STA291
Fall 2009

LECTURE 12
Tuesday, 6 October

Five-Number Summary (Review)

- Maximum, Upper Quartile, Median, Lower Quartile, Minimum
- Statistical Software SAS output (Murder Rate Data)

Quantile	Estimate
100% Max	20.30
75% Q3	10.30
50% Median	6.70
25% Q1	3.90
0% Min	1.60

Note the distance from the median to the maximum compared to the median to the minimum.

Interquartile Range

- The Interquartile Range (IQR) is the difference between upper and lower quartile
- IQR = $Q_3 - Q_1$
- IQR= Range of values that contains the middle 50% of the data
- IQR increases as variability increases

Box Plot (AKA Box-and-Whiskers Plot)

- A box plot is basically a graphical version of the five-number summary (unless there are outliers)
- It consists of a ***box*** *that contains the central 50%* of the distribution (from lower quartile to upper quartile),
- A ***line*** *within the box that marks the median,*
- And ***whiskers*** *that extend to the maximum and* minimum values, unless there are outliers

Outliers

- An observation is an outlier if it falls
  - more than 1.5 IQR above the upper quartile or
  - more than 1.5 IQR below the lower quartile
- Example: Murder Rate Data w/o DC
  - upper quartile Q3 = 10.3
  - IQR = 6.4
  - Q3 + 1.5 IQR = ________
  - Any outliers?

Illustrating Boxplot with Murder Rate Data

- (w/o DC—key: 20|3 = 20.3)

Quantile	Estimate
100% Max	20.30
75% Q3	10.30
50% Median	6.70
25% Q1	3.90
0% Min	1.60

```
Stem Leaf                      #
  20 3                         1
  19
  18
  17
  16
  15
  14
  13 135                       3
  12 7                         1
  11 334469                    6
  10 2234                      4
   9 08                        2
   8 03469                     5
   7 5                         1
   6 034689                    6
   5 0238                      4
   4 46                        2
   3 0144468999               10
   2 039                       3
   1 67                        2
     ----+----+----+----+
```

Measures of Variation

- Mean and Median only describe a typical value, but not the spread of the data
- Two distributions may have the same mean, but different variability
- Statistics that describe variability are called measures of variation (or dispersion)

Sample Measures of Variation

- Sample Range:

  Difference between maximum and minimum sample value

- Sample Variance: $s^2 = \dfrac{\sum\left(x_i - \bar{x}\right)^2}{n-1}$

- Sample Standard Deviation: $s = \sqrt{s^2} = \sqrt{\dfrac{\sum\left(x_i - \bar{x}\right)^2}{n-1}}$

- Sample Interquartile Range:

  Difference between upper and lower quartile of the sample

Population Measures of Variation

- Population Range:

  Difference between maximum and minimum population values

- Population Variance: $$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

- Population Standard Deviation:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

- Population Interquartile Range:

  Difference between upper and lower quartile of the population

Range

- Range: Difference between the largest and smallest observation
- Very much affected by outliers (one misreported observation may lead to an outlier, and affect the range)
- The range does not always reveal different variation about the mean

Deviations

- The deviation of the i^{th} observation, x_i, from the sample mean, $\bar{x}$, is $x_i - \bar{x}$, the difference between them
- The sum of all deviations is zero because the sample mean is the center of gravity of the data (remember the balance beam?)
- Therefore, people use either the sum of the absolute deviations or the sum of the squared deviations as a measure of variation

Sample Variance

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

The *variance* of *n* observations is the sum of the squared deviations, divided by *n* – 1.

Variance: Interpretation

- The variance is about the average of the squared deviations
  - "average squared distance from the mean"

- Unit: square of the unit for the original data
- Difficult to interpret

- Solution: Take the square root of the variance, and the unit is the same as for the original data

