Lecture 6: Examining/Visualizing Numerical Data Part 2

Chapter 1.6

Goals for Today

- Rule of thumb for standard deviations
- Population vs sample mean/variance/standard deviations
- Percentiles and Quartiles
- Boxplots

Rule of Thumb for Standard Deviations

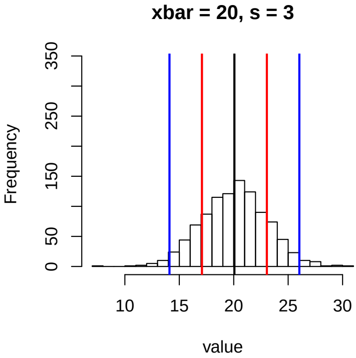
If the data distribution is bell-shaped, then

- about $\frac{2}{3}$'s of the data will be within one standard deviation of the mean
- about 95% will be within two standard deviations

Notes:

- The book has the first rule at 70%, and not $\frac{2}{3}$'s.
- This is not a hard and fast rule. Look at examples in Figure 1.27 on page 27 of text.

Going back to Second Example



Here

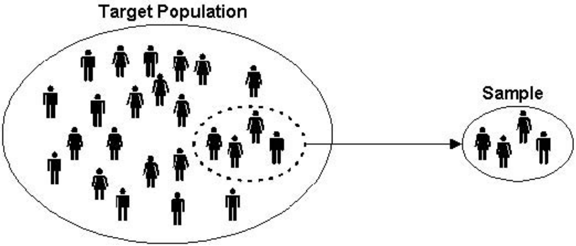
- black line is mean $\overline{x}$
- red lines mark $[\overline{x} - s, \overline{x} + s] = [20 - 3, 20 + 3] = [17, 23]$
- blue lines mark $[\overline{x} - 2s, \overline{x} + 2s] = [20 - 6, 20 + 6] = [14, 26]$

So roughly

- $\frac{2}{3}$'s of the data is between the red lines
- 95% of the data will is between the blue lines

Population vs Sample Mean/Variance/Standard Deviation

Recall from Lecture 1.3 the notion of taking a sample from a study/target population:



We slowly start articulating this concept in statistical terms. Say we are interested in the income of the individuals.

Population vs Sample Mean/Variance/Standard Deviation

The sample mean $\overline{x}$ is the mean income of the 4 individuals in our sample. However, say we didn't just ask the 4 people in the sample for their income, but rather we asked all 24 individuals in the target population. This mean would be the population mean μ (greek letter "mu").

Much in the same vein:

- The sample variance s^2 is an estimator of the true population variance σ^2 (greek letter "sigma")
- The sample standard deviation s is an estimator of the true population standard deviation σ

Population vs Sample Mean/Variance/Standard Deviation

We say that the sample mean $\overline{x}$ is an estimator of the true population mean μ (remember the notion of generalizability). This will be the basis of future lectures based on Chapter 4 from the text.

In this example, 24 is a rather small number, so what's the real leap between between $\overline{x}$ and μ? Imagine instead your population is 300 million people! Not so easy. We need $\overline{x}$ to estimate μ.

Population vs Sample Mean/Variance/Standard Deviation

	True Population Value	Sample Value
Mean	μ	$\overline{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s

The sample value is used to estimate the (true) population value.

Percentiles

A percentile (shorthand notation in my notes is %'ile) indicates the value below which a given percentage of observations in a group of observations fall.

SAT Scores from 2012
http://media.collegeboard.com/digitalServices/pdf/research/SAT-Percentile-Ranks-2012.pdf

So for example, if you scored 700 in critical reading, 95% of college-bound seniors who took the test did worse.

Quartiles

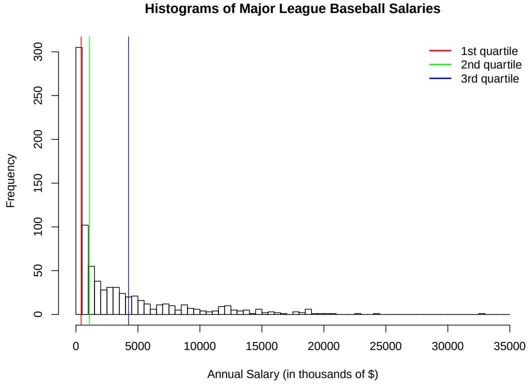
Quartiles split up the data into 4 intervals, each with (roughly) one quarter of the data: 1st (lower) quarter, 2nd quartile (median), and 3rd (upper) quartile:

So

- The lower quartile is the 25th %'ile (percentile)
- The median is the 50th %'ile
- The upper quartile is the 75th %'ile

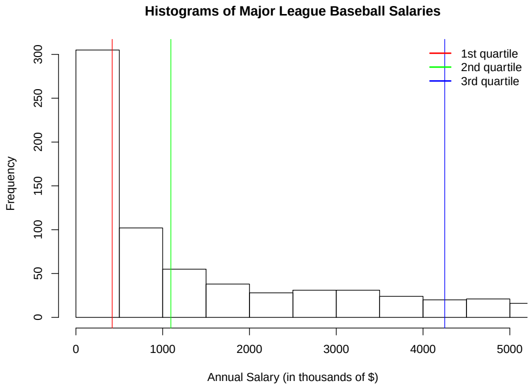
MLB Data Quartiles

```
summary(MLB$salary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  400.0   418.3  1094.0  3282.0  4250.0 33000.0
```



MLB Data Quartiles

```
summary(MLB$salary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  400.0   418.3  1094.0  3282.0  4250.0 33000.0
```



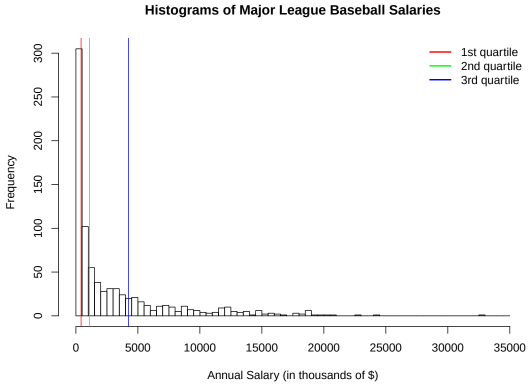
Interquartile Range

The interquartile range (IQR) is another, less-used, measure of the spread of a sample:

$$\text{IQR} = \text{upper quartile} - \text{lower quartile}$$

MLB Data Quartiles

```
summary(MLB$salary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  400.0   418.3  1094.0  3282.0  4250.0 33000.0
```



The IQR is (3rd Quartile - 1st Quartile) = 4250.0 - 418.3 = 3831.7
i.e the distance between the red and blue line.

Robust Statistics (Chapter 1.6.6)

Robust estimates are statistics where extreme observations (outliers) have less effect on their values, or stated differently:

- not as sensitive to outliers
- more "resistant to outliers"

Robust Statistics (Chapter 1.6.6)

One example illustrating the philosophy of "robustifying" to outliers is scoring in figure skating: drop the highest & lowest scores and only then take the average.

Say we have a figure skater who gets judged by judges from countries V-Z. The scores are as follows:

Country	V	W	X	Y	Z
Score	4.0	5.2	5.2	5.3	6.0

Drop the 4.0 and 6.0, then the final score is: $\frac{5.2+5.2+5.3}{3} = 5.23$

Median and IQR are Robust Statistics

The median and IQR are called robust estimates because extreme observations have little effect on their values.

Say we felt that the highest paid player in baseball, Alex Rodriguez, wasn't paid enough at 33 million. So we increase it to 45 million a season! Then the median and IQR would not change.

Boxplots

Boxplots are visual summaries of a sample $x_1, \ldots, x_n$ based on five statistics that bring to light unusual values (potential outliers):

1. lower quartile
2. median
3. upper quartile
4. smallest x_i
5. largest x_i

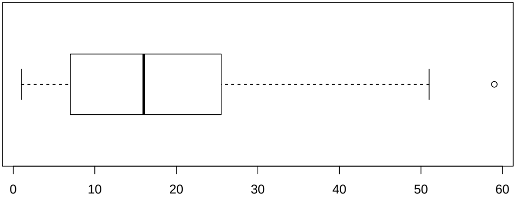
Boxplots

Example: # US Forces casualties in the war in Afghanistan for each month from 2008-2009:

7, 1, 7, 5, 16, 28, 20, 22, 27, 16, 1, 3, 14, 15, 13, 6, 12, 24, 44, 51, 37, 59, 17, 17

Boxplots

```
> summary(casualties)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.00    7.00   16.00   19.25   24.75   59.00
```



US Forces casualties in Afghanistan for each month 2008−2009

Please read page 29 on how the determine the length of the whiskers: it captures data that is no more than $1.5 \times IQR$ of both ends of the box.

Boxplots to Identify Outliers

The outlier of 59 corresponds to October 2009, when among other things:

- On October 3, 2009, a force of 300 Taliban assaulted the American Combat Outpost Keating near the town of Kamdesh of Nuristan province in eastern Afghanistan in the "Battle of Kamdesh." The attack was the bloodiest battle for US forces since the Battle of Wanat in July 2008. The attack resulted in eight Americans killed.
- 14 died in two separate helicopter crashes on October 26, 2009.

Outliers Are Relatively Extreme

An outlier is an observation that appears extreme relative to the rest of the data.

Why it is important to look for outliers? Examination of data for possible outliers serves many useful purposes, including

- Identifying strong skew in the distribution.
- Identifying data collection or entry errors.
- Providing insight into interesting properties of the data.

Next Time

We discuss examining/visualizing categorical data. In particular:

- Contingency Tables
- Barplots
- Piecharts