MSPP PPOL 501-02 & -06: Fall 2014
Course Notes #4: Measures of Central Tendency and Dispersion
Professor Carolyn Hill

I. CENTRAL TENDENCY (covered in Math Pre-Orientation)

- Three basic measures are often used to describe the "typical" observation in a sample or population: mode, median, mean.

- Which measure to use? depends on the level of measurement and characteristics of the data.

A. *Mode*

- the category or class of variable that contains more cases (or has a higher frequency) than any other category – i.e., the most commonly-occurring score, category, value, etc. Examples:
- When to use the mode?
  -- when the variable being examined is nominal, only the mode can be used as a measure of central tendency
  -- when you want to describe the most common value of the distribution (for a nominal, ordinal, or interval-ratio variable)
- Problems with the mode?
  -- there may be more than one mode in a particular frequency distribution (not a "problem" per se). A distribution with one mode is *unimodal*; with two modes, *bimodal*; with three modes, *trimodal*, etc.

B. *Median*

- The value of the variable in the middle position when the values are ranked low to high (or high to low).
- There are as many values of the variable below the median value as above the median value.
- To find the median:
  (1) Reorder the values so that they are in order from low to high (or high to low).
  (2) Find the observation in the middle position, where this position is located at the following position in the distribution you created in (1): $\frac{n+1}{2}$
  (3) The median is the value of the observation in the middle position.
  Note that when the sample size n is an odd number, the middle position is a whole number. When n is an even number, the middle position will be in-between two values.
- When to use the median?
  -- When the variable of interest is either ordinal or interval/ratio
  -- When the mean would give a misleading value of the "typical" person

C. *Mean*

- The arithmetic average of all values of a variable:

$$\bar{X} = \frac{\sum X_i}{n} \text{ for a sample}$$

$$\mu = \frac{\sum X_i}{N} \text{ for a population}$$

- Special characteristics of the mean:
  (1) If a variable's mean $(\bar{X})$ is subtracted from each observation's value for the variable (X_i), then the sum of these differences is zero. :

$$\sum_i (X_i - \bar{X}) = 0$$

  (2) If you sum the squared differences between the mean and each observation's value for the variable, you will get a lower number than by using any other point along the distribution. This will show up later!

$$\sum (X_i - \bar{X})^2 = \text{minimum}$$

  (3) Sensitive to the value of extreme cases – that is, variable values that are much higher or lower than the other values.

- The relative position of the median and mean of a distribution will indicate whether the data are skewed:

-- *positive skew* if mean > median
-- *negative skew* if mean < median
-- *no skew* if mean = median

- Extra note: it's also possible to calculate the mean from data that are grouped.

$$\bar{X} = \frac{\sum fX_i}{n}$$

II. DISPERSION (covered in Math Orientation)

A. *Range*

- The difference between the largest and smallest value in a distribution (e.g.,
- With extreme scores, the range may not give a very good picture of the distribution. Instead, a restricted range might be more meaningful:

B. *Percentile*

- Identifies the point below which a specific percentage of cases fall.
- To find the value of a case at a particular percentile, multiple the number of observations (*n)* by the proportional value of the percentile you're interested in (e.g., to find the 90th percentile, multiply (*n*0.90).*

C. *Interquartile range (IQR)*

- The difference between the values at the third and first quartiles of the distribution – the 25th and 75th percentiles (i.e., the middle half of the distribution)
- Disadvantages to using range and IQR: (1) you ignore a lot of information – you only use two scores in each measure; (2) don't get an idea about how much scores are different from the center of the distribution

D. *Variance (and standard deviation)*

$$\frac{\sum (X_i - \mu)^2}{N} = \text{variance of a population distribution} = \sigma^2$$

$$\frac{\sum (X_i - \bar{X})^2}{n-1} = \text{variance of a sample distribution} = s^2$$

Notes: Remember that N is the population size; and *n* is the sample size. Later we'll go into more detail about why there's an *n-1* in the denominator of the sample standard deviation, but not in the population standard deviation.

- Think of the variance as the average squared deviation from the mean

- As deviations grow large, so too does the variance – potentially on a different magnitude from the data in the distribution we're examining.

- We only squared the deviations to keep the sum from being zero (a property of the mean). Now that we have a non-zero number, take the square root to get a statistic that is back in the metric we started out in:

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} \quad , \quad so \quad s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

III. APPLICATION: CRIME LEVELS EXAMPLE

Suppose you measured the number of reported assaults in each of the 17 different neighborhoods in a city and came up with the following numbers:

13, 41, 15, 16, 18, 141, 20, 38, 21, 21, 27, 22, 154, 24, 25, 26, 36

* Calculate all appropriate measures of central tendency and dispersion.

* How would you describe the typical level of assault in the city? Write a brief paragraph, incorporating guidelines from Miller.

IV. PRODUCING AND INTERPRETING FREQUENCY TABLES AND CROSS-TABULATIONS IN STATA

- Frequency distributions can be used for nominal, ordinal, or interval/ratio data (recall that int/rat data may best be presented in a frequency table by grouping).

- Simple frequency distributions in STATA, and cross-tabulations, will be fundamental parts of any initial data analysis.

- Caution: Before you hit "print" on the STATA output window, scroll through to see whether you really want to print everything there! Often, you can cut and paste only those portions that you want. (e.g., you may inadvertently run a frequency table for the interval-ratio variables in the data set, and this may go on for pages and pages!)

- Let's interpret frequency tables and cross-tabs a little bit more: We used the nlsy97 data set, and the variables region97 and region99 to produce the following cross-tab, using this code:

```
. tabulate region97

cv_census_r |
 egion 1997 |      Freq.     Percent        Cum.
------------+-----------------------------------
         NE |      1,585       17.64       17.64
         NC |      2,050       22.82       40.46
          S |      3,359       37.39       77.85
          W |      1,990       22.15      100.00
------------+-----------------------------------
      Total |      8,984      100.00

. tabulate region99

cv_census_r |
 egion 1999 |      Freq.     Percent        Cum.
------------+-----------------------------------
         NE |      1,403       17.13       17.13
         NC |      1,836       22.42       39.55
          S |      3,116       38.05       77.60
          W |      1,834       22.40      100.00
------------+-----------------------------------
      Total |      8,189      100.00
```

```
tabulate region97 region99, cell col row
```

```
+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
| column percentage |
|  cell percentage  |
+-------------------+
```

cv_census_region1997	cv_census_region 1999 NE	NC	S	W	Total
NE	1,385	3	40	7	1,435
	96.52	0.21	2.79	0.49	100.00
	98.72	0.16	1.28	0.38	17.52
	16.91	0.04	0.49	0.09	17.52
NC	1	1,782	44	28	1,855
	0.05	96.06	2.37	1.51	100.00
	0.07	97.06	1.41	1.53	22.65
	0.01	21.76	0.54	0.34	22.65
S	11	34	3,005	26	3,076
	0.36	1.11	97.69	0.85	100.00
	0.78	1.85	96.44	1.42	37.56
	0.13	0.42	36.70	0.32	37.56
W	6	17	27	1,773	1,823
	0.33	0.93	1.48	97.26	100.00
	0.43	0.93	0.87	96.67	22.26
	0.07	0.21	0.33	21.65	22.26
Total	1,403	1,836	3,116	1,834	8,189
	17.13	22.42	38.05	22.40	100.00
	100.00	100.00	100.00	100.00	100.00
	17.13	22.42	38.05	22.40	100.00

V. USING STATA TO PRODUCE MEASURES OF CENTRAL TENDENCY & DISPERSION

Central Tendency and Dispersion Using STATA: Income Example

```
. summarize hhinc99

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     hhinc99 |      2188    49569.13    54394.78          0     291681
```

```
. tabstat hhinc99, statistics (count mean q min max)

    variable |         N      mean       p25       p50       p75       min
-------------+--------------------------------------------------------------
     hhinc99 |      2188  49569.13     12000     37500     70000         0
------------------------------------------------------------------------------

    variable |       max
-------------+----------
     hhinc99 |    291681
-------------------------
```

```
summarize hhinc99, detail

                  cv_income_gross_yr 1999
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0             0
 5%            0             0
10%         2500             0       Obs                2188
25%        12000             0       Sum of Wgt.        2188

50%        37500                     Mean           49569.13
                        Largest      Std. Dev.      54394.78
75%        70000        291681
90%       108000        291681       Variance       2.96e+09
95%       151500        291681       Skewness       2.320124
99%       291681        291681       Kurtosis        10.0064
```

VI. GETTING TO KNOW THE STANDARD DEVIATION AS A MEASURE OF DISPERSION

Illustration of Standard Deviations for Values in Hypothetical Samples of n=20

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

	PANEL I			*PANEL II*		
Obs #	**GROUP A**	**GROUP B**	**GROUP C**	**GROUP D**	**GROUP E**	**GROUP F**
1	1	3	5	101	103	105
2	1	4	5	101	104	105
3	1	4	5	101	104	105
4	1	4	5	101	104	105
5	2	4	5	102	104	105
6	2	4	5	102	104	105
7	2	4	5	102	104	105
8	2	5	5	102	105	105
9	3	5	5	103	105	105
10	3	5	5	103	105	105
11	4	5	6	104	105	106
12	5	5	6	105	105	106
13	5	5	6	105	105	106
14	6	5	6	106	105	106
15	7	5	6	107	105	106
16	8	6	7	108	106	107
17	9	6	7	109	106	107
18	10	6	7	110	106	107
19	12	7	8	112	107	108
20	15	7	8	115	107	108
MEAN	4.95	4.95	5.85	104.95	104.95	105.85
STDDEV	4.06	1.05	1.04	4.06	1.05	1.04
MIN	1	3	5	101	103	105
MAX	15	7	8	115	107	108
RANGE	14	4	3	14	4	3

	Frequency Table		
Value	**GROUP D**	**GROUP E**	**GROUP F**
101	4	0	0
102	4	0	0
103	2	1	0
104	1	6	0
105	2	8	10
106	1	3	5
107	1	2	3
108	1	0	2
109	1	0	0
110	1	0	0
111	0	0	0
112	1	0	0
113	0	0	0
114	0	0	0
115	1	0	0
TOTAL	20	20	20

	Frequency Table		
Value	**GROUP A**	**GROUP B**	**GROUP C**
1	4	0	0
2	4	0	0
3	2	1	0
4	1	6	0
5	2	8	10
6	1	3	5
7	1	2	3
8	1	0	2
9	1	0	0
10	1	0	0
11	0	0	0
12	1	0	0
13	0	0	0
14	0	0	0
15	1	0	0
TOTAL	20	20	20

Things to Note:

1. The populations in Panels I and II differ in that the variable values in Panel II are exactly 100 units larger than the variable values in Panel I.

2. The first two groups in each panel have identical means, but range and standard deviation in the second group in each panel are smaller, compared to the first group the panel.

3. The middle and last groups in each panel have virtually the same standard deviation, but the mean is higher in the last group in each panel.

4. Figures 1 and 2 are histograms for Groups A, B, and C; and for D, E, and F. The larger standard deviation for Group A is visually confirmed by Fig. 1; Group B, which has the same mean but a smaller standard deviation, is much less disperse. Group C, which has the same standard deviation but a higher mean than group B, also shows some clustering, but it would be difficult on visual inspection alone to know that the standard deviations of Groups B and C are virtually identical (at least I would find it difficult). I would be able to guess on visual inspection however that Group C had a higher mean than Group B. However, given the high values for Group A which might pull the mean up, I would not be able to guess that Group A's mean was less than Group C's mean.

5. Panel II is included here to prompt you to think about how your interpretation of whether data are disperse or not is affected by the possible values of the data. If I told you that the possible values for this variable for the 3 populations in Panel II ranged from 0 to 200, would you think that the differences in standard deviation among the three groups indicate important differences in dispersion? What about if I told you that the possible values for this variable were 101 to 120?

