Graphs versus Numbers

Using either graphs or numbers, we primarily want to describe the central tendency and spread of a distribution

Central tendency

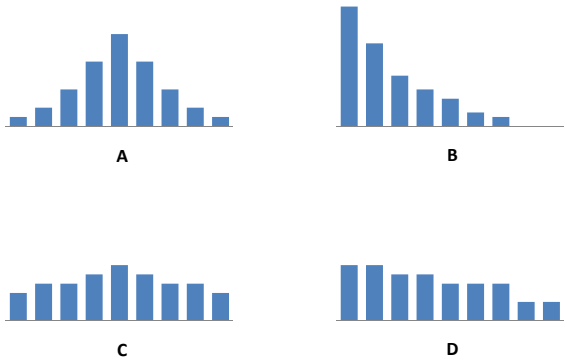
Where do the bulk of cases fall on the distribution?

Where is the middle of the distribution? (Is there one?)

Spread (or **Variation** or **Dispersion**)

How tightly clumped together are cases, especially relative to the center of the distribution?

Graphs versus Numbers



Graphs versus Numbers

Unit of Observation: Presidents of the United States

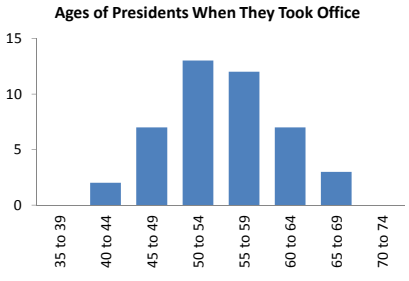
Variable: Age when they took office

President	Age at Inauguration
George Washington	57
John Adams	61
…	
George W. Bush	54
Barack Obama	47

Grover Cleveland

Graphs versus Numbers

Always start by looking at a picture of the distribution



Graphs versus Numbers

Many of the numeric summaries of central tendency and spread pertain only to *continuous* variables

Central Tendency	Continuous Variables	Discrete Variables
Mean	✓	×
Median	✓	×
Mode	✓	✓

Spread	Continuous Variables	Discrete Variables
Range	✓	×
Interquartile Range	✓	×
Variance	✓	×
Standard Deviation	✓	×
Index of Diversity	×	✓
Index of Qualitative Variation	×	✓

Central Tendency I: Mean

The **mean** (or **average**) is computed by adding together (or summing) the value of the continuous variable across all individuals and then dividing by the number of individuals

Mathematically:

$$\bar{Y} = \frac{Y_1 + Y_2 + Y_3 + \ldots Y_N}{N} = \frac{\sum_{i=1}^{N} Y_i}{N}$$

where Y_i represents the value of the variable, *Y*, for individual *i*, "*Y-bar"* is the mean, and *N* is the number of individuals in the data

Central Tendency I: Mean

Example: The **mean** age of U.S. Presidents when they took office equals:

$$\bar{Y} = \frac{57+61+\ldots+54+47}{N} = \frac{2{,}405}{44} = 54.7$$

Central Tendency II: Median

Median

- The point on the distribution at which exactly half of all individuals fall above that point and half fall below that point
- The point on the distribution at which the cumulative percentage reaches 50%

To compute the median, or *M*, arrange the observations in order of size from lowest to highest

- If the number of observations, *N*, is odd, then the median, *M*, is the middle observation
- If the number of observations, *N*, is even, then the median, *M*, is the mean of the middle two observations

Central Tendency II: Median

Example: Ages of U.S. Presidents when they took office:

42 43 46 46 47 47 48 49 49 50 51 51 51 51 51 52 52 54 54 54 54 54 55 55 55 55 56 56 56 57 57 57 57 58 60 61 61 61 62 64 64 65 68 69

Because the number of observations (44) is an even number, the median is the average of 22nd and 23rd cases

So, M = (54+55)/2 = 54.5

Central Tendency II: Median

Odd number of observation…

3 4 7

…median is 4

Even number of observations…

3 4 7 9

…median is (4+7)/2, or 5.5

Central Tendency III: Mode

Mode

The single value (or values) that appears most often in the distribution

A distribution with two modes is referred to as "bimodal"

For the ages of U.S. Presidents when they took office, the values 51 and 54 appear more frequently (5 times each) than any other value