Sample standard deviation

- The standard deviation *s* is the positive square root of the variance

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

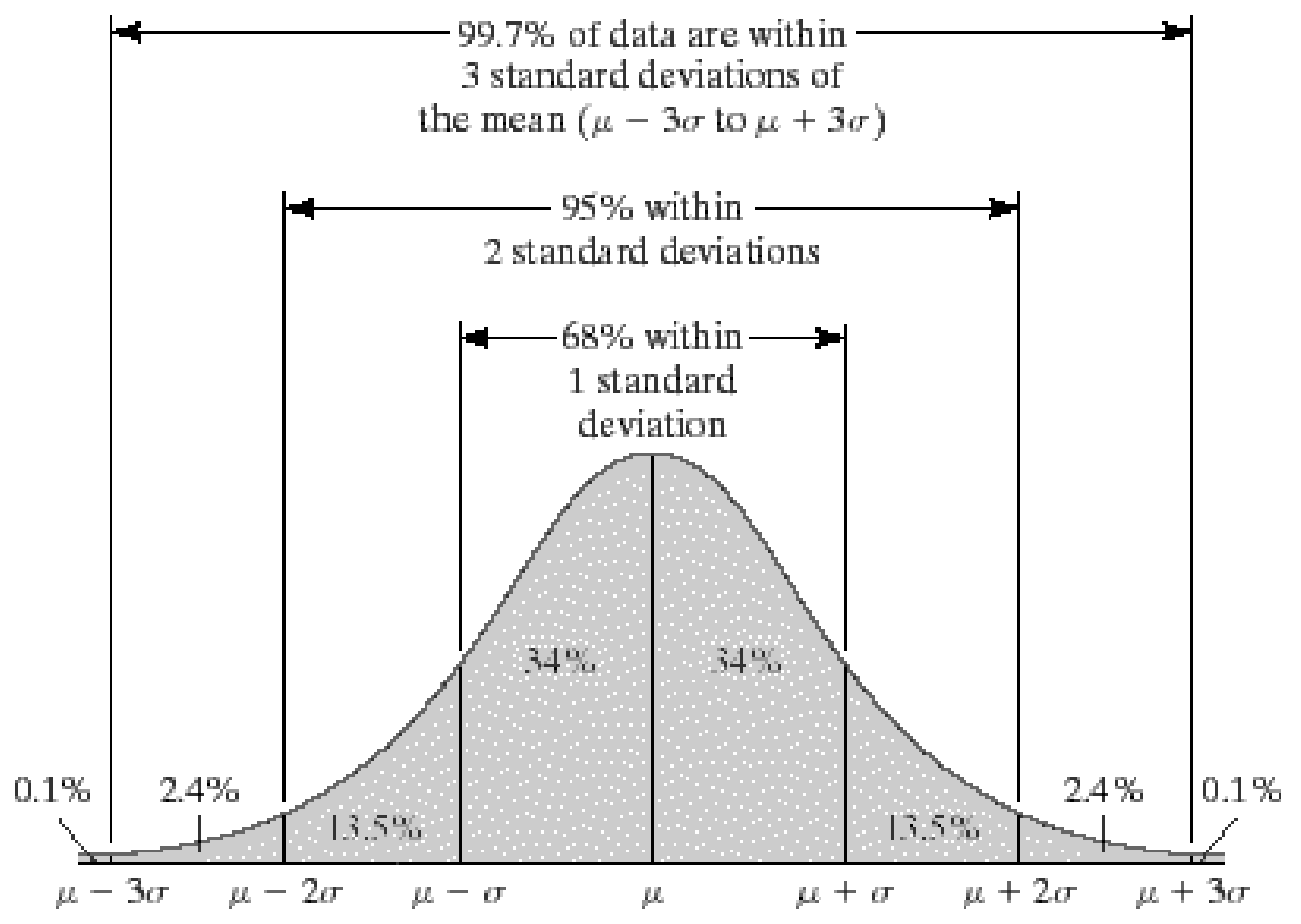
Standard Deviation: Properties

- $s \geq 0$ always
- $s = 0$ only when all observations are the same
- If data is collected for the whole population instead of a sample, then $n-1$ is replaced by n
- s is sensitive to outliers

Standard Deviation Interpretation: Empirical Rule

- ***If*** the histogram of the data is **approximately symmetric and bell-shaped**, then
  - About **68% of the data are within one** standard deviation from the mean
  - About **95% of the data are within two** standard deviations from the mean
  - About **99.7% of the data are within three** standard deviations from the mean

Standard Deviation Interpretation: Empirical Rule



Sample Statistics, Population Parameters

- Population mean and population standard deviation are denoted by the Greek letters μ (mu) and σ (sigma)
- They are unknown constants that we would like to estimate
- Sample mean and sample standard deviation are denoted by $\bar{x}$ and s
- They are random variables, because their values vary according to the random sample that has been selected

Attendance Survey Question 12

- ***On a your index card:***
  - Please write down your name and section number
  - Today's Question: