EXPLORATIONS

Chapter 2

Summarizing Distributions of Univariate Data

Topics 5 through 8 summarize the distribution of the data shown in Chapter 1. This chapter answers questions such as *What is the center of the data and how much does the data vary from the center value?* and *How does changing units affect the summary measures?* Boxplots are used to provide a schematic view of a five-number summary.

Topic 5—Measures of Center and Spread

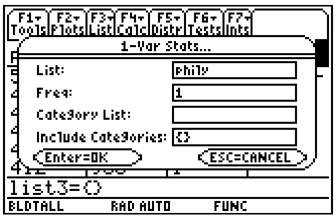
The most common measures of the center of a distribution are the mean and the median. The measures of spread are range, interquartile range, and standard deviation. These measures are given for data collected in a list or grouped in a frequency table.

Using the Data List

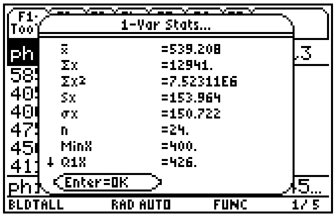
Example: Use the list of building heights from Topic 1 (list **phily** in folder **BLDTALL**).

1. From the Stats/List Editor, press F4 **Calc** and select **1:1-VarStats**.
2. Select List: **phily** with Freq: **1** (screen 1).

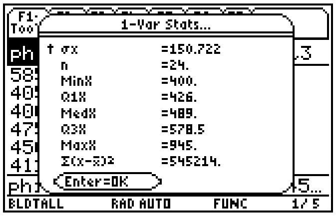
(1)



3. Press ENTER to display screen 2.

(2) 

4. Scroll down using ⊖ to display the second page of 1-VarStats (screen 3).

(3) 

Measures of Center

Mean = $\bar{x} = \dfrac{\sum x}{n}$ = 12941/24 = 539.208 feet for the data set, **phily** list of building heights (the first value in screen 2).

This is the balance point of the dotplot from Topic 2, shown below. Think of the dots as bowling balls of equal weight.

Note: *$\bar{x} = \mu$, for this case, since you are calculating the population mean of all 16 tall buildings 400 feet or taller in Philadelphia, PA.*

Median = MedX = 489 feet is the middle value (50^{th} percentile). Twelve values are below and 12 values are above 489 (the fifth value in screen 3).

Measures of Spread

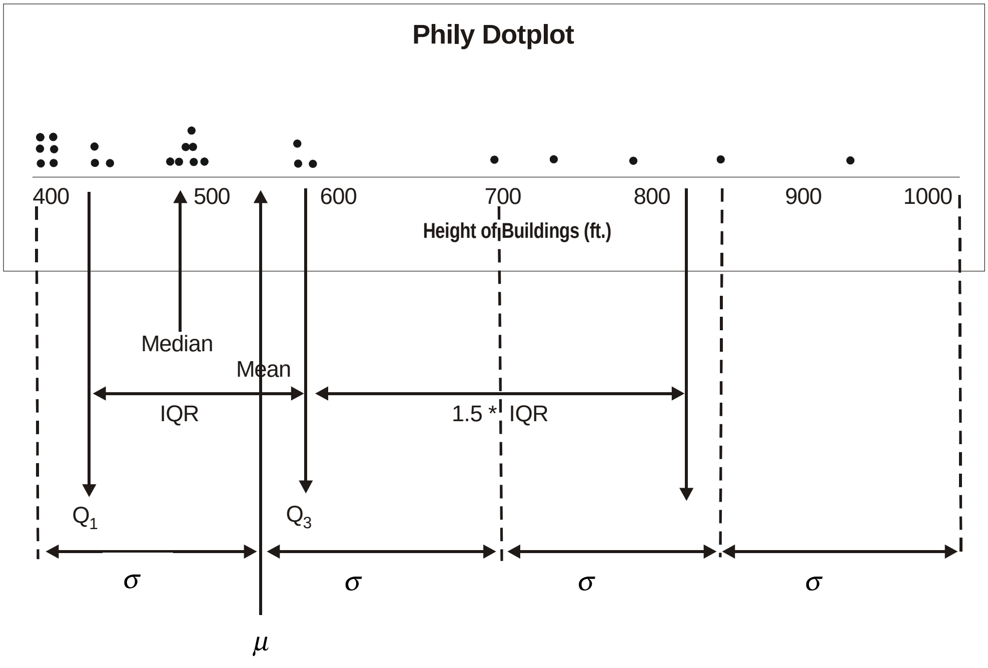
Measures of spread are:

Range = maxX - minX = 945 - 400 = 545 feet.