42 43 46 46 47 47 48 49 49 50 51 51 51 51 51 52 52 54 54 54 54 54 55 55 55 55 56 56 56 57 57 57 57 58 60 61 61 61 62 64 64 65 68 69

Central Tendency: Outliers

Outliers … or extreme values … can strongly influence the mean, but not the mode or median

4 4 6 8

Median: 5 *Mode*: 4 *Mean*: 5.5

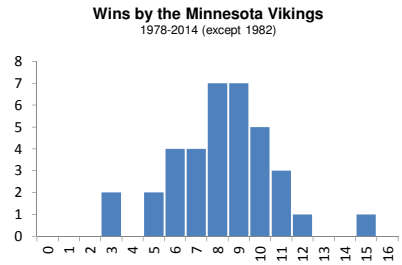
4 4 6 80

Median: 5 *Mode*: 4 *Mean*: 23.5

4 4 6 800

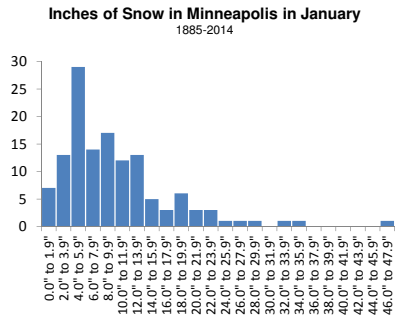
Median: 5 *Mode*: 4 *Mean*: 203.5

Measures of Central Tendency



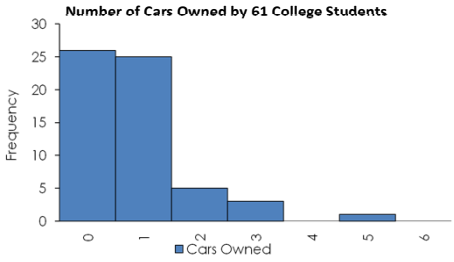
Mean: **8.3** **Median**: **8** **Mode**: **8 & 9**

Measures of Central Tendency



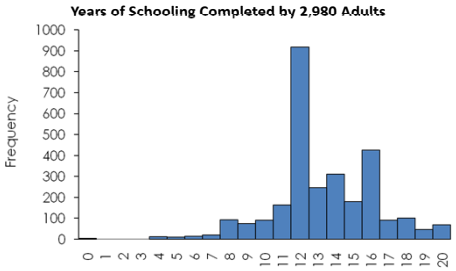
Mean: **10.0”** **Median**: **8.3”** **Mode**: **4.6”**

Measures of Central Tendency



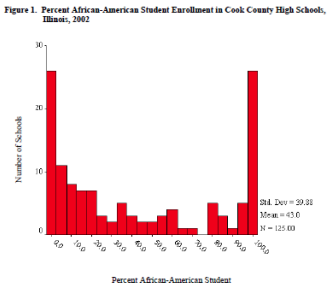
Mean: **0.75** **Median**: **1** **Mode**: **0**

Measures of Central Tendency



Mean: **13.2** **Median**: **13** **Mode**: **12**

Measures of Central Tendency



Worksheet

I surveyed 20 graduate students and asked them how many pets they own. Eight of them had no pets at all. Six had just one pet, five had two pets, and one had four

1. What is the mean of this distribution?
2. What is the median?
3. What is the mode?

Measures of Spread I: Range

Range

The difference between the largest and the smallest observation

Example: Ages of U.S. Presidents when they took office:

42 43 46 46 47 47 48 49 49 50 51 51 51 51 51 52 52 54 54 54 54 54 55 55 55 55 56 56 56 57 57 57 57 58 60 61 61 61 62 64 64 65 68 69

In this example, the range is the difference between 42 and 69 … so, 69 – 42 = 27

Measures of Spread II: Inter-Quartile Range

The **inter-quartile range** is the range in the distribution that marks out the middle 50% of the distribution

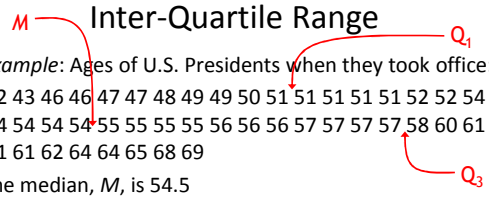
- 25% of the distribution falls below the first quartile
- 25% of the distribution falls above the third quartile

To calculate the inter-quartile range:

1. Sort the value from lowest to highest and find the median, *M*
2. The 1st quartile (Q_1) is the median of the values to the left of *M*
3. The 3rd quartile (Q_3) is the median of the values to the right of *M*

The inter-quartile range (or IQR) equals Q_3 minus Q_1

Measures of Spread II: Inter-Quartile Range



- *Example*: Ages of U.S. Presidents when they took office:
  42 43 46 46 47 47 48 49 49 50 51 51 51 51 51 52 52 54
  54 54 54 54 55 55 55 55 56 56 56 57 57 57 57 58 60 61
  61 61 62 64 64 65 68 69
- The median, *M*, is 54.5
- The 1st quartile, Q_1, is the median of the 22 values below the median … so, (51+51)/2 = 51
- The 3rd quartile, Q_3, is the median of the 22 values above the median … so, (57+58)/2 = 57.5
- The inter-quartile range is 57.5 – 51 = 6.5

Measures of Spread III: Variance & Standard Deviation

Variance (s^2)

Basically, the average of the squared differences between each observation and the mean

$$s_Y^2 = \frac{(Y_1-\bar{Y})^2 + (Y_2-\bar{Y})^2 + \ldots + (Y_N-\bar{Y})^2}{N\text{-}1} = \frac{\sum_{i=1}^{N}(Y_i-\bar{Y})^2}{N-1}$$

Standard Deviation (s)

The square root of the variance

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

Measures of Spread III: Variance & Standard Deviation

Example: Variance (s^2) of the age of U.S. Presidents

$$s_Y^2 = \frac{\sum_{i=1}^{44}(Y_i-54.659)^2}{44-1} = \frac{1{,}683.886}{43} = 39.160$$

Example: Standard deviation (s) of age of U.S. Presidents

$$s_Y = \sqrt{\frac{\sum_{i=1}^{N}(Y_i-\bar{Y})^2}{N-1}} = \sqrt{39.160} = 6.258$$

Measures of Spread III: Variance & Standard Deviation

Example: 4 4 6 8

Y	Y – Y	(Y – Y)²
4	(4 – 5.5) = -1.500	$(4-5.5)^2 = 2.250$
4	(4 – 5.5) = -1.500	$(4-5.5)^2 = 2.250$
6	(6 – 5.5) = 0.500	$(6-5.5)^2 = 0.250$
8	(8 – 5.5) = 2.500	$(8-5.5)^2 = 6.250$
		$\Sigma(Y-Y)^2 = 11.000$

so s^2 equals 11/(*N*-1) = 11/3 = 3.667

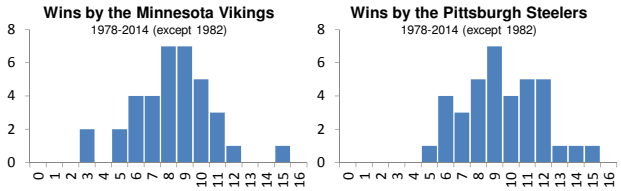
Measures of Spread III: Variance & Standard Deviation

Variance has no inherent unit of measurement; standard deviations are in the same unit of measurement as the original observations

Variance and standard deviation measure spread around the mean, not spread around the median (or the mode)

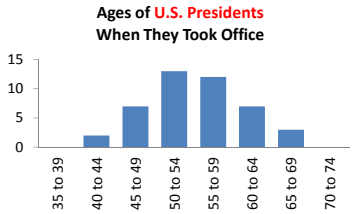
Because means are strongly affected by outliers, and because s and s^2 are computed using the mean, s and s^2 are also strongly affected by outliers

Measures of Spread

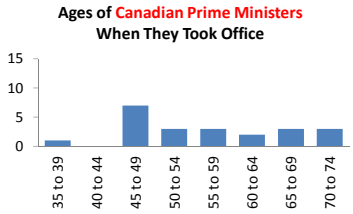


Mean = 8.3
Range = 12 (3 to 15)
IQR = 3 (7 to 10)
s^2 = 5.85
s = 2.42

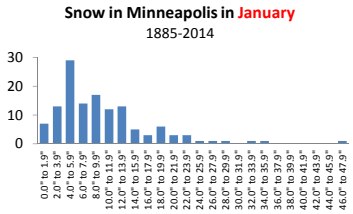
Mean = 9.6
Range = 10 (5 to 15)
IQR = 3 (8 to 11)
s^2 = 5.57
s = 2.36



