Announcements

- Quiz 1 available at 4pm (until Mon, Jan 14 at 3pm). If you *do not* receive an email later today telling you it is available, you need to contact me so I can add you to the list.
- Yan He's new office hours are Tues/Thurs 2:00 – 3:30
- Today we need to finish Lecture 2. If time, will show R Commander. It will also be covered next Fri in discussion.
- If you plan to use R Commander in the ICS labs, you need to get an account. See course webpage for information. In the meantime, you can use a temporary account:

    Username: ics-temp , Password: Anteat3r

Today's Homework (due Mon, Jan. 14):

Chapter 2: #96, 128, 130

Today:

- Finish Lecture 2 (but read on your own slides about detecting cheating with a dotplot, slides 38 and 39).
- Cover Section 2.7
- If time, go over how to install and use R Commander. See handouts on course webpage. First assignment using R Commander will be given next week. Strongly encourage you to install the software now!

2.5 More Numerical Summaries of Quantitative Data

Notation for Raw Data:

n = number of individuals in a data set

$x_1, x_2, x_3, \ldots, x_n$ represent individual raw data values

Example: A data set consists of heights for the first 4 students in the UCDavis1 dataset.

So $n = 4$, and

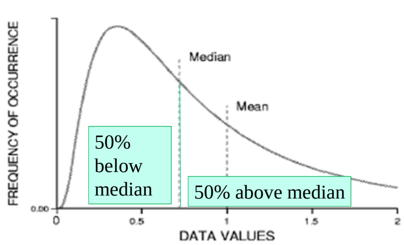
$x_1 = 66$, $x_2 = 64$, $x_3 = 72$, $x_4 = 68$

Describing the "Location" of a Data Set

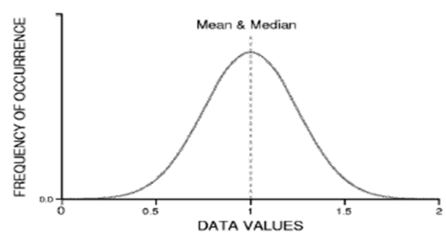
- **Mean:** the numerical average
- **Median:** the middle value (if n odd) or the average of the middle two values (n even)

Symmetric: mean = median
Skewed Left: usually mean < median
Skewed Right: usually mean > median

Mean versus Median for Data values skewed to the right



Bell-shaped distribution



Determining the Mean and Median

The Mean $\bar{x} = \frac{\sum x_i}{n}$

where $\sum x_i$ means "add together all the values"

The Median

If n is odd: *Median* = middle of ordered values.
Count $(n + 1)/2$ from top or bottom of ordered list.
Example: 5, 7, $\boxed{10,}$ 13, 15 $(n + 1)/2 = 6/2 = 3$

If n is even: *Median* = average of middle two ordered values. Average the values that are $(n/2)$ and $(n/2) + 1$ down from top of ordered list.

7

The Mean, Median, and Mode

Ordered Listing of 28 Exam Scores

32, 55, 60, 61, 62, 64, 64, 68, 73, 75, 75, 76, 78, 78, 79, 79, 80, 80, 82, 83, 84, 85, 88, 90, 92, 93, 95, 98

- **Mean (numerical average): 76.04**
- **Median: 78.5 (halfway between 78 and 79)**
- **Mode (most common value): no single mode exists, many occur twice.**

8

The Influence of Outliers on the Mean and Median

- **Larger influence on mean** than median.
- High outliers and data skewed to the *right* will increase the mean.
- Low outliers and data skewed to the *left* will decrease the mean.

Ex: Suppose ages at death of your eight great-grandparents are: 28, 40, 75, 78, 80, 80, 81, 82.
Skewed to the *left*, so mean is lower than median.
Mean age is 544/8 = 68 years old
Median age is (78 + 80)/2 = 79 years old

9

Caution: *Normal* does not mean *Average*

Common mistake to confuse "average" with "normal".
Is woman 5 ft. 10 in. tall 5 inches taller than <u>normal</u>??