Interquartile Range = IQR = Q_3 - Q_1 = 578.5 - 426 = 152.5 feet.

Note: *The Interquartile Range spread of 152.5 feet covers the middle 50% of the data, with 25% or six values below and 25% or six values above, as shown in the dotplot from Q_1 to Q_3.*

1.5 ∗ IQR = 1.5 ∗ 152.5 = 228.75 *feet is used to measure the distance from the quartiles, Q_1 and Q_3, to detect possible outliers. From the dotplot with Q_3 x = 578.5 feet (with 75% of the values below it and 25% or six values above it) measure* 573.5 + 228.75 = 807.3 *feet. Only two buildings are taller than 807 feet and are identified as possible outliers at 848 and 945 feet. There are no possible outliers to the left of Q_1.*



Standard Deviation

The *standard deviation* is a measure of the deviations from the mean, and its value can be highly influenced by values a long way from the mean. The following table compares values by dropping the largest value (945) from the list.

	n = 24	n = 23	Difference
Mean	539	522	17
Median	489	488	1*
Range	545	448	97
IQR	153	155	2*
s_x	154	130	24
σ_x	151	127	24

* Robust Measures: Values that do not change much by the presence of extreme values are called *robust*. Notice that the median and interquartile ranges are more robust than the mean, range, and standard deviation.

s_x = sample standard deviation = $\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\frac{545214}{23}}$
= 153.96 ft.

σ_x = population standard deviation = $\sqrt{\frac{\Sigma(x-\mu)^2}{n}} = \sqrt{\frac{545214}{24}}$
= 150.72 ft.

Note: *If your list was a random sample of values from a population, you could use s_x to estimate σ_x.*

The standard deviation is a measure of the spread from the mean, and several intervals from the mean are given below the dotplot.

Topic 6—Measures of Position: Percentiles and Quartiles, Standard Scores (*z* scores)

Example: Use the building heights in Philadelphia, PA stored in list **phily** and folder **BLDTALL** in Topic 1.

Finding Standard Scores (*z* scores)

Standard scores (z scores) measure the number of standard deviations a value is from the mean, $z = \dfrac{x - \mu}{\sigma}$.

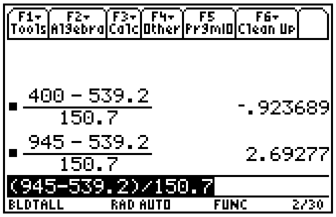
1. Press $\boxed{\text{HOME}}$ to perform Home screen calculations. If there are computations on the Home screen from previous work, press $\boxed{\text{F1}}$ **Tools** and **8** ↓ $\boxed{\text{CLEAR}}$ $\boxed{\text{HOME}}$ to clear the screen.

From Topic 5 you found the mean = **539.2**, and the standard deviation = **150.7**. Use these values to find the *z* score of the smallest (400 feet) and the tallest (945 feet) buildings in your list.

2. Type **(400 - 539.2)/150.7** and press $\boxed{\text{ENTER}}$.
3. Type **(945 - 539.2)/150.7** and press $\boxed{\text{ENTER}}$ (screen 4).

Smallest Building: ***z*** **= (400 - 539.2)/150.7 = -.923689**
Largest Building: ***z*** **= (945 - 539.2)/150.7 = 2.69277**

The smallest value is less than one standard deviation below the mean, and the largest value is almost three standard deviations above the mean.