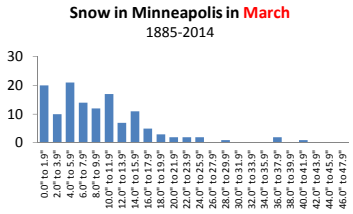
Mean: 54.7
Range: 27 (42 to 69)
IQR: 6.5 (51 to 57.5)
s^2: 39.2
s: 6.3



Mean: 55.6
Range: 35 (39 to 74)
IQR: 18 (47 to 65)
s^2: 95.6
s: 9.8



Mean: 10.0
Range: 45.8 (0.6 to 46.4)
IQR: 7.9 (4.9 to 12.8)
s^2: 55.4
s: 7.4



Mean: 9.3
Range: 40 (0 to 40)
IQR: 9.3 (4.1 to 13.4)
s^2: 57.0
s: 7.6

Worksheet

I surveyed 5 people and asked them how many brothers and sisters they have. The five people had zero, zero, two, three, and five siblings, respectively

1. What is the range of this distribution?
2. What is the variance?
3. What is the standard deviation?

Measuring Skew

For highly skewed distributions the mean may not accurately reflect central tendency; in these situations, many people prefer to use the median to summarize central tendency

How do you know when there is a "lot" of skew?

$$Skewness = \frac{3(\bar{Y}\text{-}Median)}{s_Y}$$

(Note: This is just one measure of skew. Lots of others exist.)

Measuring Skew

4 4 6 8

Median: 5 Mean: 5.5 s_Y: 1.91 Skew: 0.78

4 4 6 800

Median: 5 Mean: 203.5 s_Y: 397.67 Skew: 1.50

4 4 6 -800

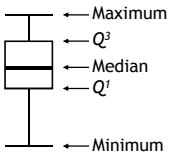
Median: 4 Mean: -196.5 s_Y: 402.33 Skew: -1.50

Summarizing Distributions I: Five Number Summary and Box Plot

One convenient way to describe the central tendency and spread of a distribution is to use five numbers to make a "box plot"

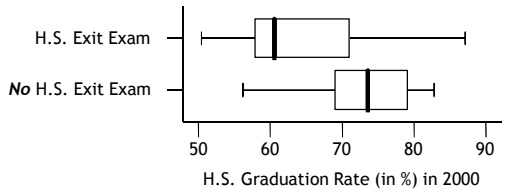
These numbers are the minimum value, Q_1, M, Q_3, and the maximum value

A box plot:



Summarizing Distributions I: Five Number Summary and Box Plot

Are state high school graduation rates related to whether states require students to pass "exit examinations" as a requirement for graduation? (All data as of 2000)



Summarizing Distributions I: Five Number Summary and Box Plot

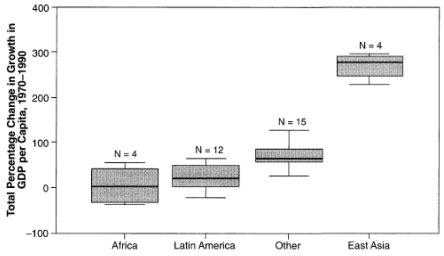
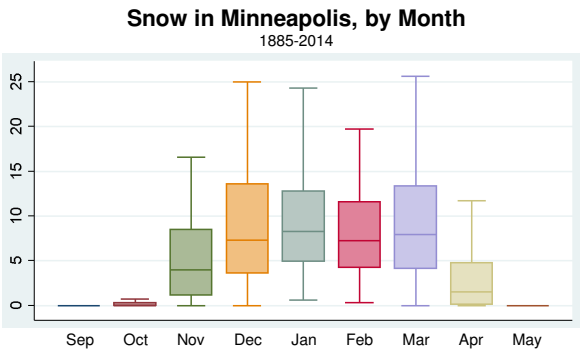


Figure 2. Boxplot of Growth in GDP per Capita by Region, 1970 to 1990

Source: Evans, Peter Evans and James E. Rauch. 1999. "Bureaucracy and Growth: A Cross-National Analysis of the Effects of 'Weberian' State Structures on Economic Growth Bureaucracy and Growth: A Cross-National Analysis of the Effects of "Weberian" State Structures on Economic Growth." *American Sociological Review* 64: 748-765.



