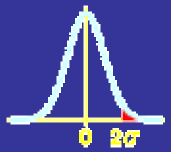
Measures of Central Tendency

SOCY601—Alan Neustadtl

Measures of Central Tendency

- A *measure of central tendency* is a single number used to represent the "center" of a group of data.
- Different variables may possess different numerical characteristics. So different measures of central tendency may better summarize the variable. The basic measures are the:
  - mode
  - median
  - mean
- This class of measures can be calculated on grouped or ungrouped data. The difference is in how the data values are weighted.



The Mode

- The mode is the:
  - most frequently occurring value in a group or raw scores
  - value of the group that contains the most cases in grouped data

The Mode—Example

Frequency Distribution of Sex in the 2000 General Social Survey

Race	f
White	2,244
Black	404
Other	170
Total	*2,817*

The Median

- The median is defined as the middle value (case) of n values of X objects arranged in order of size.
- For an *odd* number of cases, the middle case will be equal to the $\frac{n+1}{2}$ case.
- For an *even* number of cases, the middle case will be halfway between the $\frac{n}{2}$ and the $\frac{n}{2}+1$ case.

The Median—Example

Rank	Name	Interlocking Directorates
1	Equitable Life Insurance	75
2	Morgan Guaranty Trust	72
3	Chemical Bank of New York	70
4	First National Bank	69
5	Chase Manhattan Bank	66

The Median—Example

Rank	Name	Interlocking Directorates
1	Equitable Life Insurance	75
2	Morgan Guaranty Trust	72
3	Chemical Bank of New York	70
4	First National Bank	69
5	Chase Manhattan Bank	66
6	New York Life	61

The Median—Grouped Data

Stated Limits	True Limits	f	F	Number of Cases Less Than:
2,000 - 2,900	1,950 - 2,950	17	17	$2,950
3,000 - 3,900	2,950 - 3,950	26	43	$3,950
4,000 - 4,900	3,950 - 4,950	38	81	$4,950
5,000 - 5,900	4,950 - 5,950	51	132	$5,950
6,000 - 6,900	5,950 - 6,950	36	168	$6,950
7,000 - 7,900	6,950 - 7,950	21	189	$7,950

Look for the interval containing the median or the $\frac{n}{2}$ case.

$$\frac{n}{2} = \frac{189}{2} = 94.5$$

The Median—Grouped Data

Stated Limits	True Limits	*F*	F	Number of Cases Less Than:
2,000 - 2,900	1,950 - 2,950	17	17	$2,950
3,000 - 3,900	2,950 - 3,950	26	43	$3,950
4,000 - 4,900	3,950 - 4,950	38	81	$4,950
5,000 - 5,900	4,950 - 5,950	51	132	$5,950
6,000 - 6,900	5,950 - 6,950	36	168	$6,950
7,000 - 7,900	6,950 - 7,950	21	189	$7,950

There are 51 cases in this interval. We divide the interval into 51 equal sub-intervals equal to $19.61.

$$\frac{\$1,000}{51} = \$19.61$$

The Median—Grouped Data

Then we simply count the sub-intervals from the lower class limit until we come to the median.

We could also get this number by subtracting 81 from 94.5, the location of the median.

The Median—Grouped Data

$$md = l + \left(\frac{\frac{n}{2} - F}{f} \right) i$$

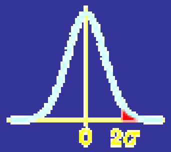
Where: l=lower limit of the interval containing the median

F=cumulative frequency corresponding to the lower limit

f=number of cases in the interval containing the median

i=width of the interval containing the median

The Mean



For Populations:

$$\mu = \frac{X_1 + X_2 + \ldots X_N}{N}$$

For Samples:

$$\bar{X} = \frac{X_1 + X_2 + \ldots X_n}{n}$$

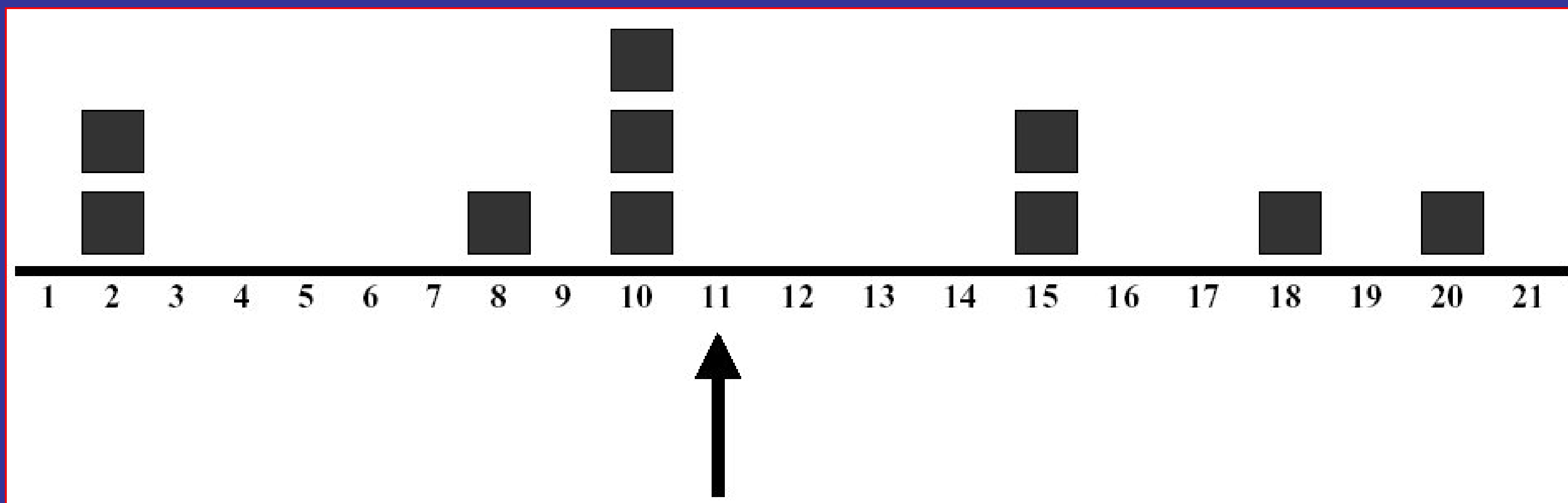
The Mean

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

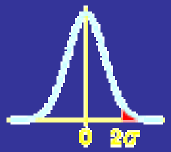
$$\bar{X} = \frac{\sum X}{n}$$

$$\bar{X} = \frac{1}{n}\sum X$$

The Mean as the Center of Gravity



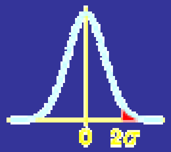
Properties of the Mean



- The mean has the algebraic property that the sum of the deviations of each score from the mean will always be zero. Symbolically:

$$\sum_{i=1}^{n}\left(X_i - \bar{X}\right) = 0$$

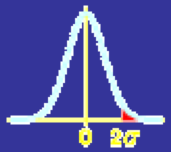
Properties of the Mean



- The sum of the squared deviations of each score from the mean is less than the sum of the squared deviations from any other constant (number). Symbolically:

$$\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \text{minimum}$$

Proof that: $\sum\left(X-\bar{X}\right)=0$



Given:

$$\sum\left(X-\bar{X}\right)$$

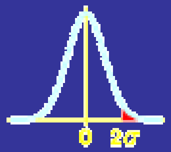
By distribution, we can rewrite this expression as:

$$\sum X_i-\sum\bar{X}$$

The mean is a constant. The sum of a constant is equal to *n* times that constant. So, we can rewrite this expression as:

$$\sum X_i-n\bar{X}$$

Proof that: $\sum\left(X-\bar{X}\right)=0$



The mean is a constant. The sum of a constant is equal to n times that constant. So, we can rewrite this expression as:

$$\sum X_i - n\bar{X}$$

We also know the basic definition of the mean and can substitute it:

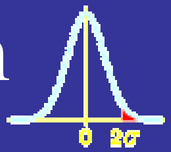
$$\sum X - n\left(\frac{\sum X}{n}\right)$$

The n's cancel:

$$\sum X - \cancel{n}\left(\frac{\sum X}{\cancel{n}}\right)$$

$$\sum X - \sum X = 0$$

Sum of Squared Deviations About the Mean



The logic of this proof is that if we subtract any number <u>other</u> than the mean from each value of *X*, square that amount, and sum up these values for all values of *X*, we will get a number that is larger than if we had carried out the same procedure using the mean of *X*.

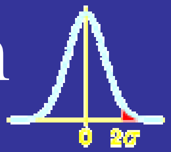
Let's call this other number "X-bar prime" and start with the original expression:

$$\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$$

However, we will strip out the summation and exponentiation, and substitute "X-bar prime" for the mean:

$$\left(X_i-\bar{X}'\right)$$

Sum of Squared Deviations About the Mean



Now we can add and simultaneously subtract the actual mean to and from this expression. This has no "net" effect on the expression. This is equal to:

$$\left(\bar{X}-\bar{X}'\right)=\left(X_i-\bar{X}\right)+\left(\bar{X}-\bar{X}'\right)$$