Figure 1: Distributions of Samples A, B, and C

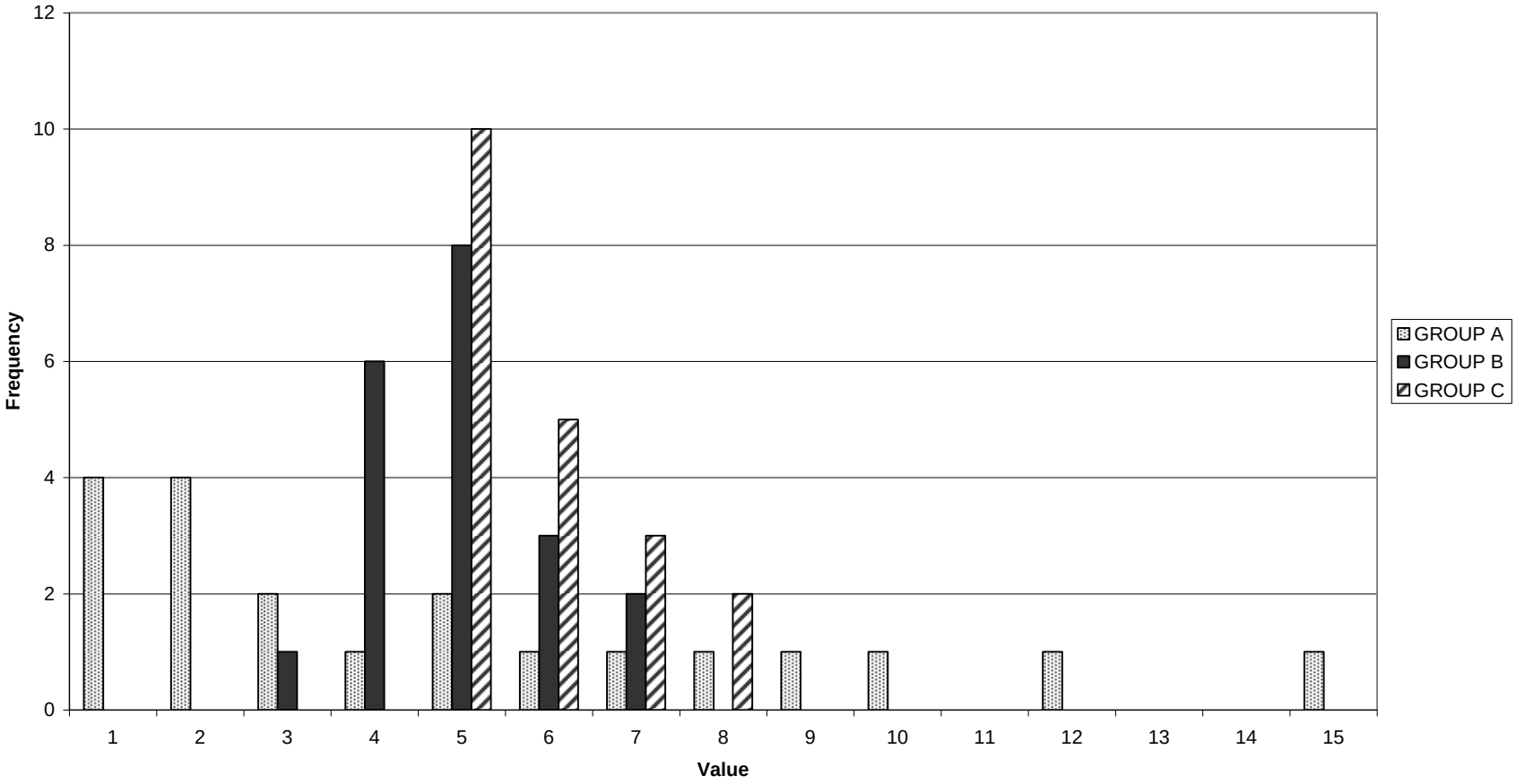

Figure 2: Distributions of Samples D, E, and F