Chapter 2

Describing Variables

2.5 Measures of Dispersion

Measures of Dispersion

Measures of dispersion indicate the amount of variation or "average differences" among the scores in a frequency distribution.

We're less familiar with such concepts in daily life, although a <u>range of values</u> is sometimes reported:

- **Today's forecast high temp will be 59-62 degrees**
- **N. Korea's Taepodong missile has a reported range of 2,400 to 3,600 miles**
- **Gallup Poll reported 51% of a national sample agree that President Obama is doing a good job, with a "margin of error" of ±3%**

Discrete Variable Dispersion Measures

Index of Diversity (D) measures whether two randomly selected observations are likely to fall into the same or different categories

$$D = 1 - \sum_{i=1}^{K} p_i^2$$

Higher D indicates the cases are more equally spread across a variable's *K* categories (i.e., they are less concentrated)

Calculate D for these four GSS regions of residence:

Region	p_i	$(p_i)^2$
NORTH EAST	.175	________
MIDWEST	.215	________
SOUTH	.361	________
WEST	.248	________

$$\sum_{i=1}^{K} p_i^2 = \underline{\qquad\qquad\qquad}$$

$$D = 1 - \Sigma p_i^2 = \underline{\qquad\qquad\qquad\qquad\qquad\qquad}$$

The **Index of Qualitative Variation (IQV)** adjusts D for the number of categories, *K*

$$\mathrm{IQV} = \frac{\mathrm{K}}{\mathrm{K}-1}(\mathrm{D})$$

IQV gives a bigger “boost” to *D* for a variable with fewer categories, thus allowing comparison of its dispersion to a variable that has more categories

Sally and three friends buy a 12-pack of beer (144 oz.). Ted and seven friends buy two 12-packs (288 oz.). Which distribution of beer is “fairer” (more equally distributed within each set of drinkers)?

Sally: **20, 28, 46, 50 oz.**

Ted: **20, 28, 32, 36, 40, 40, 44, 48 oz.**

$$IQV = \left(\frac{K}{K-1}\right)\left(1 - \sum_{i=1}^{K} p_i^2\right)$$

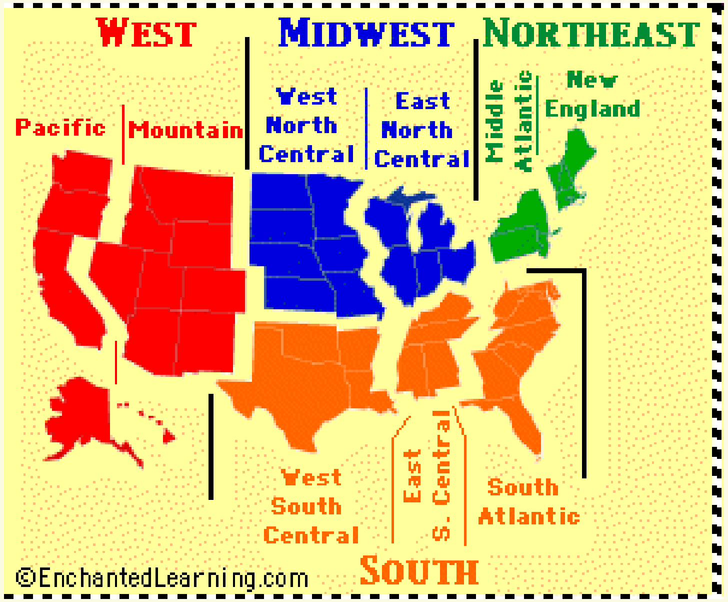
IQV_{Sally} = ______________________________

IQV_{Ted} = ______________________________

Indices of Diversity for proportions of U.S. population living in 4 Census regions and the distribution in 9 Census regions:

Four-region D = 0.731

Nine-region D = 0.855



The population seems more equally spread among the 9 regions than among the 4 regions. However, …

calculate the IQVs for both measures. Do these two population distributions now seem differently dispersed?

Four-region IQV = ______________________

Nine-region IQV = ______________________

Range the difference between largest and smallest scores in a continuous variable distribution

What are the ranges for these GSS variables?

	Min.-Max.	**Range**
EDUC:	0 to 20 years	___________
AGE:	18 to 89 years	____________
PRESTG80:	17 to 86 points	____________
PAPRES80:	17 to 86 points	____________

Average Absolute Deviation (AAD)

Read this subsection (pp. 48-49) for yourself, as background info for the variance & standard deviation

Because ADD is never used in research statistics, we won’t spend any time on it in lecture

Variance and Standard Deviation

Together with the mean, the variance (and its kin, the standard deviation) are the workhorse statistics for describing continuous variables

Variance the mean (average) squared deviation of a continuous distribution

The deviation (*d_i*) of case i is the difference between its score Y_i and the distribution's mean:

$$d_i = Y_i - \bar{Y}$$

To calculate the variance of a sample of N cases:

- Compute and square each deviation
- Add them up
- Divide the sum by N - 1

$$s_Y^2 = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}{N-1} = \frac{\sum d_i^2}{N-1}$$

Reason for using N-1, not N, will be explained later.