Squaring both sides of the expression brings us a step closer to the original equation:

$$\left(\bar{X}-\bar{X}'\right)^2=\left[\left(X_i-\bar{X}\right)+\left(\bar{X}-\bar{X}'\right)\right]^2$$

Because $(a+b)^2=a^2+2ab+b^2$, when expanded this is equal to:

Sum of Squared Deviations About the Mean

Because $(a+b)^2 = a^2 + 2ab + b^2$, when expanded this is equal to:

$$\left(\bar{X}-\bar{X}'\right)^2 = \left(X_i-\bar{X}\right)^2 + 2\left(X_i-\bar{X}\right)\left(\bar{X}-\bar{X}'\right) + \left(\bar{X}-\bar{X}'\right)^2$$

Now, add the summation symbol back in on both sides of the expression and with a little bit of algebraic manipulation we get:

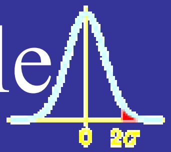
$$\begin{aligned}
\sum\left(\bar{X}-\bar{X}'\right)^2 &= \sum\left(X_i-\bar{X}\right)^2 + \sum 2\left(X_i-\bar{X}\right)\left(\bar{X}-\bar{X}'\right) + \sum\left(\bar{X}-\bar{X}'\right)^2 \\
&= \sum\left(X_i-\bar{X}\right)^2 + 2\left(\bar{X}-\bar{X}'\right)\sum\left(X_i-\bar{X}\right) + \sum\left(\bar{X}-\bar{X}'\right)^2 \\
&= \sum\left(X_i-\bar{X}\right)^2 + \sum\left(\bar{X}-\bar{X}'\right)^2 \\
&= \sum\left(X_i-\bar{X}\right)^2 + n\left(\bar{X}-\bar{X}'\right)^2
\end{aligned}$$

The Mean from Grouped Data

Ungrouped Data	Grouped Data
$\overline{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$	$\overline{X} = \dfrac{\sum_{i=1}^{n} w_i X_i}{\sum w_i}$

- The only difference in these formulas is the weight, w_i.
- With ungrouped data, the weight is implicitly equal to one.

The Mean from Grouped Data—Example



Type of PAC	Contributions	N	Weighted
Corporations	$37,100.78	1,875	$69,563,962.50
Labor	$112,757.12	371	$41,832,891.52
Non-Connected	$13,832.66	1,318	$18,231,445.88
T/M/H	$62,029.20	852	$52,848,878.40
Cooperative	$54,178.63	56	$3,034,003.28
W/O Stock	$27,272.09	149	$4,063,541.41
Sums	$307,170.48	4,621	$189,574,722.99
Averages	$51,195.08		$41,024.61

