Describing Distributions with Numbers

- **Using graphs, we could determine the *center*, *spread*, and *shape* of the distribution of a quantitative variable.**
- **We can also use numbers (called *summary statistics*) to describe the *center*, *spread*, and *shape* of the distribution of a quantitative variable.**
- ***Example 1:* Barry Bonds vs. Hank Aaron (baseball home run kings)**
- **Consider two data sets: Yearly home run counts for Barry Bonds (1986-2007) and yearly home run counts for Hank Aaron (1954-1976).**
- **How can we characterize the *center*, *spread*, and *shape* of these two data sets?**

Histograms of Home Run Data

- **What is the rough *midpoint* of Bonds' home run distribution?**
- **How can we describe the *spread* of Bonds' home run distribution?**
- **How do these compare to the corresponding characteristics of Aaron's distribution?**
- **Now let's examine some more precise numerical measures of center and spread.**

Histogram of Bonds yearly home run values:

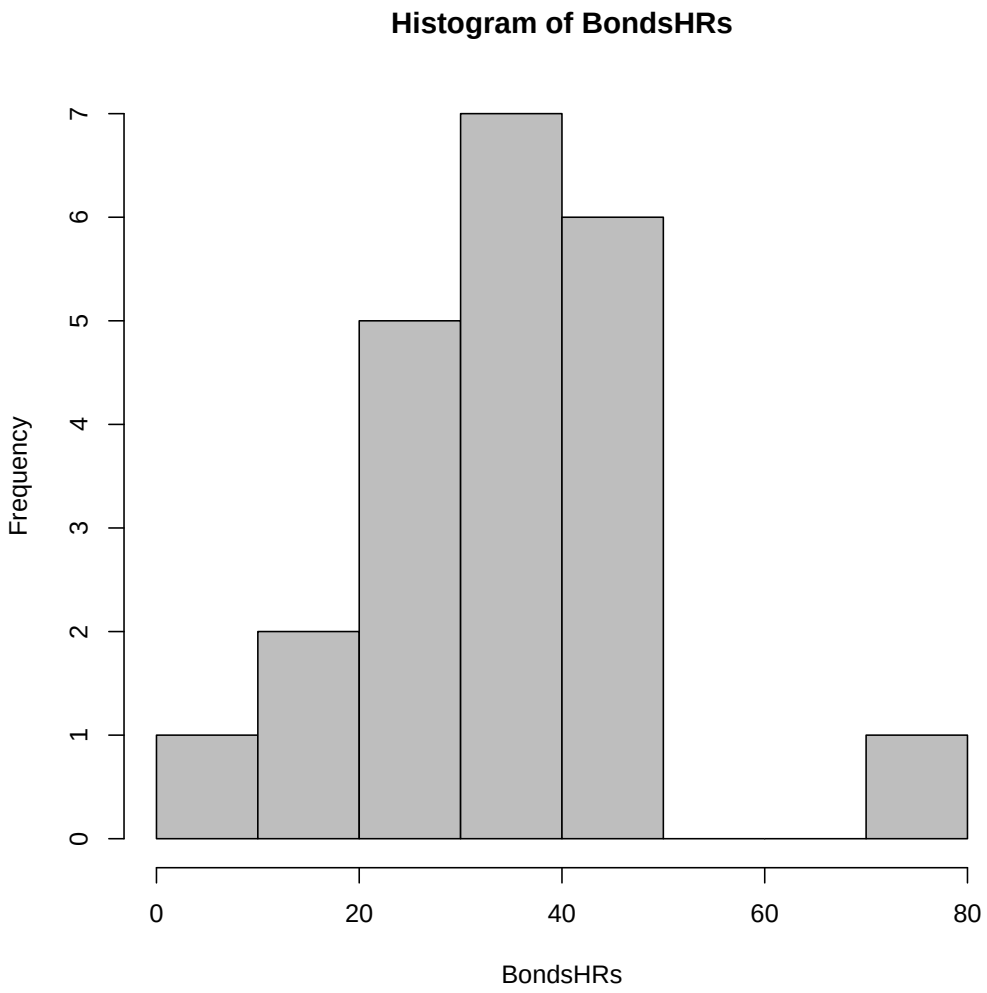


Figure 1: Barry Bonds' yearly home run totals (1986-2007).