(4)


Note: *The difference in these z scores is another indication of the non-symmetric nature of the distribution (the data is skewed to the right). This is clearly shown below the dotplot in Topic 5.*

Percentiles and Quartiles

The *percentile* of a score indicates what percent of the data values are smaller in magnitude. Not all textbooks or computer programs use the same procedure for finding percentiles, but all methods give similar answers as the sample size increases.

25th Percentile = $P_{25} = Q_1$ = 426 feet in the 1-VarStats value of Topic 5 (screens 2 and 3) with 25% of 24, or six values below it.

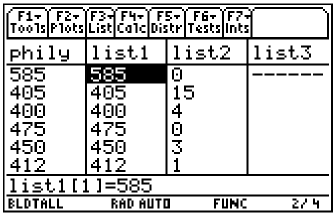
50th Percentile = P_{50} = Med = 489 feet in Topic 5 with 50% of 24, or 12 values below (and 12 values above).

75th Percentile = $P_{75} = Q_3$ = 578.5 feet in Topic 5 with 75% of 24, or 18 values below it (and six values above).

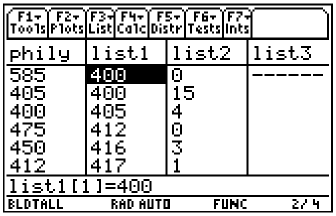
Finding the 90th Percentile

Example: Copy list **phily** data into **list1**. Sort **list1** in ascending order.

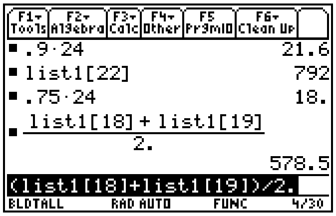
1. Return to the Stats/List Editor using [◆] [APPS].
2. Select Stats/List Editor and press [ENTER].
3. Highlight the **list1** heading, paste or type **phily** and press [ENTER] (screen 5).

(5) 

4. To sort **phily** data in ascending order, press [F3] **List**, **2:Ops** and then **1:Sort List**. Make sure the list is **list1** and press [ENTER] (screen 6).

(6) 

5. Calculate 90% of *n*, or .90 ∗ 24 = 21.6. On the Home screen, enter **.90** [×] **24** and then press [ENTER] to display the top of screen 7.
6. Use **22**, since 21.6 values are below 22. (If the result is a decimal, round up.)
7. On the Home screen, paste **list1** and then press [2nd] [[] **22** [2nd] []].
8. Press [ENTER] for **792** (screen 7).

(7) 

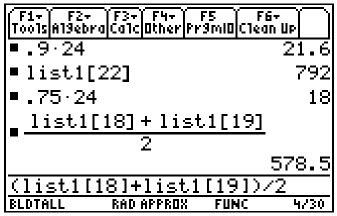
Finding the 75th Percentile

Example: Copy list **phily** data into **list1**. Sort **list1** in ascending order.

1. Calculate 75% of *n*, or .75 ∗ 24 = 18. On the Home screen, enter **.75** [×] **24** and then press [ENTER] to display the middle of screen 8.

   Since this is an integer, you will take the average of the 18th and 19th values so type **(list1[18] + list1[19])/2** and press [ENTER] (screen 8), which agrees with Q_3x of **1-VarStats** in Topic 5, screen 3.

(8)

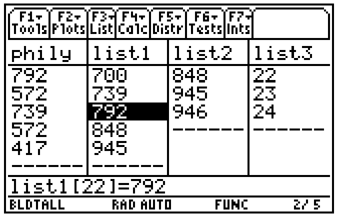


Finding the Percentile Value of a Number

Example: What percentile is the height 792 feet?

1. Return to the Stats/List Editor using [◆] [APPS].
2. Select Stats/List Editor and press [ENTER].
3. Observe that in **list1** (with the data in order), **792** is the 22nd value (screen 9). Thus, there are 21 values below it for 21/24.0 = .875 = 87.5% or the 87.5 percentile.

