

CHAPTER 5

Describing Distributions Numerically

Describing the Distribution

- Center
 - Median
- Spread
 - Range
 - Interquartile Range

Median

- The middle number of a data set
- Different calculation when
 - n is odd
 - n is even
 - n is the number of observations

Median - n is odd

- Order the data from smallest to largest
- The median is the middle number on this list
- $(n+1)/2$ number from the bottom
 - Ex. If $n=17$, then median is the $(17+1)/2=9$ th number from the bottom
 - Ex. If $n=25$, then median is the $(25+1)/2=13$ th number from the bottom

Barry Bonds' HR

- Order the $n = 19$ values
- 16, 19, 24, 25, 25, 33, 33, 34, 34, 37, 37, 40, 42, 45, 45, 46, 46, 49, 73
- Median = $(19+1)/2 = 10$ th number on the list => 37
- So Barry Bonds' median HR total is 37

Median - n is even

- Again, order the data from smallest to largest
- Median is the average of the middle two numbers
- $(n+1)/2$ is halfway between these two numbers
- Ex. If $n=14$, $(14+1)/2=7.5$, so median is the average of the 7th and 8th numbers from the bottom

Sammy Sosa's HR

- Order the $n=16$ values
- 4, 8, 10, 15, 25, 33, 35, 36, 36, 40, 40, 49, 50, 63, 64, 66
- Median = $(16+1)/2 = 8.5$ number on the list, the average of the 8th and 9th numbers on the list from the bottom
- Median = $(36+36)/2 = 36$
- So Sammy Sosa's median HR total is 36

Spread

- Range
 - Find minimum value
 - Find maximum value
- Range = maximum - minimum

Examples

- Sammy Sosa
 - Min = 4, Max = 66, Range = $66 - 4 = 62$
- Barry Bonds
 - Min = 16, Max = 73, Range = $73 - 16 = 57$
- The range is larger for Sammy Sosa than for Barry Bonds

Range

- Very basic measure of spread
 - Range is highly affected by outliers.
 - Makes spread appear larger than reality.
 - Ex. Barry Bonds has an outlier at 73
 - Range without outlier = 33
 - Range with outlier = 57

Spread

- Interquartile Range (IQR)
 - First Quartile (Q1)
 - Larger than 25% of the data
 - Third Quartile (Q3)
 - Larger than 75% of the data.
- $IQR = Q3 - Q1$
 - Center 50% of the values

Finding Quartiles

- The method in the book is different than most computer programs, including your calculators
 - the answers will very close
 - the book's method is easier

Finding Quartiles

- Order the data
- Split the data into two halves at the median
 - When n is odd, include the median in both halves
- Q1 = median of the lower half
- Q3 = median of the upper half

Sammy Sosa

- Order the n=16 values
- 4, 8, 10, 15, 25, 33, 35, 36, 36, 40, 40, 49, 50, 63, 64, 66
- Lower half = 4, 8, 10, 15, 25, 33, 35, 36
 - Q1 = median of lower half = $(15+25)/2 = 20$
- Upper half = 36, 40, 40, 49, 50, 63, 64, 66
 - Q3 = median of upper half = $(49+50)/2 = 49.5$
- IQR = Q3 - Q1 = $49.5 - 20 = 29.5$

Barry Bonds

- Order the n=19 values
- 16, 19, 24, 25, 25, 33, 33, 34, 34, 37, 37, 40, 42, 45, 45, 46, 46, 49, 73
- Lower half = 16 19 24 25 25 33 33 34 34 37
 - Q1 is the median of the lower half = 29
- Upper half = 37 37 40 42 45 45 46 46 49 73
 - Q3 is the median of the upper half = 45
- IQR = Q3 - Q1 = $45 - 29 = 16$