Example: How much hotter than normal is normal?

"October came in like a dragon Monday, hitting 101 degrees in Sacramento by late afternoon. That temperature tied the record high for Oct. 1 set in 1980 – and was 17 degrees *higher than normal for the date*. (Korber, 2001, italics added.)"

Article had thermometer showing "normal high" for the day was 84 degrees. High temperature for Oct. 1st is quite variable, from 70s to 90s. While 101 was a record high, it was not "17 degrees higher than normal" if "normal" includes the range of possibilities likely to occur on that date.

10

Describing Spread (Variability): Range, Interquartile Range and Standard deviation

- **Range** = high value – low value
- **Interquartile Range (IQR)** = upper quartile – lower quartile = Q_3 - Q_1 (to be defined)
- **Standard Deviation**
(will cover with Section 2.7)

11

Example 2.13 *Fastest Speeds Ever Driven*

Five-Number Summary for 87 males

	Males (87 Students)	
Median	110	
Quartiles	95	120
Extremes	55	150

- ***Median*** = 110 mph measures the center of the data (there were many values of 110, see page 42)
- Two ***extremes*** describe spread over 100% of data
Range = 150 – 55 = 95 mph
- Two ***quartiles*** describe spread over middle 50% of data
Interquartile Range = 120 – 95 = 25 mph

12

Notation and Finding the Quartiles

Split the ordered values at median:

•half at or below the median ("at" if ties)

•half at or above the median

Q_1 = **lower quartile**
= median of data values
that are (at or) *below* the median

Q_3 = **upper quartile**
= median of data values
that are (at or) *above* the median

Example 2.13 *Fastest Speeds (cont)*

Ordered Data (in rows of 10 values) for the 87 males:

55	60	80	80	80	80	85	85	85	85
90	90	90	90	90	92	94	95	95	95
95	***95***	95	100	100	100	100	100	100	100
100	100	101	102	105	105	105	105	105	105
105	105	109	**110**	110	110	110	110	110	110
110	110	110	110	110	112	115	115	115	115
115	115	120	120	120	***120***	120	120	120	120
120	120	124	125	125	125	125	125	125	130
130	140	140	140	140	145	150			

- ***Median*** = (87+1)/2 = 44th value in the list = 110 mph
- Q_1 = median of the 43 values below the median = (43+1)/2 = 22nd value from the start of the list = 95 mph
- Q_3 = median of the 43 values above the median = (43+1)/2 = 22nd value from the end of the list = 120 mph

Percentiles

The ***k*th percentile** is a number that has k% of the data values at or below it and $(100 - k)$% of the data values at or above it.

- Lower quartile: 25th percentile
- Median: 50th percentile
- Upper quartile: 75th percentile

Describing Spread with Standard Deviation

Standard deviation measures variability by summarizing how far individual data values are from the mean. It's most useful for bell-shaped data.

Think of the standard deviation as ***roughly the average distance values fall from the mean***.

Describing Spread with Standard Deviation: A very simple example

Numbers	Mean	Standard Deviation
100, 100, 100, 100, 100	100	0
90, 90, 100, 110, 110	100	10

Both sets have same mean of 100.

Set 1: all values are equal to the mean so there is *no variability* at all.

Set 2: one value equals the mean and other four values are 10 points away from the mean, so the *average distance away from the mean is about* 10.