Standard deviation the positive square root of the variance

This transformation avoids the unclear meaning of squared measurement units; e.g., years-squared

The standard deviation of a sample:

$$s_Y = \sqrt{s_Y^2}$$

Calculate s^2 and s for these 10 scores

Y_i	$-$	$\bar{Y}$	$=$	d_i	$(d_i)^2$
2	-	2	=	______	______
0	-	2	=	______	______
4	-	2	=	______	______
1	-	2	=	______	______
6	-	2	=	______	______
3	-	2	=	______	______
1	-	2	=	______	______
2	-	2	=	______	______
1	-	2	=	______	______
0	-	2	=	______	______

$$\sum_{i=1}^{10} (d_i)^2 = \_\_\_\_\_\_\_\_\_\_\_\_$$

$$s_Y^2 = \sum (d_i)^2 \;/\; (N-1) =$$

$$\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

$$s_Y = \sqrt{s_Y^2} = \_\_\_\_\_\_\_\_\_\_\_\_$$

To calculate the variance of a dichotomy, just multiply both proportions:

$$s_Y^2 = (p_0)(p_1)$$

The 2008 GSS asked, "Do you favor or oppose the death penalty for persons convicted of murder?" What is its variance?

CAPPUN	p_i
1 FAVOR	.66
0 OPPOSE	.34

$$s_Y^2 = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

A item about having ever used crack cocaine was split more unevenly. Is its variance larger or smaller than CAPPUN's?

EVCRACK	p_i
1 YES	.06
0 NO	.94

$$s_Y^2 = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

Variance of a Grouped Frequency Distribution

Use the variance formula but multiply each squared deviation by its relative frequency (f_i), then sum the products across all *K* categories:

$$s_Y^2 = \frac{\sum_{i=1}^{K} (Y_i - \overline{Y})^2 (f_i)}{N-1} = \frac{\sum (d_i^2)(f_i)}{N-1}$$

What is the variance of these grouped data?

HOMOSEX1 “What about sexual relations between two adults of the same sex; is it …”

[Mean = 2.15 for N = 1,309]

Response	Y_i	f_i	$(d_i)^2(f_i)$
Always wrong	1	733	______________
Almost always	2	67	______________
Only sometimes	3	88	______________
Not wrong at all	4	421	______________

$$s_Y^2 = \frac{\sum_{i=1}^{K} (d_i)^2 (f_i)}{N-1} = \_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

Skewness describes nonsymmetry (lack of a mirror-image) in a continuous distribution

It compares the mean and the median:

$$\mathbf{Skewness} = \frac{3(\overline{\mathbf{Y}} - \mathbf{Mdn})}{\mathbf{s_Y}}$$

- Positive skew has a "tail" to right of Mdn
- Negative skew has a "tail" to left of Mdn