Five Number Summary

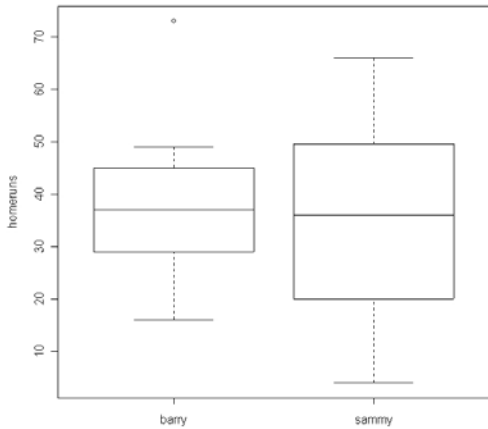
- Minimum
- Q1
- Median
- Q3
- Maximum

Examples

- | | |
|---------------|---------------|
| • Sammy Sosa | • Barry Bonds |
| – Min = 4 | – Min = 16 |
| – Q1 = 20 | – Q1 = 29 |
| – Median = 36 | – Median = 37 |
| – Q3 = 49.5 | – Q3 = 45 |
| – Max = 66 | – Max = 73 |

Sammy vs. Barry

- Boxplot of Sammy Sosa HRs.
 - Box from 20 to 49.5
 - Line in box at 36.
 - Lines extend out from box to 4 and 66.
- Boxplot of Barry Bonds HRs.
 - Box from 29 to 45
 - Line in box at 37
 - Lines extend out from box to 16 and 49



Mean

- Mean is the ordinary average
 - add up all the observations
 - then divide by the number of observations

$$\bar{y} = \frac{\sum y}{n} = \frac{y_1 + y_2 + y_3 + \dots + y_n}{n}$$

Examples

- Sammy Sosa

$$\frac{(4 + 8 + 10 + 15 + \dots + 66)}{16} = 35.875$$

- Barry Bonds

$$\frac{(16 + 19 + 24 + 25 + \dots + 73)}{19} = 37$$

Mean vs. Median

- Median is the middle number
- Mean is the point where the histogram balances

Mean vs. Median

- Mean and Median similar when
 - Data are symmetric
- Mean and median different when either
 - Data are skewed
 - Outliers are present

Mean vs. Median (Outliers)

- Mean influenced by unusually high or unusually low values.
 - Example: Income in a small town of 6 people.

\$25,000	\$27,000	\$29,000
\$35,000	\$37,000	\$38,000

**The mean income is \$31,830

**The median income is \$32,000

Mean vs. Median

- Bill Gates moves to town
\$25,000 \$27,000 \$29,000
\$35,000 \$37,000 \$38,000 \$40,000,000
- **The mean income is \$5,741,571
- **The median income is \$35,000
- Outlier pulls the mean towards it
- Median is not
- Mean is not a good center of these data

Mean vs. Median (Skewness)

- Mean is pulled in direction of tail.
 - Skewed to the right = mean > median
 - Skewed to the left = mean < median

Mean vs. Median

- Always question when means are reported for skewed data
 - Income
 - Housing prices
 - Course grades
- Do not automatically take means at their face value

Spread

- Standard deviation
 - “Average” spread from mean.
 - Most common measure of spread
 - Denoted by letter s

Standard Deviation

$$s = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}$$

$$s = \sqrt{\frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \dots + (y_n - \bar{y})^2}{n-1}}$$

Standard Deviation

- Usually calculate using computer or calculator
 - Choose n-1 option on calculator
- By hand
 - Make a table

Sammy Sosa

4	-31.93	1019.5249
8	-27.93	780.0849
10	-25.93	672.3649
15	-20.93	438.0649
25	-10.93	119.4649
33	-2.93	8.5849
36	0.07	0.0049
36	0.07	0.0049
40	4.07	16.5649
40	4.07	16.5649
49	13.07	170.8249
50	14.07	197.9649
63	27.07	732.7849
64	28.07	787.9249
66	30.07	904.2049
		5864.9335

Sammy Sosa

$$s = \sqrt{\frac{5864.9335}{14}} = 20.47$$

Properties of s

- $s = 0$ only when all observations are equal.
Otherwise, $s > 0$
- s has the same units as the data
- s is not resistant
 - Skewness and outliers affect s, just like mean

Which summaries are the best?

- Five Number Summary
 - Skewed Data
 - Data with outliers
- Mean and Standard Deviation
 - Symmetric Data
- ALWAYS PLOT YOUR DATA