Histogram of Aaron yearly home run values:

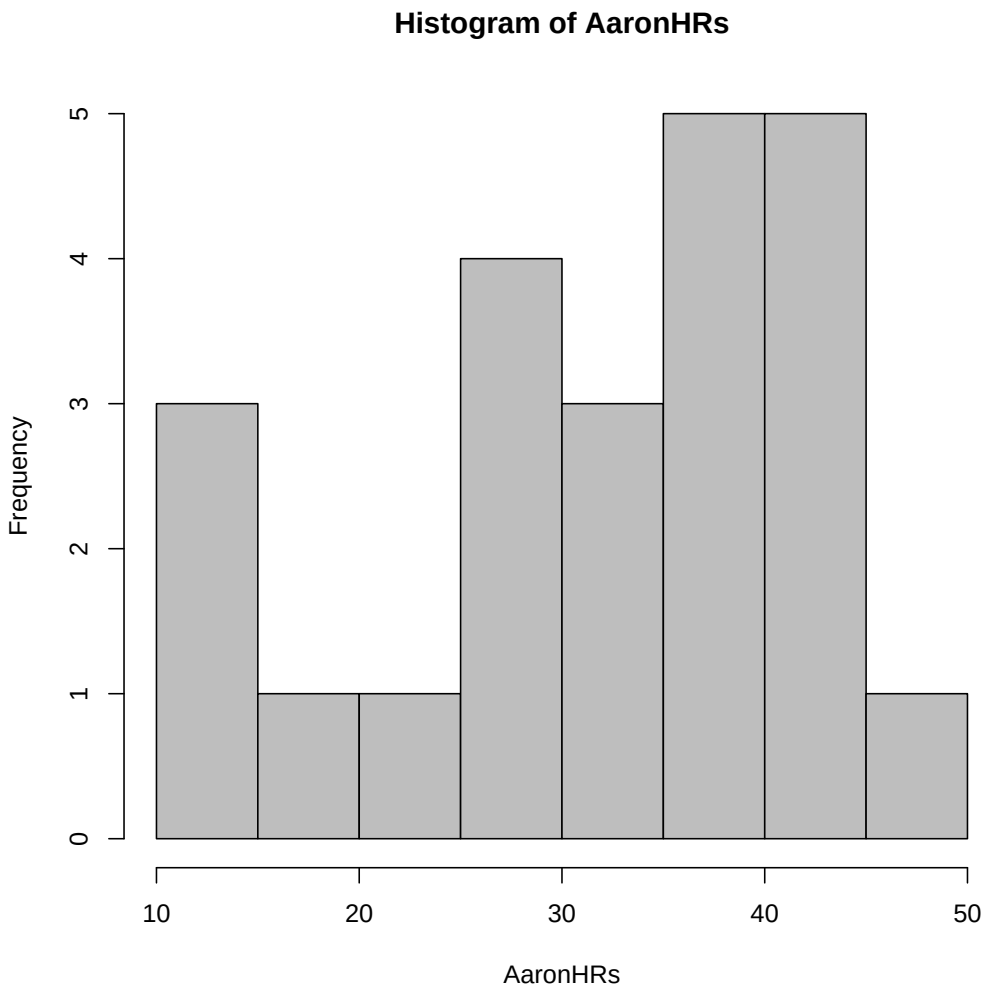


Figure 2: Hank Aaron's yearly home run totals (1954-1976).