The Mean from Grouped Data—Example

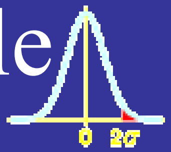
Annual Income of American Men in 1975

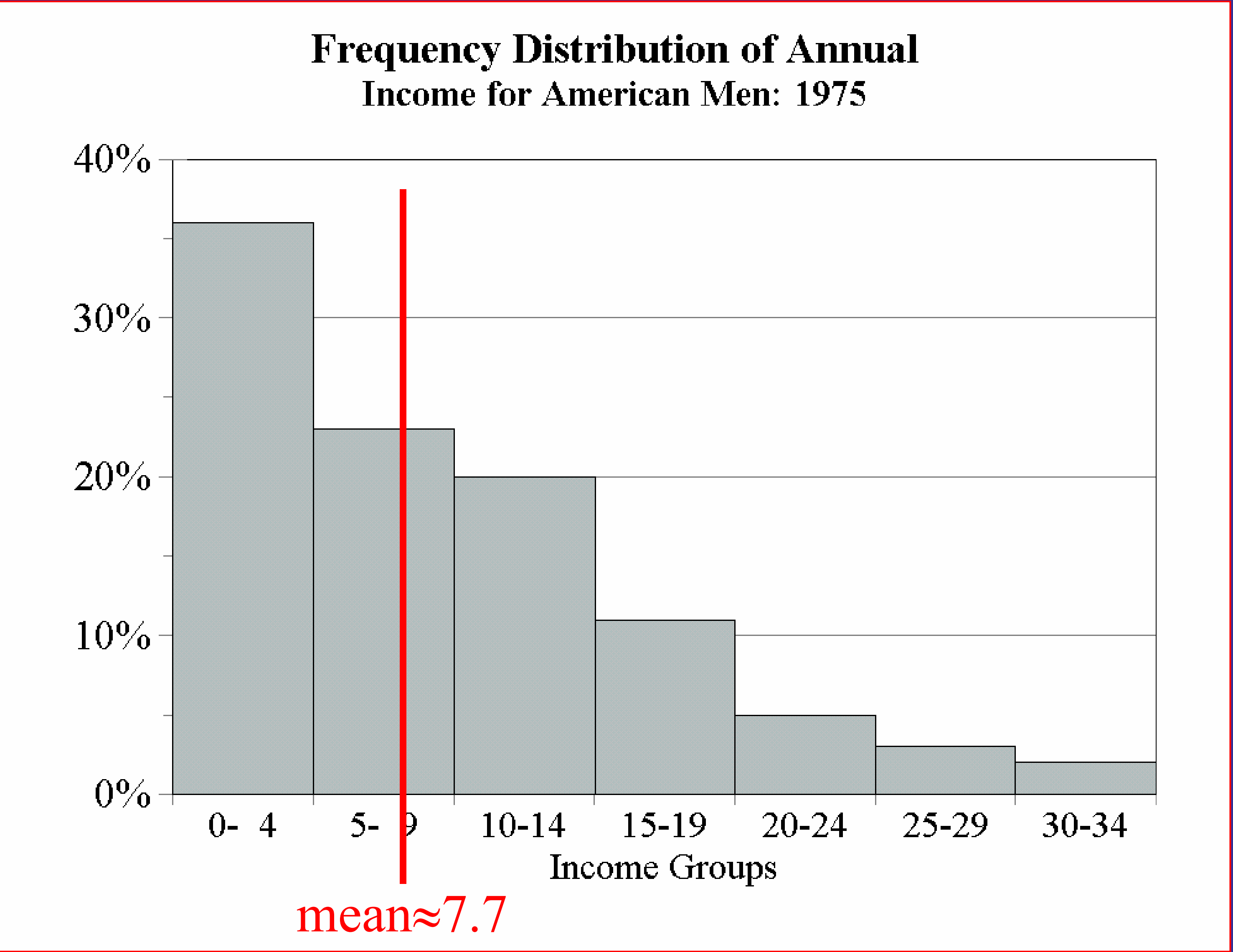
Annual Income	Percent of Men	Midpoint	Weighted by Midpoint
0- 4	36%	2	72
5- 9	23%	7	161
10-14	20%	12	240
15-19	11%	17	187
20-24	5%	22	110
25-29	3%	27	81
30-34	2%	32	64
		119	915

average=7.69

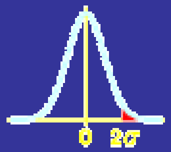
$$\overline{X} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum w_i}$$

The Mean from Grouped Data—Example





Percentiles



- A *percentile* is the outcome or score below which a given percentage of observations fall,

$$P_i = L_p + \left(\frac{(p_i)(n_i) - c_p}{f_p} \right) W_i$$

Where: P_i=the score if the ith percentile

L_p=the true lower limit of the interval containing the ith percentile

p_i= the ith percentile written as a proportion (e.g. 75th = 0.75)

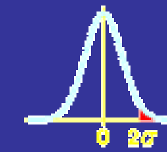
n=the total number of observations

c_p=the cumulative frequency up to but *not* including the interval containing P_i

f_p=the frequency in the interval containing the ith percentile

W_i=the width of the interval containing P_i; $W=U_p-L_p$

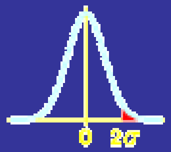
Percentiles



Stated Limits	True Limits	f	F	Number of Cases Less Than:
2,000 - 2,900	1,950 - 2,950	17	17	$2,950
3,000 - 3,900	2,950 - 3,950	26	43	$3,950
4,000 - 4,900	3,950 - 4,950	38	81	$4,950
5,000 - 5,900	4,950 - 5,950	51	132	$5,950
6,000 - 6,900	5,950 - 6,950	36	168	$6,950
7,000 - 7,900	6,950 - 7,950	21	189	$7,950

$$P_{50} = \$4,950 + \left(\frac{(.5)(189) - 81}{51} \right) 1,000 = \$5,214.70$$

Percentiles



Stated Limits	True Limits	f	F	Number of Cases Less Than:
2,000 - 2,900	1,950 - 2,950	17	17	$2,950
3,000 - 3,900	2,950 - 3,950	26	43	$3,950
4,000 - 4,900	3,950 - 4,950	38	81	$4,950
5,000 - 5,900	4,950 - 5,950	51	132	$5,950
6,000 - 6,900	5,950 - 6,950	36	168	$6,950
7,000 - 7,900	6,950 - 7,950	21	189	$7,950

$$P_{75} = \$5,950 + \left(\frac{(.75)(189) - 132}{36} \right) 1,000 = \$5,220.83$$