UCLA STAT 13

Introduction to Statistical Methods for the Life and Health Sciences

Instructor: Ivo Dinov,

Asst. Prof. of Statistics and Neurology

Teaching Assistants:

Fred Phoa, Kirsten Johnson,

Ming Zheng & Matilda Hsieh

University of California, Los Angeles, Fall 2005

http://www.stat.ucla.edu/~dinov/courses_students.html

Parameters and Statistics

- Variables can be summarized using statistics.
- **Definition**: A *statistic* is a numerical measure that describes a characteristic of the sample.
- **Definition**: A *parameter* is a numerical measure that describes a characteristic of the population.
- We use *statistics* to estimate *parameters*

Measures of Centrality

- Recall that center is #2 of the BIG three.
- Measures of center include:
  - the mean
  - the median
  - the mode (the value with the highest frequency)
- These measures all describe the center of a distribution in a slightly different way

Measures of Center

- The Mean
  - aka the average
  - can be thought of as the balancing point of a distribution

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Measures of Center

Example: In an experiment with some statistics students, 8 male students were randomly selected and asked to perform the standing long jump. In reality every student participated, but for the ease of calculations below we will focus on these eight students. The long jumps were as follows:

long jump (in.)
74
78
106
80
68
64
60
76

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{74+78+...+60+76}{8} = 75.75 inches$$

Measures of Center

- The Median
  - can be thought of as the point that divides a distribution in half (50/50)
- Steps to find the median:
  1. Arrange the data in ascending order $\left(\text{observation } \frac{(n+1)}{2}\right)$

  2a. If n is odd, the median is the middle value

  2b. If n is even, the median is the average of the middle two values

$$\left(\text{the average of observations } \frac{n}{2} \text{ and } \left(\frac{n}{2}+1\right)\right)$$

Measures of Center

- Example: Long Jump (cont')
  - Because n is even, the median will be the average of the middle two values

long jump (in.)		→	long jump (in.)	
74	68		60	76
78	64		64	78
106	60		68	80
80	76		74	106

$$median = \frac{74+76}{2} = 75 inches$$

Resistance

- **Definition:** A statistic is said to be ***resistant*** if the value of the statistic is relatively unchanged by changes in a small portion of the data
- Referencing the formulas for the median and the mean, which statistic seems to be more resistant?
- Example: Long Jump (cont')
  - Let's remove the student with the long jump distance of 106 and recalculate the median and mean.

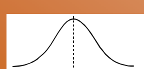
```
Descriptive Statistics: distance

Variable  N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum
distance  7   0  71.43     2.85   7.55    60.00  64.00   74.00  78.00    80.00
```

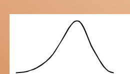

Center vs. Shape

- We can also use the mean and median to help interpret the shape of a distribution
- In a unimodal distribution:
  - mean = median
  - mean > median
  - mean < median

Positively skewed (long upper tail)



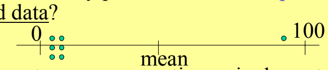
(d) Symmetric



Negatively skewed (long lower tail)

Mean, Median, Mode, Quartiles, 5# summary

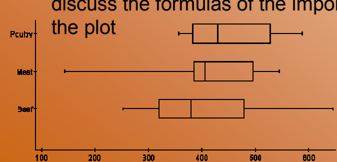
- The sample mean is the average of all numeric obs's.
- The sample median is the obs. at the index $(n+1)/2$ (note take avg of the 2 obs's in the middle for fractions like 23.5), of the observations ordered by size (small-to-large)?
- The sample median usually preferred to the sample mean for <u>skewed data</u>?



- Under what circumstances may quoting a <u>single center</u> (be it mean or median) not make sense?(multi-modal)
- What can we say about the sample mean of a qualitative variable? (meaningless)

Boxplots

- An additional graphical display for the data that utilizes some of these measures of center is called a boxplot.
  - slightly painful to construct by hand
  - we will rely on the computer, but we will still discuss the formulas of the important aspects of the plot



Boxplots

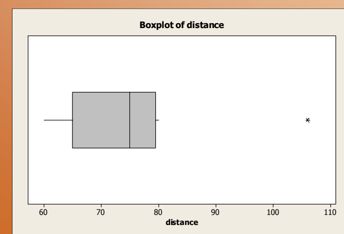
- The five number summary:
  - minimum: the smallest observation
  - maximum: the largest observation
  - median: splits the data into 50/50
  - quartiles: split the data into quarters
    - Q1 is the lower quartile and Q3 is the upper quartile
- A boxplot is a visual representation of the five number summary

Boxplots

- There are four additional features of a boxplot
  - Interquartile range (IQR): Q3 – Q1, the spread of the middle 50% of the data
  - whiskers
    - extend from Q1 and Q3 to the smallest* and largest* observations within the *fences
  - *fences
    - used to identify extreme observations
    - lower fence (LF): Q1 – 1.5(IQR)
    - upper fence (UF): Q3 + 1.5(IQR)
  - outliers
    - extreme observations that fall outside the fences

Boxplots

- Example (cont'): Using the long jump data a boxplot of distance would be:



Measures of Spread

- Recall that spread is #3 of the BIG three.
- Measures of spread include:
  - the range
  - the variance
  - the standard deviation

Measures of Spread

- The range
  - easiest measure of spread to calculate
  - not the "best" measure of spread
  - range = max - min
- Example: Long Jump (cont')
  - Calculate the range for the long jump data

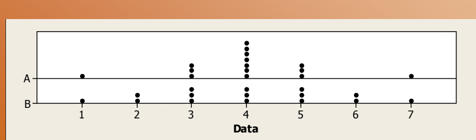
Descriptive Statistics: distance

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
distance	8	0	75.75	4.98	14.08	60.00	65.00	75.00	79.50	106.00

Range = 106 – 60 = 46

Measures of Spread – A & B Dot plot

- The range (cont')
- Why is the range not the best measure of spread?
  - Suppose we have the following data sets, dotplots below.
  - Intuitively which plot (A or B) seems to have more spread (ie. less cluster)?