Calculating the Standard Deviation

Formula for the (***sample***) **standard deviation**:

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

The value of s^2 is called the (*sample*) **variance.** An equivalent formula, easier to compute, is:

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

Calculating the Standard Deviation

Example: 90, 90, 100, 110, 110

Step 1: Calculate $\overline{x}$, the sample mean. *Ex*: $\overline{x}$ = 100

Step 2: For each observation, calculate the difference between the data value and the mean.
Ex: -10, -10, 0, 10, 10

Step 3: Square each difference in step 2.
Ex: 100, 100, 0, 100, 100

Step 4: Sum the squared differences in step 3, and then divide this sum by $n - 1$. Result = *variance* s^2
Ex: 400/(5 – 1) = 400/4 = 100

Step 5: Take the square root of the value in step 4.
Ex: s = *standard deviation* = $\sqrt{100} = 10$

Population Standard Deviation

Data sets usually represent a sample from a larger population. If the data set includes measurements for an ***entire population***, the notations for the mean and standard deviation are different, and the formula for the standard deviation is also slightly different.

A **population mean** is represented by the Greek μ ("mu"), and the **population standard deviation** is represented by the Greek "sigma" (lower case)

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

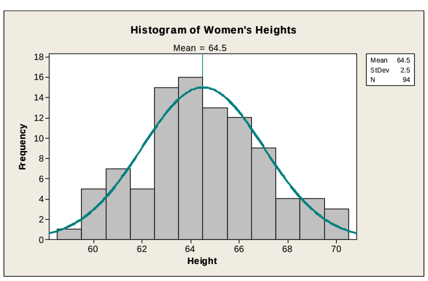
Bell-shaped distributions

- Measurements that have a bell-shape are so common in nature that they are said to have a *normal distribution*.
- Knowing the mean and standard deviation *completely determines* where all of the values fall for a normal distribution, assuming an infinite population!
- In practice we don't have an infinite population (or sample) but if we have a large sample, we can get good approximations of where values fall.

Examples of bell-shaped data

- Women's heights
  - mean = 64.5 inches, s = 2.5 inches
- Men's heights
  - mean = 70 inches, s = 3 inches
- IQ scores
  - mean = 100, s = 15 (or for some tests, 16)
- High school GPA for intro stat students
  - mean = 3.1, s = 0.5
- Verbal SAT scores for UCI incoming students
  - mean = 569, s = 75

Women's heights from UCDavis data, n = 94
Note approximate bell-shape of histogram
"Normal curve" with mean = 64.5, s = 2.5
superimposed over histogram



Interpreting the Standard Deviation for Bell-Shaped Curves:
The Empirical Rule

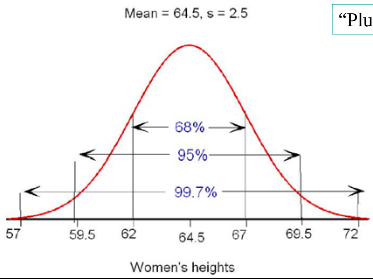
For any bell-shaped curve, approximately

- **68%** of the values fall within **1 standard deviation** of the mean in either direction
- **95%** of the values fall within **2 standard deviations** of the mean in either direction
- **99.7%** (almost all) of the values fall within **3 standard deviations** of the mean in either direction

Ex: Population of women's heights

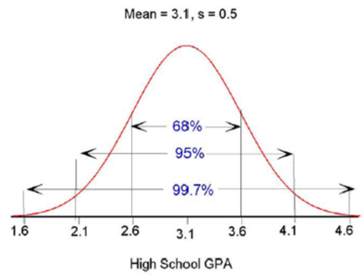
- 68% of heights are between 62 and 67 inches (64.5 ± 2.5)
- 95% of heights are between 59.5 and 69.5 inches
- 99.7% of heights are between 57 and 72 inches

"Plus and minus"



Ex: Population of high school GPAs

- 68% of GPAs are between 2.6 and 3.6
- 95% of GPAs are between 2.1 and 4.1
- 99.7% of GPAs are between 1.6 and 4.6



Other Examples

- Men's heights
  - mean = 70 inches, s = 3 inches
- IQ scores
  - mean = 100, s = 15
- Verbal SAT scores for UCI incoming students
  - mean = 569, s = 75

Women's Heights: How well does the Empirical Rule work?

