Math 221: Describing Distributions with Numbers

S. K. Hyde
Chapter 3 (Gould & Ryan)

2002 Texas Ranger's Salaries

Player	Salary	Ordered Stats
Reynaldo Garcia	$ 300,000	$x_{(1)}$
Jermaine Clark	300,000	$x_{(2)}$
Colby Lewis	302,500	$x_{(3)}$
Hank Blalock	302,500	$x_{(4)}$
Kevin Mench	327,500	$x_{(5)}$
Michael Young	415,000	$x_{(6)}$
Mike Lamb	440,000	$x_{(7)}$
C.J. Nitkowski	550,000	$x_{(8)}$
Ruben Sierra	600,000	$x_{(9)}$
Aaron Fultz	600,000	$x_{(10)}$
Chad Kreuter	750,000	$x_{(11)}$
Todd Greene	750,000	$x_{(12)}$
Francisco Cordero	900,000	$x_{(13)}$
Doug Glanville	1,000,000	$x_{(14)}$
John Thomson	1,300,000	$x_{(15)}$
Herbert Perry	1,300,000	$x_{(16)}$
Esteban Yan	1,500,000	$x_{(17)}$
Einar Diaz	1,837,500	$x_{(18)}$
Mark Teixeira	1,875,000	$x_{(19)}$
Todd Van Poppel	2,500,000	$x_{(20)}$
Ismael Valdes	2,500,000	$x_{(21)}$
Jay Powell	3,250,000	$x_{(22)}$
Jeff Zimmerman	3,366,667	$x_{(23)}$
Ugueth Urbina	4,500,000	$x_{(24)}$
Rusty Greer	7,000,000	$x_{(25)}$
Rafael Palmeiro	9,000,000	$x_{(26)}$
Carl Everett	9,150,000	$x_{(27)}$
Chan Ho Park	12,884,803	$x_{(28)}$
Juan Gonzalez	13,025,000	$x_{(29)}$
Alex Rodriguez	22,000,000	$x_{(30)}$

Measures of Center

1. Mean

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum_{k=1}^{n} x_i \\ &= \frac{1}{n}(300{,}000 + \cdots + 22{,}000{,}000) \\ &= \frac{1}{n}(\$104{,}526{,}470) = \$3{,}484{,}215.67\end{aligned}$$

2. Median

Choose the middle score. Since $n = 30$, then find the average of the 15th and the 16th salaries.

$$\begin{aligned}M &= \frac{x_{(15)} + x_{(16)}}{2} = \frac{\$1{,}300{,}000 + \$1{,}300{,}000}{2} \\ &= \$1{,}300{,}000\end{aligned}$$

3. Mode

The mode is defined as the value that appears most often. It is most appropriate to use when you have categorical data. If there is only one value that appears the most, then the distribution is said to be unimodal. If two values appear, then it is bimodal. If three values appear, then it is trimodal. If more than four values appear, then the distribution has no mode. The value that appears the most for the Rangers data is $300,000, $302,500, $600,000, $750,000, $1,300,000, and $2,500,000. Hence, this set of data has no mode.

4. Midrange

The midrange is the average of the min and the max values. So

$$\begin{aligned}\text{Midrange} &= \frac{x_{(\min)} + x_{(\max)}}{2} \\ &= \frac{\$300{,}000 + \$22{,}000{,}000}{2}\end{aligned}$$

Measures of Variation

1. Range

The range is the difference between the maximum and minimum values. Hence, it is

$$\begin{aligned}R &= x_{(\max)} - x_{(\min)} \\ &= \$22{,}000{,}000 - \$300{,}000 \\ &= \$21{,}700{,}000\end{aligned}$$

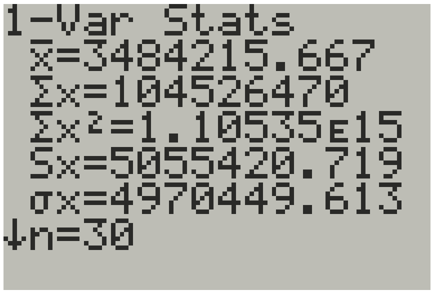
2. Standard Deviation To compute the standard deviation, use a calculator. When using the TI-83 or the TI-84, you first need to edit the data. To do this, press

$$\boxed{\texttt{STAT}} \longrightarrow \boxed{\texttt{EDIT}} \longrightarrow \boxed{\texttt{Edit...}}.$$

Enter your data into the `L1` column. Next, calulate the one variable statistics by pressing

$$\boxed{\texttt{STAT}} \longrightarrow \boxed{\texttt{CALC}} \longrightarrow \boxed{\texttt{1-Var Stats}},$$

and finally press `Enter`. The standard deviation is the variable `Sx` found at the top of the `1-Var Stats` screen (Figure 1). For the Rangers data, the standard deviation is $5,055,420.72.



```
1-Var Stats
 x̄=3484215.667
 Σx=104526470
 Σx²=1.10535E15
 Sx=5055420.719
 σx=4970449.613
↓n=30
```

Figure 1: 1-Var Stats (top)

3. Variance

The variance is the square of the standard deviation. First, find the standard deviation using the previous section. Then the variance can be found by pressing

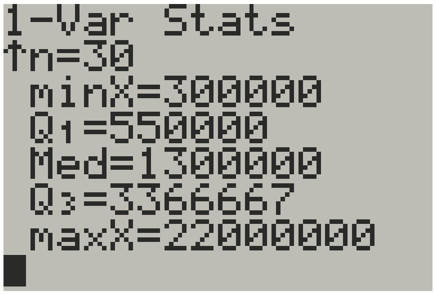
$$\boxed{\texttt{VARS}} \longrightarrow \boxed{\texttt{5}} \longrightarrow \boxed{\texttt{3}} \longrightarrow \boxed{x^2}.$$

For the Ranger's data set, the variance is 2.555×10^{13}.

4. IQR

The IQR is the difference between the first and third quartile, which are computed in the next section. You can also find them using the calculator. From the bottom of the `1-Var Stats` screen (Figure 2), the first and third quartile can be found. Thus, the IQR is

$$\begin{aligned} \text{IQR} &= Q_3 - Q_1 \\ &= 3{,}366{,}667 - 550{,}000 \\ &= 2{,}816{,}667 \end{aligned}$$



```
1-Var Stats
↑n=30
 minX=300000
 Q₁=550000
 Med=1300000
 Q₃=3366667
 maxX=22000000
```

Figure 2: 1-Var Stats (bottom)

Measures of Relative Standing

1. First Quartile (Q_1)

To find the first quartile, you find the median of the lower half of your data (excluding the median). Since the median is the average of the 15th and 16th, then find the median of the first fifteen values. This will be the 8th element. It follows that the first quartile is

$$Q_1 = x_8 = \$550{,}000$$

2. Midquartile (median)

To find the Midquartile or the median, follow what was done on the previous page.

3. Third Quartile (Q_3)

To find the third quartile, you find the median of the upper half of your data (excluding the median). Since the median is the average of the 15th and 16th, then find the median of the last fifteen values. This will be the 23th element. It follows that the first quartile is

$$Q_1 = x_{23} = \$3{,}366{,}667$$

4. Z-Score

The z-score for a data value is defined as

$$z = \frac{x - \text{mean}}{\text{standard deviation}} = \frac{x - \bar{x}}{s}$$

For the largest data value (Alex Rodriguez's salary), the z score is

$$z = \frac{\$22{,}000{,}000 - \$3{,}484{,}215.6667}{\$5{,}055{,}420.72} = 3.6626$$

5. Five number summary

   The five number summary is the five numbers:

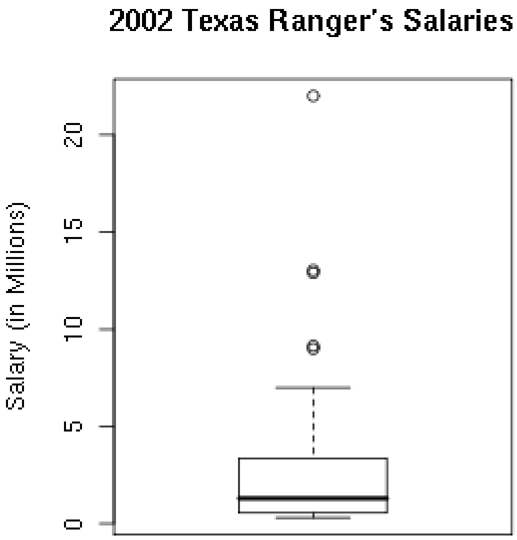
   $$\min, Q_1, \text{Median}, Q_3, \max.$$

   For this data set, the five number summary is

min	\$300,000
Q_1	\$550,000
M	\$1,300,000
Q_3	\$3,366,667
max	\$22,000,000.

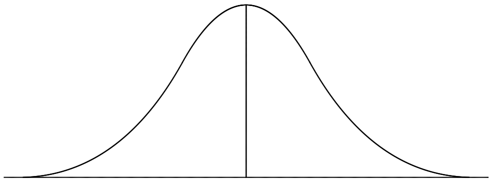
6. Boxplot

   The boxplot is a graphical plot of the five number summary. The five number summary is plotted with the middle three being connected in a box, and the remaining having lines attaching to the box as follows:

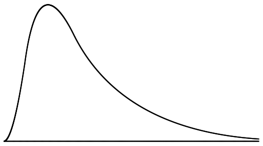


7. Bell Shaped, Skewed and Uniform distributions

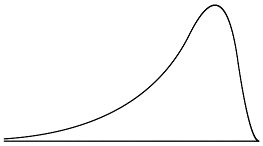
   A distribution that has most of its observations surrounding its mean, and which the farther you are from the mean, the smaller the amount of data, but equally in both directions, is called a symmetric distribution. The distribution is bell-shaped and looks like:



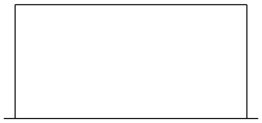
Sometimes the distribution is either skewed to the left or skewed to the right. A distribution that is skewed to the right, it has a larger percentage of values that are substantially larger than most of the data. In a skewed right distribution, the mean is larger than the median. For example, a distribution which is skewed to the right will look like



When a distribution is skewed to the left, then there is a large percentage of values that are substantially less than most of the data. In a skewed left distribution, the mean is less than the median. When this occurs, the distribution looks like



When the distribution of the values is the same over the entire range of data, then the distribution is uniform. This occurs when the same percentage of values occurs over all intervals of fixed length. The distribution look like



When a distribution is bell shaped and symmetric, the mean and the standard deviation are chosen to represent the distribution. They do a very good job then, because the mean, median, and mode are the same in bell shaped distributions.

When a distribution is skewed, the mean does not represent the distribution well. In this case, the five number summary is chosen to represent the distribution.