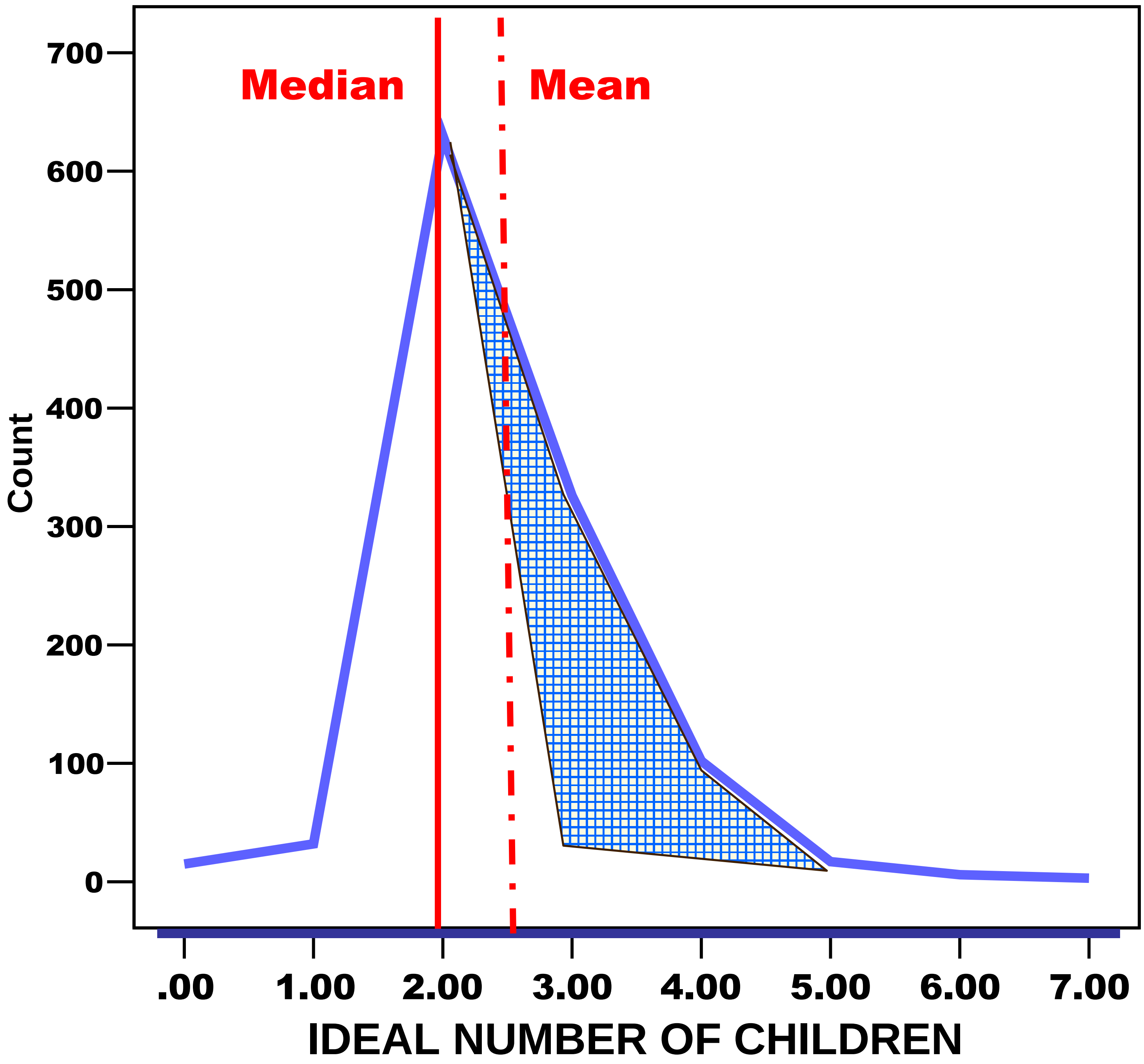
For most continuous variables, a positively skewed distribution typically has a mean much larger than its median. A negatively skewed distribution typically has a mean smaller than its median.

<u>U.S. household income</u> is positively skewed: in 2006 the median was $48,201 but the mean was $66,570. What produced this gap?

The 2008 GSS asked, "What do you think is the ideal number of children for a family to have?"

