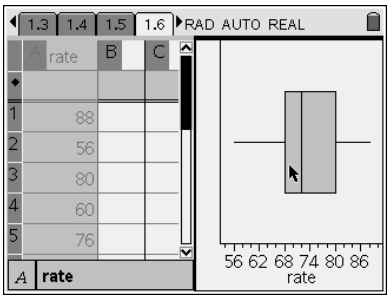


Here is the time plot. We can clearly a decline in alcohol-related fatalities in the early years of the data. However, the rate seems to have stabilized of late. What can be done to encourage less alcohol consumption combined with driving?

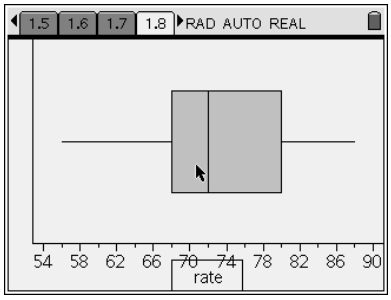
WHAT CAN GO WRONG?

Primarily, errors here are due to data entry mistakes. Always double check what you have entered. If a boxplot fails to display, the plot may not have been turned on. Always be sure to turn off any extra plots after copying them to paper. If not, you probably will receive either the Invalid Dim (referring an empty list) or Stat(Plot Setup on a TI-89) which is the incompatible window ranges for two plots error message. These were discussed in the previous chapter.

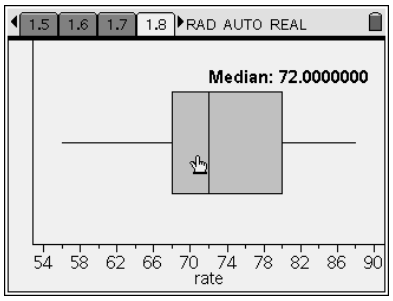
Commands for the TI-Nspire™ Handheld Calculator



To create a boxplot, first enter the data into a list. In the previous chapter we created a list of pulse rates named *rate*. Press (home) and insert a Lists and Spreadsheet page. Move the cursor to the top of the column A and type the name "rate". You should see the list of pulse rates appear on column A. With the cursor in column A, press ▲ until that the list is highlighted. Then press (menu) , select Data , and then select Quick Graph. A dotplot appears. Now press (menu) , select Plot Type, and then Boxplot.



If you prefer to see the plot on an entire screen, rather than a split screen, press (home) and then select Data & Statistics. At first you will see a plot of dots. Use the arrows to move to the bottom of the display until "Click to add variable" appears. Press (click) , highlight the variable, in this case *rate*, and (enter) again. You will see the dotplot. Now press (menu) , select Plot Type, and then select Boxplot.



Move the cursor onto the boxplot and special values such as the median will appear on the screen.