Summarizing Distributions II: Mean and Standard Deviation

By far the most common way to summarize the central tendency and spread in a distribution is to use the **mean** and **standard deviation**

However, this way of describing a distribution only works well if the distribution is fairly symmetrical and if there are no outliers (or if outliers have been removed) … this is because the mean and standard deviation are heavily influenced by outliers and skew

Summarizing Distributions II: Mean and Standard Deviation

	Min	Max	$\bar{x}$ or $\hat{p}$	Standard Deviation
Dependent Variable				
Estimated high school completion rate %	49.0	97.6	75.0	(8.80)
Minimum Wage Variables				
In senior year – $6.25 (2005$)	-.92	2.18	.00	(.55)
Mean in junior & senior years – $6.23 (2005$)	-.82	2.16	.00	(.55)
Mean in four high school years – $6.19 (2005$)	-.71	2.51	.00	(.58)
Any change in junior or senior years? (yes = 1)	0	1	.23	
Any change in four high school years? (yes = 1)	0	1	.49	
State Education Policy Variables				
Maximum Age of Compulsory Schooling				
16 or less (yes = 1)	0	1	.64	
17 (yes = 1)	0	1	.16	
18 (yes = 1)	0	1	.20	
Carnegie Units Required for Graduation				
No statewide requirement	0	1	.11	
Fewer than 18	0	1	.23	
18 to 22	0	1	.45	
More than 22	0	1	.21	
State High School Exit Examination Policy				
No state exit exam	0	1	.71	
"Minimum Competency" state exit exam	0	1	.16	
"More Challenging" state exit exam	0	1	.13	
State Unemployment Rate – 5.8% (%)	-3.5	11.6	.0	(2.0)

Note: Sample of 1,224 state-years constructed by cross-classifying the 51 states (including DC) by the 24 years from 1982 through 2005 (inclusive). See text for description of variables.

John Robert Warren and Caitlin Hamrock. "The Effect of Minimum Wage Rates on High School Completion." *Social Forces* 88:1379-1392.

Table 1. Descriptive Statistics for Analytic Sample of Value-Added Model Predicting Student Achievement

Characteristic	*N*	*M*	*SD*	Min	Max
Student characteristics					
Female	95,988	0.487	NA	0	1
Black	95,988	0.560	NA	0	1
Hispanic	95,988	0.233	NA	0	1
White	95,988	0.125	NA	0	1
Asian	95,988	0.049	NA	0	1
Other race	95,988	0.032	NA	0	1
Special education	95,988	0.177	NA	0	1
Eligible for free or reduced-price lunch	95,988	0.800	NA	0	1
English language learner	95,988	0.088	NA	0	1
Reading score (z score)	95,988	0.033	0.977	−3.668	5.632
Math score (z score)	95,988	0.042	0.977	−4.191	4.672
School characteristics					
Percentage female	95,988	48.8	4.1	0.0	100.0
Percentage black	95,988	55.7	34.7	0.0	100.0
Percentage Hispanic	95,988	22.9	28.7	0.0	99.4
Percentage white	95,988	13.1	16.6	0.0	74.1
Percentage Asian	95,988	4.9	9.7	0.0	96.6
Percentage special education	95,988	17.1	5.8	0.0	81.8
Percentage eligible for free or reduced-price lunch	95,988	79.8	15.1	0.0	100.0
Percentage English language learner	95,988	8.0	12.0	0.0	65.9
Enrollment	95,988	384.6	224.6	1	983
Average reading score (z score)	95,988	0.017	0.381	−1.990	1.143
Average math score (z score)	95,988	0.018	0.345	−3.148	1.100

Deven Carlson and Joshua M. Cowen. 2015. "Student Neighborhoods, Schools, and Test Score Growth Evidence from Milwaukee, Wisconsin." *Sociology of Education* 88: 38-55

Summarizing Distributions

Many of the numeric summaries of central tendency and spread pertain only to *continuous* variables

Central Tendency	Continuous Variables	Discrete Variables
Mean	✓	×
Median	✓	×
Mode	✓	✓

Spread	Continuous Variables	Discrete Variables
Range	✓	×
Interquartile Range	✓	×
Variance	✓	×
Standard Deviation	✓	×
Index of Diversity	×	✓
Index of Qualitative Variation	×	✓

Summarizing Distributions

You interview 10 fellow students and ask them:

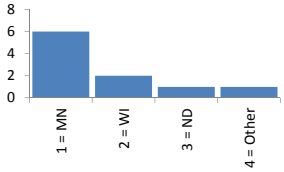
What is your home state?