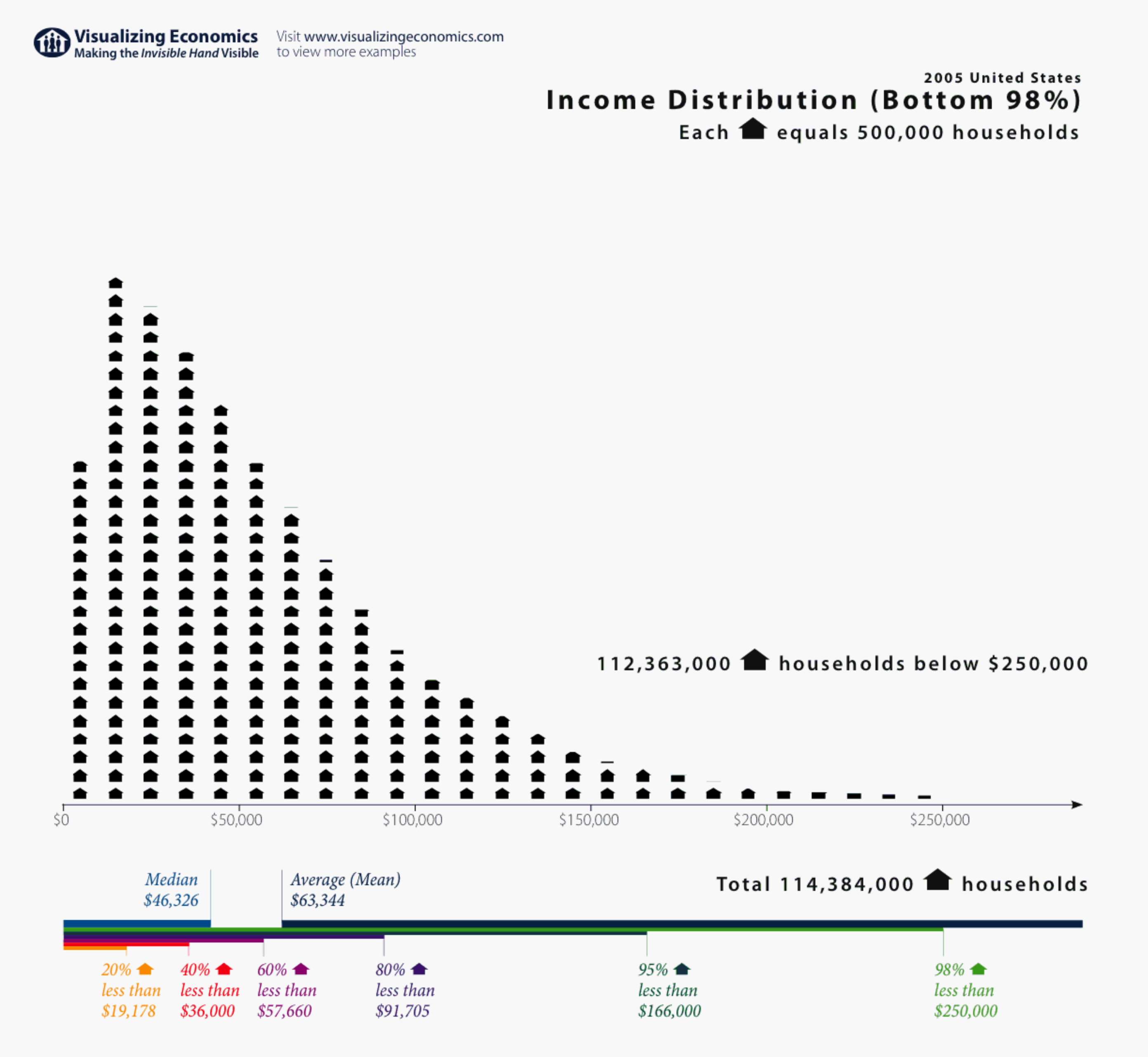
Mdn = 2.00 Mean = 2.49 Std dev = 0.88 N = 1,131

Skewness = ____________



Median
Mean
Count
700
600
500
400
300
200
100
0
.00
1.00
2.00
3.00
4.00
5.00
6.00
7.00
IDEAL NUMBER OF CHILDREN

What type of skew does this income distribution have?



Calculate s^2 and s for these 8 ungrouped scores

Y_i	$-$	$\overline{Y}$	$=$	d_i	$(d_i)^2$
1	**-**	**5**	**=**	______	______
3	**-**	**5**	**=**	______	______
4	**-**	**5**	**=**	______	______
5	**-**	**5**	**=**	______	______
6	**-**	**5**	**=**	______	______
6	**-**	**5**	**=**	______	______
7	**-**	**5**	**=**	______	______
8	**-**	**5**	**=**	______	______

$$\sum_{i=1}^{8}(d_i)^2 = \_\_\_\_\_\_\_\_\_\_\_\_$$

$$s_Y^2 = \sum(d_i)^2 \;/\; (N-1) =$$

____

$$s_Y = \sqrt{s_Y^2} = \_\_\_\_\_\_\_\_\_\_\_\_$$

Calculate variance & standard deviation of NATEDUC

"Are we spending too much money, too little money, or about the right amount on the nation's education system?"

N = 993 **Mean = 1.34**

Category	Y_i	f_i	$(d_i)^2(f_i)$
TOO LITTLE	1	707	______
ABOUT RIGHT	2	232	______
TOO MUCH	3	54	______

$$\sum_{i=1}^{K}(d_i)^2(f_i) = \_\_\_\_\_\_\_\_$$

$$s_Y^2 = \frac{\sum_{i=1}^{K}(d_i)^2(f_i)}{N-1} = \_\_\_\_\_\_\_\_\_\_\_\_$$

$$s_Y = \sqrt{s_Y^2} = \_\_\_\_\_\_\_\_$$

Calculate variance & standard deviation of SEXFREQ

N = 1,686 **Mean = 57.3**

Category	Y_i	f_i	$(d_i)^2(f_i)$
NOT AT ALL	0	416	______
ONCE OR TWICE	2	149	______
ONCE A MONTH	12	176	______
2-3 per MONTH	36	243	______
WEEKLY	52	285	______
2-3 per WEEK	156	309	______
3+ per WEEK	208	108	______

$$\sum_{i=1}^{K} (d_i)^2 (f_i) = \_\_\_\_\_\_\_\_\_\_$$

$$s_Y^2 = \frac{\sum_{i=1}^{K} (d_i)^2 (f_i)}{N-1} = \_\_\_\_\_\_\_\_\_\_$$

$$s_Y = \sqrt{s_Y^2} = \_\_\_\_\_\_\_\_\_\_$$