Mean height for the 94 UC Davis women was 64.5, and the standard deviation was 2.5 inches. Let's compare actual with ranges from Empirical Rule:

Range of Values:	Empirical Rule	Actual number	Actual percent
Mean ± 1 s.d.	68% in 62 to 67	70	70/94 = 74.5%
Mean ± 2 s.d.	95% in 59.5 to 69.5	89	89/94 = 94.7%
Mean ± 3 s.d.	99.7% in 57 to 72	94	94/94 = 100%

Another Example

Final averages from one of my Stat 7 classes, for students who passed. Here is a stemplot:

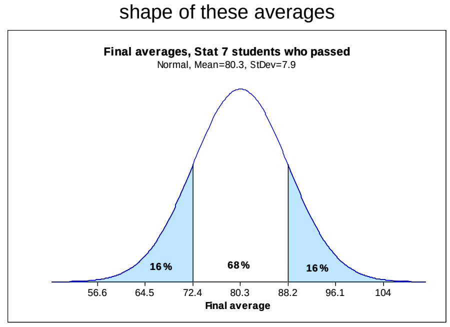
```
6|24
6|5556666667788899999
7|00000001111112233333344444449
7|555555555566666667777777788888888888899999
8|0000000011111111111222222223333333333344444444444
8|55555566666666777888899999
9|0001111111122233344
9|555568
```

Mean = 80.3, s = 7.9, n = 190; Empirical Rule intervals are 72.4 to 88.2; 64.5 to 96.1; 56.6 to 104

Normal curve showing *theoretical* 68% range

Actual in that range: 124/190 = .6526 or 65.3%

So a bell-shaped curve is a good approximation for the shape of these averages



The Empirical Rule, the Standard Deviation, and the Range

- From Empirical Rule: range from the minimum to the maximum data values is about 4 to 6 standard deviations for data sets with an approximate bell shape.
- *For a large data set, you can get a rough idea of the value of the standard deviation by dividing the range by 6 (or 4 or 5 for a smaller dataset)*

$$s \approx \frac{Range}{6}$$

Ex: Stat 7 scores, $s = 7.9$,
Range = 98 – 62 = 36 = 4.6 s

Standardized *z*-Scores

Standardized score or ***z*-score**:

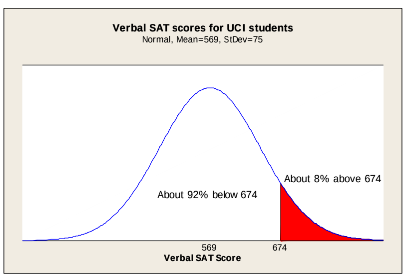
$$z = \frac{\text{Observed value} - \text{Mean}}{\text{Standard deviation}}$$

Example: UCI Verbal SAT scores had mean = 569 and s = 75. Suppose someone had SAT = 674:

$$z = \frac{674 - 569}{75} = +1.40$$

Verbal SAT of 674 for UCI student is 1.40 standard deviations ***above*** the mean for UCI students.

Verbal SAT of 674 is 1.40 standard deviations above mean.
To find proportion above or below, use Excel or R Commander
For Excel, see page 52. For R Commander, see webpage.



The Empirical Rule Restated for Standardize Scores (z-scores):

For bell-shaped data,

- About **68%** of the values have *z*-scores between –1 and +1.
- About **95%** of the values have *z*-scores between –2 and +2.
- About **99.7%** of the values have *z*-scores between –3 and +3.

Installing and Using R and R Commander

- "R" is a sophisticated and free statistical programming language.
- *R Commander* is an add-on, also free, that is menu-driven. It doesn't do everything R does.
- You can use R Commander in the ICS Computer labs, or install it on your computer.
- See handouts on course web page for installing R and R Commander, and for using R Commander for Chapters 2 and 3.
- If time, do R Commander demo.

Today's Homework (due Mon, Jan. 14):
Chapter 2: #96, 128, 130

36