(9)



Topic 7—Boxplots (or Box-and-Whisker Plots)

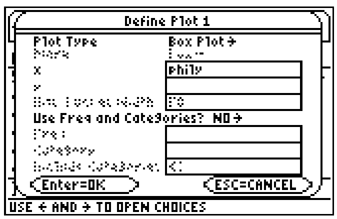
In Topic 5, screen 3, you used 1-VarStats on list **phily** for the following five-number summary:

MinX = 400	Q_1 = 426	Med = 489	Q_3 = 578.5	MaxX = 945 ft

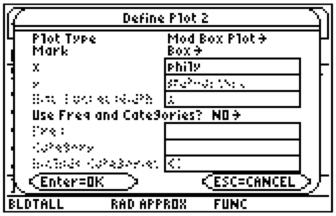
These five values define a boxplot as follows. From the Stats/List Editor and folder **BLDTALL**:

1. Set up **Plot 1** with [F2] **Plots**, **1:PlotSetup**.
2. Highlight **Plot 1**, press [F1] **Define**, and select Plot Type: **BoxPlot** and X List: **phily** (screen 10).
3. Press [ENTER] to return to the Plot Setup screen.

(10)

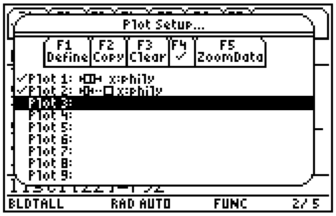


4. Highlight **Plot 2**, press [F1] **Define**, and select Plot Type: **Mod Box Plot**, Mark: **Box**, and X List: **phily** (screen 11).



(11)

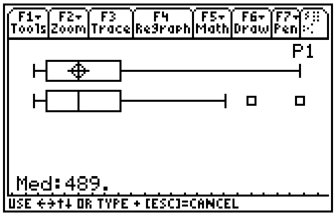
5. Press [ENTER] to return to the Plot Setup screen. Note that two plots have check marks in the left margin (screen 12).



(12)

6. Press [F5] **ZoomData** and then press [F3] **Trace**, with **Med = 489** (screen 13).

7. Use ▲ and ▶ to find the other values of the five-number summary.

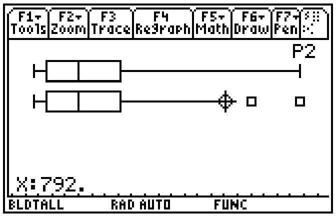


(13)

The middle box has the middle 50% of the data, while each whisker has 25% of the data. The right whisker is much longer than the left because of the positive skewness. Even the middle 50% is not symmetrical, since the median is not in the middle of the box.

Note: *If the x-axis is in the way, press* [◆] [WINDOW] *and set* **ymin = 1** *and* **ymax = 10***, and then press* [◆] [GRAPH] *and* [F3] **Trace***.*

8. Press ▼ to move the flashing cursor to the modified boxplot. The **P2** in the upper right corner represents **Plot 2** (screen 14).



(14)

Notice that the right whisker of the plot stops at 792 feet. This is the third largest value. The two boxes at **x = 848** and **maxX = 945** indicate the two possible outliers (as identified to the right of $\mathbf{Q_3 + 1.5 * IQR}$ in the dotplot of Topic 5). No outliers are identified to the left of $\mathbf{Q_1 - 1.5 * IQR}$.

There is no need to construct both boxplots. You will use the modified boxplot in this book because it shows more information.

The modified boxplots are also very helpful in comparing different distributions, as will be shown with Parallel Boxplots in Topic 9.

Topic 8—The Effect of Changing Units on Summary Measures

Changing Units with y = kx + a

Example: Change the following sample of body temperatures of 20 adults measured in degrees Fahrenheit to degrees Celsius.

98.0	**99.2**	**97.2**	**98.6**	**99.0**	**99.7**	**97.2**	**97.7**	**98.6**	**98.2**
97.0	**99.1**	**99.3**	**99.0**	**99.2**	**98.7**	**99.6**	**99.4**	**99.2**	**97.6**

$$°\mathrm{C} = \frac{5}{9}(°\mathrm{F} - 32) = \frac{5}{9}°\mathrm{F} - \frac{160}{9}$$

so in this case $k = \dfrac{5}{9}$ and $a = \dfrac{-160}{9}$

1. For measures of position (median and mean):

   a. Med $y = k *$ Med $x + a$

   b. $\bar{y} = \dfrac{y_1 + y_{2+} \ldots y_n}{n} = \dfrac{(kx_1 + a) + (kx_2 + a) + \ldots (kx_n + a)}{n} = \dfrac{k(x_1 + x_2 + \ldots x_n)}{n} + \dfrac{na}{n} = k\bar{x} + a$

   Results: Measures of position x becomes $k * x + a$

2. For measures of spread (interquartile range and standard deviation):

   a. IQRy = Q$_3y$ - Q$_1y$ = (kQ$_3x + a$) – (kQ$_1x + a$) = k(Q$_3x$ - Q$_1x$) = $k *$ IQRx

   b. $s_y = \sqrt{\dfrac{[(kx_1 + a) - (k\bar{x} + a)]^2 + \ldots [(kx_n + a) - (k\bar{x} + a)]^2}{n-1}} = \sqrt{\dfrac{k^2 \sum (x - \bar{x})^2}{n-1}} = k * s_x$

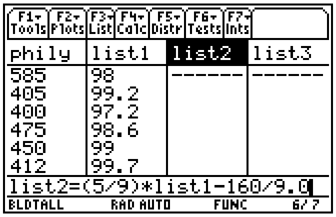
   Results: Measures of spread W becomes $k * W$

3. Press [2nd] [APPS] to return to the Stats/List Editor.

4. Clear all data in **list1**, **list2**, **list3**, and **list4** by highlighting each heading and pressing [CLEAR] [ENTER].

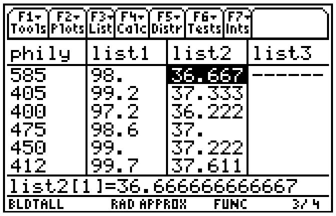
5. Type the temperatures in °F in **list1**.

6. Highlight the **list2** heading, type the temperature conversion formula: **(5/9)** ∗ **list1 – 160/9.0** (screen 15).

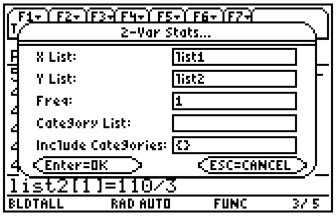
(15) 

Note: *The last number is* 9.0*, not* 9*, so the results are not fractions.*

7. Press [ENTER] and observe the Celsius temperatures in **list2** (screen 16).

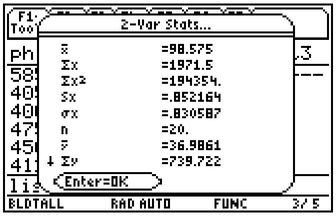
(16) 

8. Press [F4] **Calc**, and select **2: 2-VarStats**.
9. Select X List: **list1**, Y List: **list2**, and Freq: **1**.

(17) 

Note: *This does the same calculations as* 1-VarStats*, but on two lists of equal length. (If the lists are not of equal length, a dimension mismatch error will be displayed.)*

10. Press [ENTER] [ENTER] (screen 18). Pressing [2nd] ⊝ displays a second page of output and pressing [2nd] ⊝ again displays a third page of output.

(18) 

11. Check:

$$\bar{y} = 36.9861 = k\bar{x} + a = \left(\frac{5}{9}\right) * 98.575 - \frac{160}{9}$$

$$s_y = 0.473425 = k * s_x = \frac{5}{9} * 0.852164$$

$$\text{Med } y = 37.1389 = k * \text{Med } x + a = \left(\frac{5}{9}\right) * 98.85 - \frac{160}{9}$$

$$\text{IQR } y = \text{Q}_3 y - \text{Q}_1 y = .75 = k * \text{IQR}x = \left(\frac{5}{9}\right) * 1.35$$

$$\Sigma(y - \bar{y})^2 = 4.2589 = k^2\, \Sigma(x - \bar{x})^2 = \left(\frac{5}{9}\right)^2 * 13.7975$$

Changing Units by Multiplying by a Constant (y = kx)

Example: Use the building heights in Philadelphia, PA stored in list **phily** and folder **BLDTALL** in Topic 1. These heights are in feet; change them to meters.

