

This handout introduces three numerical summaries for quantitative variables as well as one more graphical display, the boxplot. The boxplot is included in this handout because it is based on one of the numerical summaries, the five-number summary.

Numerical summaries for quantitative variables

- 5-number summary
- median and interquartile range (IQR)
- mean and standard deviation

Percentiles and quantiles

For any number p between 0 and 1, the $100p^{\text{th}}$ percentile is the value such that $100p$ percent of the data values are less than or equal to that value. For example, the 90^{th} percentile is the value such that 90 percent of the data values are less than or equal to that value. The $100p^{\text{th}}$ percentile is the same as the p quantile. Thus the 90^{th} percentile is the same as the 0.9 quantile.

The 50^{th} percentile of a data set is the median, denoted M (half the values below, half above). The quartiles are the 25^{th} , 50^{th} and 75^{th} percentiles, denoted Q_1 , M , and Q_3 . They divide the distribution into quarters. Q_1 is sometimes called the first quartile or the lower quartile and Q_3 the third quartile or the upper quartile.

Calculating percentiles from a set of data usually requires some interpolation. For example, there usually isn't a value such that exactly 90 percent of the data are less than or equal to that value. Different software programs may use slightly different methods that may result in different values for percentiles. The differences are usually minor, however.

In R:

```
> quantile(birthwt$bwt,.9) # birthweight data
 90%
3864.8
```

5-number summary

The 5-number summary consists of the minimum, the first quartile (25^{th} percentile), the median, the third quartile (75^{th} percentile), and the maximum, in order; that is, min, Q_1 , M , Q_3 , max. The 5-number summary divides the data approximately into quarters. The most common use of the 5-number summary is as the basis for creating a boxplot, discussed later on in this handout, a handy graphical tool for comparing two or more distributions.

What do the 5-number summaries below tell you about the shapes of these distributions? Are they as informative as a histogram?

```
> fivenum(birthwt$bwt)
[1] 709 2414 2977 3487 4990
```

```
> fivenum(geyser$waiting)
[1] 43 59 76 83 108
```

```
> fivenum(MLB$Salary)/1000000
[1] 0.4000 0.4194 1.1510 4.2500 33.0000
```

Computing the 5-number summary by hand

The median is the value which divides the ordered data values in half. A general formula for the position of the median is $(n+1)/2$. Example: $n = 5$ gives 3 as the position of the median (the 3rd ordered value); $n = 6$ gives 3.5 which means halfway between the 3rd and 4th ordered values (the average of the two middle values).

While the method for computing the median is standard, there are several different algorithms for finding the first and third quartiles by hand. The one described here is the one R uses in its `fivenum` command. Some books give a slightly different algorithm.

To find Q_1 and Q_3 , first order the observations from smallest to largest and find the position of the median. Q_1 is then the median of all observations at or below the position of the median (essentially, the median of the lower half of the data values) and Q_3 is the median of all values at or above the position of the median (the median of the upper half of the values). I have underlined the word position because it does not matter if there are tied values; only values to the left (or right) of the position of the median are counted.

Example: Barry Bonds' and Alex Rodriguez' yearly home run counts (complete seasons only):

Bonds	ARod
1 69	1
2 45568	2 3
3 334477	3 05566
4 0255669	4 12278
5	5 247
6	6
7 3	7

Bonds, $n = 21$:

- median position is $(21+1)/2 = 11$, 11th ordered value. Hence, $M = 34$.
- Q_1 is median of the 11 values is at or below the position of M ; hence Q_1 is 6th value; $Q_1 = 26$.
- Q_3 is median of the 11 values at or above the position of M ; hence Q_3 is 6th value up starting at M (or 6th value down starting from the top); $Q_3 = 45$.
-

So the 5-number summary is 16, 26, 34, 45, 73

ARod: $n = 14$:

2-number summaries

Often a more compact numerical summary of a set of data values is used, consisting of a measure of center and a measure of spread (or variability). We'll look at the two most common "2-number" summaries ("2-number summary" is not a standard term like 5-number summary).

Median and IQR

One 2-number summary consists of the median as a measure of center and the interquartile range (IQR) as a measure of spread, where $IQR = Q_3 - Q_1$. IQR is the range of the middle half of the data. Note: IQR is a single number, the difference between the first and third quartiles.

```
> bonds =  
c(16, 25, 24, 19, 33, 25, 34, 46, 37, 33, 42, 40, 37, 34, 49, 73, 46, 45, 45, 26, 28)  
> median(bonds)  
[1] 34  
> IQR(bonds)  
[1] 19
```

Mean and standard deviation

The most common numerical summary of a distribution is the mean (a measure of center) and the standard deviation (a measure of spread).

If we denote the data values as x_1, x_2, \dots, x_n , then the mean is denoted by \bar{x} ("x-bar") and is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

The standard deviation is

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

The square of the standard deviation, s^2 , is called the variance. The standard deviation is roughly the average distance of the data values from the mean. I say "roughly" because it is actually the square root of almost the average squared distance of the data values from the mean. Taking the square root puts it back in the original units.

Why not simply take the average distance to the mean, $\frac{1}{n} \sum |x_i - \bar{x}|$? This is a legitimate measure of spread, but is not commonly used because the standard deviation has some nice properties which we'll discover later.

Note: later in the course, we'll talk about the mean and standard deviation of random variables. To distinguish these from the mean and standard deviation of a set of data values, we sometimes refer to the mean and standard deviation of a set of data values as the sample mean and sample standard deviation (with the idea that the data often come from a sample from a population).

```
> mean(bonds)  
[1] 36.04762  
> sd(bonds)  
[1] 12.68651
```

Comparison of Bonds, Sosa, and ARod

	<i>n</i>	Median	IQR	Mean	SD
Bonds	21	34.0	19.0	36.0	12.7
ARod	14	41.5	12.5	41.3	9.6

Relationship between mean and median

What's the relationship between the mean and median for the following distribution shapes?

Symmetric

Skewed to the right

Skewed to the left

Examples:

Birthweights:

MLB salaries:

Resistance: A measure is said to be resistant if it is not much affected by changes in the numerical values of a small proportion of the observations at either extreme of the distribution. For example, if the smallest value were made much smaller or the largest value were made much larger, would the measure change much? (Is it resistant to outliers?)

Is the median a resistant measure of center? How about the mean?

Is the IQR a resistant measure of spread? How about the standard deviation?

Summarizing a distribution with a measure of center and spread – which measures should you use? Since the mean and standard deviation are not resistant, they are not appropriate for skewed distributions or distributions with outliers. They're most appropriate for symmetric distributions with no outliers.

- Symmetric distribution with no outliers: mean and standard deviation (possibly, median and IQR also)
- Skewed distributions: median and IQR
- Symmetric distributions with outliers: median and IQR or mean and standard deviation with and without outliers

Why use the mean and standard deviation at all if the median and IQR are always appropriate? Because the mean and standard deviation have a nice interpretation if the data are approximately normally distributed with no outliers. This will be discussed in the next handout.

Boxplots

A boxplot is a graphical display of a 5-number summary. Its primary use is to compare two or more distributions since a histogram gives a more informative plot for a single variable.

The ends of the central “box” are the first and third quartiles and the line within the box is at the median (see Bonds’ boxplot below). The “whiskers” extend to the most extreme values that are not identified as “outliers.” For the Bonds data, no outliers were identified by the boxplot algorithm, so the whiskers extend to the minimum and maximum data values.

Any point that is greater than $Q_3 + 1.5 \text{ IQR}$ or less than $Q_1 - 1.5 \text{ IQR}$ is considered an outlier for the purposes of a boxplot. If there is an outlier on either end, then the whisker extends to the most extreme point that is not an outlier and the outliers are plotted individually.

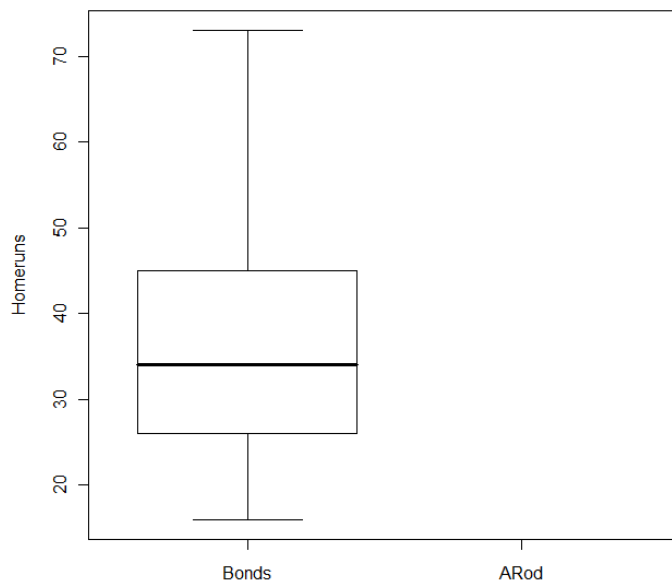
For Bonds:

$$Q_3 + 1.5 \text{ IQR} = 45 + 1.5(19) = 73.5$$

$$Q_1 - 1.5 \text{ IQR} = 25 - 1.5(19) = -3.5$$

Since none of Bonds’ counts are outside these boundaries, there are no outliers by this definition and the whiskers extend to the minimum and maximum values. If Bonds highest season had been 75, say, instead of 73, then 75 would have been an outlier and the whisker on the upper end would have extended to 49, the highest value that’s not an outlier. The 75 would have been plotted as an individual point.

Add ARod’s boxplot next to Bonds’ below. Be sure to check for outliers.



In R, the above boxplot can be obtained by:

```
bonds =  
c(16, 25, 24, 19, 33, 25, 34, 46, 37, 33, 42, 40, 37, 34, 49, 73, 46, 45, 45, 26, 28)  
arod = c(36, 23, 42, 42, 41, 52, 57, 47, 36, 48, 35, 54, 35, 30)  
boxplot(bonds, arod, names=c("Bonds", "ARod"), ylab="Homeruns")
```

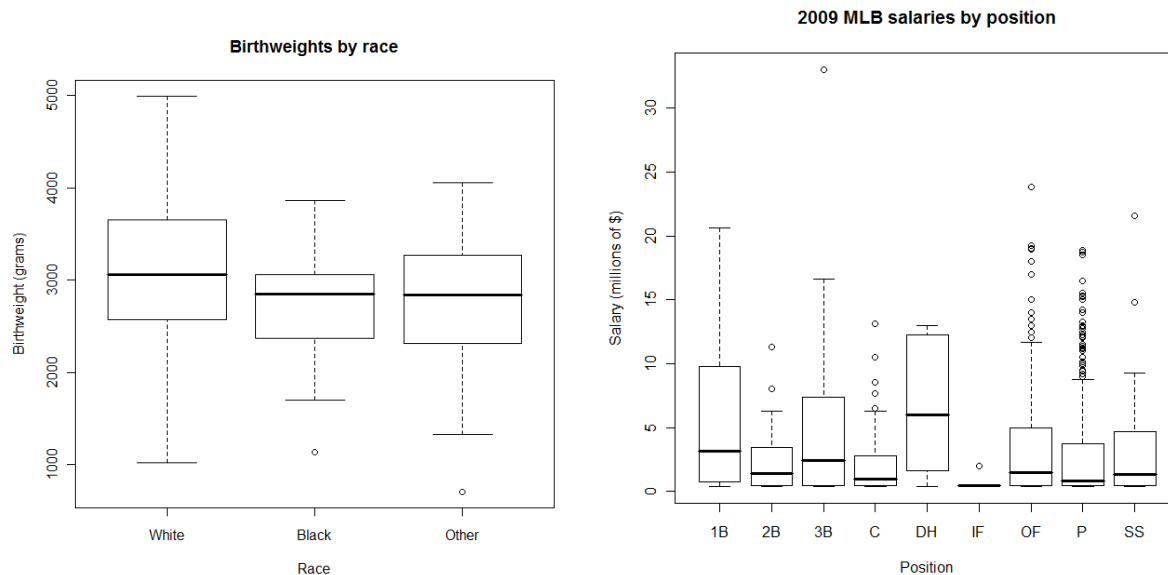
Note: the definition of an outlier used by the boxplot algorithm (the 1.5 x IQR criterion) is simply an automatic one that a computer can use; it is not “the” definition. I consider the 73 to be an outlier even though it does not make it (barely) by this definition.

The primary use of boxplots is to compare two or more distributions side-by-side. The key elements to compare are center and spread. For Bonds and ARod, ARod has a higher median number of homeruns per season and has much less variability in his counts than Bonds. Bonds had the highest individual season by far and also the lowest individual season.

Here are a couple of more examples of using boxplots for comparisons. The first one comes from the data frame `birthwt` which has data on 189 newborn babies and their mothers. The second one comes from the 2009 Major League Baseball salary data frame `MLB`.

```
boxplot(bwt~race, data=birthwt, names=c("White", "Black", "Other"),  
xlab="Race", ylab="Birthweight (grams)", main="Birthweights by race")
```

```
boxplot(Salary/1000000~Position, data=MLB, xlab="Position", ylab="Salary  
(millions of $)", main="2009 MLB salaries by position")
```



Remember, a boxplot does not show the shape of the distribution as well as a histogram. Its big advantage is that it can be used to compare several distributions easily. Can you think of a distributional feature that would be readily apparent in a histogram but not in a boxplot?

Note: it doesn't make sense to use boxplots if there are a small number of data values in each group (10 or less); you should consider making side-by-side dotplots that show the actual values instead. Histograms should also not be used for very small data sets.