Measures of Spread

- The standard deviation
  - The logic behind the standard deviation is to measure the difference (ie. deviation) between each observation and the mean
  - A deviation is $y_i - \bar{y}$
  - What seems like a reasonable way to find an "average" deviation?
  - Big problem, why?

$$\sum (y_i - \bar{y}) = 0$$

  - How could we solve this problem?

Measures of Spread

- The variance

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

- The standard deviation (sd)

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

  - Why use the sd and not the variance?

Measures of Spread

- Example (cont'): Calculate the sd

long jump (in.)	
74	68
78	64
106	60
80	76

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

$$= \sqrt{\frac{(74 - 75.75)^2 + (78 - 75.75)^2 + \ldots + (76 - 75.75)^2}{8-1}}$$

$$= 14.079 \ \textit{inches}$$

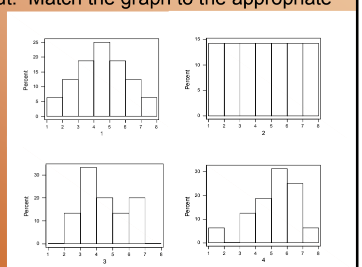
Descriptive Statistics: distance

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
distance	8	0	75.75	4.98	14.08	60.00	65.00	75.00	79.50	106.00

Measures of Spread

- Below we have four relative frequency histograms and four portions of output. Match the graph to the appropriate output.

	Mean	Median	StDev
A	4.688	5.000	1.493
B	4.000	4.000	1.633
C	3.933	4.000	1.387
D	4.000	4.000	2.075



The Empirical Rule

- The empirical rule is useful when talking about a distribution, using the standard deviation in terms of it's distance from the mean.
- In general, for symmetric distributions:

$$\bar{y} \pm s \approx 68\%$$

$$\bar{y} \pm 2s \approx 95\%$$

$$\bar{y} \pm 3s \approx > 99\%$$

- NOTE: If the distribution is not unimodal symmetric the empirical rule may not hold.

The Empirical Rule

- Example (hotdogs cont'): From the hotdog data we have the following output:

Descriptive Statistics: Calories

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Calories	54	0	145.44	4.00	29.38	86.00	131.75	145.00	173.50

Variable	Maximum	Range
Calories	195.00	109.00

$$\bar{y} \pm s = 145.44 \pm 29.38 = (116.06, 174.82)$$

$$\bar{y} \pm 2s = 145.44 \pm 2(29.38) = (86.68, 204.20)$$

$$\bar{y} \pm 3s = 145.44 \pm 3(29.38) = (57.30, 233.58)$$

The Empirical Rule

- Example (hotdogs cont'): From the hotdog data we have the following intervals:

$$\bar{y} \pm s = 145.44 \pm 29.38 = (116.06, 174.82)$$

$$\bar{y} \pm 2s = 145.44 \pm 2(29.38) = (86.68, 204.20)$$

$$\bar{y} \pm 3s = 145.44 \pm 3(29.38) = (57.30, 233.58)$$

- 30/54 = 55% is this close to 68%?

Character Stem-and-Leaf Display

```
Stem-and-leaf of Calories  N  = 54
Leaf Unit = 1.0

   2    8 67
   4    9 49
   9   10 22677
  11   11 13
  12   12 9
  22   13 1225556899
 (11)  14 01234667899
  21   15 223378
  15   16
  15   17 235569
   9   18 1246
   5   19 00015
```

The Goal

- **Definition:** A *statistical inference* is the process of drawing conclusions about a population based on observations in a sample.
  - To make a statistical inference we want the sample to be representative of the population.
- How could we ensure this?

The Goal

- **Definition:** *Random* means that each subject of the population must have an equal chance of being selected.
- Why does this seem important for statistics?
- How can we ensure random selection?

More Notation

- Both samples and populations have numeric quantities of interest, such as:
  - mean (the average)
  - standard deviation (the spread)
  - proportion (percent)
- For what type of variable(s) would each of these numeric quantities be appropriate?

More Notation

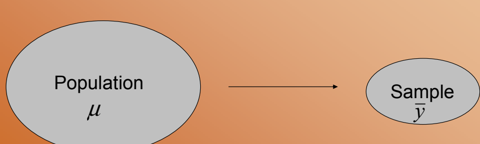
- **Recall:** A characteristic of the population is called a parameter and a characteristic of a sample is called a statistic.

	Mean	Standard Deviation	Proportion
Population	μ	σ	p
Sample	$\bar{x}$	s	$\hat{p}$

- Under what circumstances would we know μ ?
- What seems like a good estimate of μ ?

More Notation

- **Recall:** Statistics estimate parameters.



- The big question is: how good of an estimate are these values?