Note: *This is a special case of $y = kx + a$ using $a = 0$. Measures of position and measures of spread (or both) are multiplied by a conversion factor k to change units.*

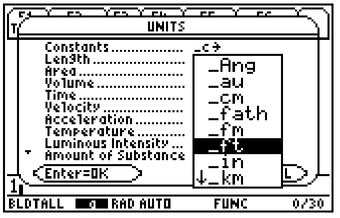
Notice that 3 ft = 1 yd, so

$$12\text{ft} = 12\text{ft} * \frac{1 \text{ yd}}{3 \text{ ft}} = 4 \text{ yd or } k = \frac{1}{3} = 0.33\overline{3}$$

to change feet to yards.

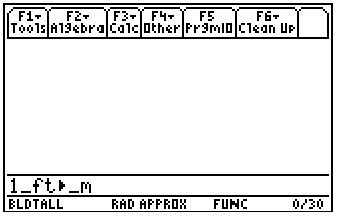
One meter is approximately a yard, but what is the conversion factor? The TI-89 has the answer stored with [2nd] [UNITS].

1. From the Home screen in folder **BLDTALL**, type **1** and then press [2nd] [UNITS] to display the UNITS screen.
2. From the Length submenu, highlight _**ft** (screen 19).
3. Press [ENTER], and then press [2nd] [▸], followed by [2nd] [UNITS]. (The [▸] arrow is above the [MODE] key.)



(19)

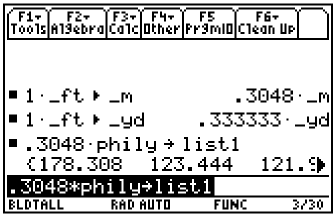
4. From the Length submenu, highlight _**m** and press [ENTER] for **1 _ft ▸_m** in the input line (screen 20).



(20)

5. Press [ENTER] for **0.3048_m** in the first line of screen 21 (or 1 ft = 0.3048 meters).

6. Repeat steps 1 through 5, except select **_yd** instead of **_m** in step 4. The second line in screen 21 shows the ft to yards conversion.

(21)


7. Multiply **0.3048 ∗ phily**, press [STO▸] for an arrow on the input line of screen 21, and then paste or type **list1**.

Note: *You could have done this step in the Stats/List Editor.*

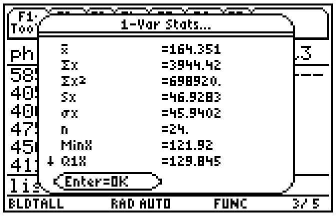
8. Press [ENTER] and the first result of **178.308** indicates that the first height in list **phily** of 585 ft = 178.308 meters (screen 21).

From the Stats/List Editor ([2nd] [APPS]):

1. Press [F4] **Calc** and select **1:1-VarStats**.

2. Enter List: **list1**and Freq: **1**.

3. Press [ENTER] [ENTER] (screen 22). Mean = **164.351** and s_x = **46.9283** meters.

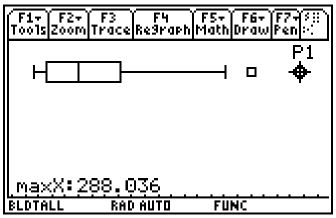
(22)


4. Compare these values in meters with the values in feet of Topic 5, screen 2, with $\bar{x}$ = **539.208** feet and s_x = **153.964** feet.

5. Check: Mean = 164.351 meters = $k * \bar{x}$ = 0.3048 ∗ 539.208 ft.

Sample standard deviation = 46.9283 meters = $k * s_x$ = 0.3048 ∗ 153.964 ft.

All the plots will look the same in each unit, but with a different scale. Screen 23 shows a modified boxplot using **ZoomData** to fit the plot to the screen. This screen resembles Topic 7, screen 14, but now **maxX = 945 ft = 288** meters. A building 1000 feet tall is 304.8 meters — 305 does not sound as grand as 1000!

(23)


Note: *To get the modified boxplot, repeat steps 4 and 6 in Topic 7 with* X List: **list1** *instead of* **phily** *(screen 23).*