The Median: A Measure of Center

- **The *median* (denoted M) of a data set is a numerical measure of the midpoint.**
- **The number of data values *less than* M is always the same as the number of data values *greater than* M.**
- **To find the median of a data set, we first order the data from smallest to largest.**
- **Example: Bonds data (ordered from smallest to largest):**

```
 5 16 19 24 25 25 26 28 33 33 34 34 37 37 40
42 45 45 46 46 49 73
```

Finding The Median (Continued)

- **Once the data are ordered, use the "n-plus-1 over 2" rule!**
- **Find $\frac{n+1}{2}$ (Remember n = the overall number of data values).**
- **This will tell you the *position* of the median in the ordered data set.**
- **If n is odd, then the median is an actual data value (the one in position $\frac{n+1}{2}$ in the ordered data set).**
- **If n is even, then $\frac{n+1}{2}$ is not a whole number, and so the median will be halfway between two data values.**

Clicker Quiz 1

In the Aaron data set, there are 23 observations. Using the "n-plus-1 over 2" rule, what is the position of the median in the ordered data set?

A. Halfway between 11th and 12th

B. Halfway between 12th and 13th

C. 12th

D. 23rd

Clicker Quiz 2

In the Bonds data set, there are 22 observations. Using the "n-plus-1 over 2" rule, what is the position of the median in the ordered data set?

A. 11th

B. Halfway between 11th and 12th

C. 12th

D. 22nd

Examples of Finding Medians

Aaron data set (in order):

10 12 13 20 24 26 27 29 30 32 34 34 38 39 39 40 40 44 44 44 44 45 47

- **Since $n = 23$, $\frac{n+1}{2} = 12$.**
- **The 12th value in the ordered data set is 34, so the median of this data set is 34.**

Examples of Finding Medians (Continued)

Bonds data set (in order):

5 16 19 24 25 25 26 28 33 33 34 34 37 37 40
42 45 45 46 46 49 73

- **Since $n = 22$, $\frac{n+1}{2} = 11.5$.**
- **The 11th value in the ordered data set is 34, and the 12th value in the ordered data set is also 34!**
- **So the median of this data set is 34 (halfway between 34 and 34).**

Examples of Finding Medians (Continued Once More)

Babe Ruth data set (in order):

0 2 3 4 6 11 22 25 29 34 35 41 41 46 46 46 47 49 54 54 59 60

- **Since again $n = 22$, $\frac{n+1}{2} = 11.5$.**
- **The 11th value in the ordered data set is 35, and the 12th value in the ordered data set is 41**
- **So the median of the Ruth data set is 38 (halfway between 35 and 41).**
- **For Aaron, M = 34. For Bonds, M = 34. For Ruth, M = 38. Conclusions?**

Quartiles

- **The median is essentially a number that divides the data set into *halves*.**
- **The quartiles (denoted Q_1, Q_2, Q_3) are numbers that divide the data set into *quarters*.**
- **Q_2 is simply the median.**
- **Q_1 is the median of all the observations that are to the *left of the position of the overall median* M in the ordered data set.**
- **Q_3 is the median of all the observations that are to the *right of the position of the overall median* M in the ordered data set.**

Examples of Finding Quartiles

Aaron data set (in order):

10 12 13 20 24 26 27 29 30 32 34 34 38 39 39 40 40 44 44 44 44 45 47

- **Since $n = 23$, $\frac{n+1}{2} = 12$.**
- **The observations to the left of the 12th value in the ordered data set are simply the first 11 values.**
- **The median of these first 11 values is 26 (check it!), so Q_1 = 26.**
- **The observations to the right of the 12th value in the ordered data set are simply the *last* 11 values.**
- **The median of these *last* 11 values is 44 (check it!), so Q_3 = 44.**

Examples of Finding Medians (Continued)

Bonds data set (in order):

```
  5 16 19 24 25 25 26 28 33 33 34 34 37 37 40
42 45 45 46 46 49 73
```

- **The observations to the left of the 11.5 *position* in the ordered data set are simply the first 11 values.**
- **The median of these first 11 values is 25 (check it!), so Q_1 = 25.**
- **The observations to the right of the 11.5 *position* in the ordered data set are simply the *last* 11 values.**
- **The median of these *last* 11 values is 45 (check it!), so Q_3 = 45.**

The Five-Number Summary

- **A lot of information about a distribution can be summarized in the *5-number summary*.**
- **This *5-number summary* consists of: The minimum value; Q_1; the median; Q_3; and the maximum value.**
- **Summarizes information about the center, the spread, and the tails of the distribution.**
- **The median describes the center of the distribution.**
- **The distance between Q_1 and Q_3 describes the spread of the middle 50% of the data.**
- **The minimum and maximum give information about the "tails" and possible outliers.**

Clicker Quiz 3

The 5-number summary for the Aaron data set is:

10 26 34 44 47.

The 5-number summary for the Bonds data set is:

5 25 34 45 73.

How could we accurately compare the two distributions?

A. They have very different centers and spreads.

B. The centers are the same, but the Aaron data set is somewhat more spread out.

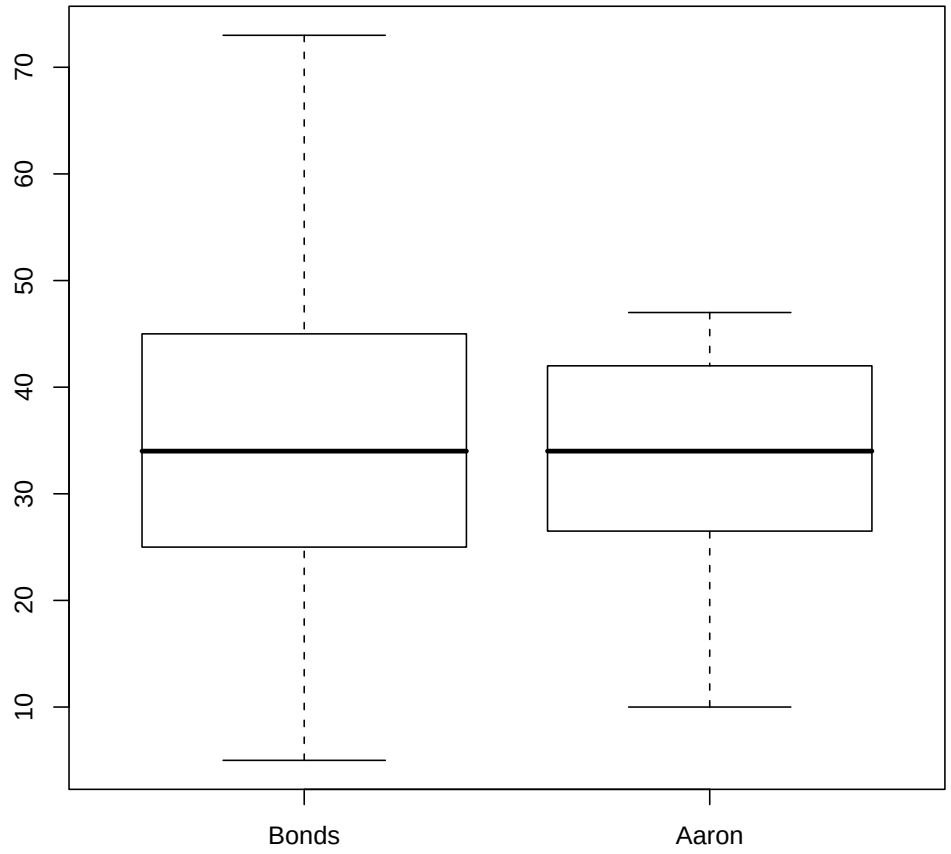
C. The centers are the same, but the Bonds data set is somewhat more spread out.

D. The Aaron data set seems to have more outlying values.

Boxplots

- **A *boxplot* is a graphical presentation of the *5-number summary*.**
- **The minimum value; Q_1; the median; Q_3; and the maximum value are plotted on one axis:**
- **A box is drawn whose ends range from Q_1 to Q_3.**
- **A line is drawn inside the box where the median is located.**
- **Lines extend outside the box to the smallest and largest values in the data set.**
- **Often multiple boxplots are placed in the same graph (using same axes) to compare multiple distributions (Bonds / Aaron example)**

Boxplots of Bonds and Aaron yearly home run values:



A Measure of Spread: Interquartile Range

- **Recall the distance between Q_1 and Q_3 describes the spread of the middle 50% of the data.**
- **This distance is simply $Q_3 - Q_1$, which we call the Interquartile Range (IQR) of the data set.**
- **The IQR is a numerical measure of the spread in a data set.**
- **Note that the IQR is simply the length of the "box" part of a boxplot.**

Clicker Quiz 4

The 5-number summary for the Aaron data set is:

10 26 34 44 47.

The 5-number summary for the Bonds data set is:

5 25 34 45 73. **What is the IQR of the Aaron data set?**

A. 37

B. 20

C. 34

D. 18

Clicker Quiz 5

The 5-number summary for the Aaron data set is:

10 26 34 44 47.

The 5-number summary for the Bonds data set is:

5 25 34 45 73. **What is the IQR of the Bonds data set?**

A. 68

B. 20

C. 34

D. 18

More on Boxplots

- **Some computer packages produce boxplots whose extra lines (or "whiskers") don't necessarily extend all the way to the minimum or maximum values.**
- **They may only extend to the "non-outlying values" and the outliers may be marked with separate symbols on the plot.**
- **Typically an observation is labeled an outlier if it lies more than** $1.5 \times IQR$ **above** $\boldsymbol{Q_3}$ **or below** $\boldsymbol{Q_1}$**.**
- **Example for Bonds data set:** $1.5 \times IQR$ **= ?**

Even More on Boxplots

- **A boxplot can indicate whether a distribution is symmetric or skewed.**
- **Does one half of the box extend farther out than the other half? Does one half of the overall boxplot extend farther out than the other half?**
- **It's typically easier to determine shape and symmetry/skewness using a histogram or a stemplot than a boxplot, though.**
- **Boxplots are good for quick summaries and comparisons of distributions.**

Other Measures of Center and Spread: Mean, Variance and Standard Deviation

- **The *mean* of the data set (denoted $\bar{x}$) is simply the sum of the observations divided by the total number of observations.**
- **Like the median, the mean is a measure of a data set's *center*.**
- **The *standard deviation* of a data set (denoted *s*) measures roughly how far each observation is from the mean, *on average*.**
- **Like the IQR, the standard deviation is a measure of a data set's *spread*.**

Mean, Variance and Standard Deviation (Continued)

- **If *s* is large, then the data set is very spread out. If *s* = 0, then all data values are the same (absolutely no spread!).**
- **The *variance* is the square of the standard deviation.**
- **Page 271 in book describes how to find the variance and standard deviation by hand (we won't do this).**

Median or Mean?

- **Which is a better measure of center, the median or the mean?**
- **An advantage of the mean: It uses the actual values of all the observations.**
- **A disadvantage of the mean: It is more affected by *outliers*.**
- **One outlying value can greatly affect the mean.**
- **An outlier won't affect the median as much – the median is more *robust* to outliers.**
- **Similarly, the IQR is a more robust measure of spread than the standard deviation.**

Median or Mean? (Example)

- **Example: 2014-15 New York Knicks salaries (in millions):** `0.1 0.3 0.5 0.5 0.5 0.7 0.9 1.0 1.3 1.6 3.2 3.3 7.1 12.0 22.5`
- **The mean salary is $3.7 million; the median salary is $1.0 million.**
- **Which is a better reflection of a "typical player's" salary?**
- **The mean is greatly affected by the outlier(s) at the "high end" of the salary values.**

Median or Mean? (More)

- **Usual rule: For skewed distributions or distributions with outliers, use median as a measure of center. (Examples: income data, house price data)**
- **For symmetric distributions, the mean and standard deviation are reasonable measures.**
- **NOTE: It's always good to look at a graph of the data, not just to rely on numerical measures alone!**