1 = Minnesota
2 = Wisconsin
3 = North Dakota
4 = Someplace else

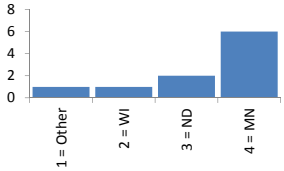
Here are the responses you obtained:

1 1 1 1 1 1 2 2 3 4

Summarizing Distributions



Mean: 1.7
Range: 3 (1 to 4)
IQR: 1 (1 to 2)
s^2: 1.122
s: 1.059



Mean: 3.2
Range: 3 (1 to 4)
IQR: 2 (2 to 4)
s^2: 1.289
s: 1.135

Summarizing Distributions

You interview 10 fellow students and ask them:

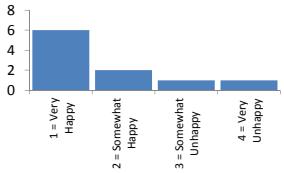
How happy are you?

1 = Very happy
2 = Somewhat happy
3 = Somewhat unhappy
4 = Very unhappy

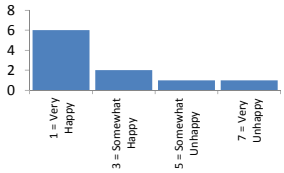
Here are the responses you obtained:

1 1 1 1 1 1 2 2 3 4

Summarizing Distributions



Mean: 1.7
Range: 3 (1 to 4)
IQR: 1 (1 to 2)
s^2: 1.122
s: 1.059



Mean: 2.4
Range: 6 (1 to 7)
IQR: 2 (1 to 3)
s^2: 4.489
s: 2.119

Measures of Spread for Discrete Variables

Index of Diversity (D)

Measure of spread (or dispersion) for discrete variables

Measures whether two observations selected randomly from a population are likely to fall into the same or different categories

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

where *K* indexes the categories in discrete variable *Y* and *p* is the proportion of observation in the i^{th} category

Measures of Spread for Discrete Variables

	Cat. 1	Cat. 2	Cat. 3	D
Y	10	10	10	0.667
Z	20	5	5	0.500

The higher the value of D, the more dispersed the cases across categories (and thus the less variability there is); D = 0 when all cases are in one category

The maximum value of D depends on how many categories there are … thus D cannot be compared across discrete variable with different numbers of categories

Measures of Spread for Discrete Variables

D doesn't work when variables have different numbers of categories

	Cat. 1	Cat. 2	Cat. 3	Cat. 4
Y	10	10	10	10
Z	10	10	10	

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

$D_Y = 1 - [\ (0.25^2) + (0.25^2) + (0.25^2) + (0.25^2)\ ] = 0.75$

$D_Z = 1 - [\ (0.33^2) + (0.33^2) + (0.33^2)\ ] = 0.67$

But they are both uniformly distributed…

Measures of Spread for Discrete Variables

Index of Qualitative Variation (IQV)

A measure of dispersion in the distribution of a discrete variable that is standardized such that it can be compared across discrete variables with different numbers of categories

$$IQV = [K/(K-1)](D)$$

The maximum value of IQV is always 1.0 (when the cases are distributed evenly across categories)

The minimum value of IQV is always 0.0 (when all of the cases fall into one category)

Thus IQV can be compared across discrete variables that differ with respect to how many categories they have

Measures of Spread for Discrete Variables

D doesn't work when variables have different numbers of categories

	Cat. 1	Cat. 2	Cat. 3	Cat. 4
Y	10	10	10	10
Z	10	10	10	

$$IQV = [K/(K-1)](D)$$

$IQV_Y = [4/(4\text{-}1)] \times 0.75 = 1.00$

$IQV_Z = [3/(3\text{-}1)] \times 0.67 = 1.00$

Want More?

David Lane's Text

http://onlinestatbook.com/2/summarizing_distributions/summarizing_distributions.html

Richard Lowry's Text

http://vassarstats.net/textbook/ch2pt1.html and

http://vassarstats.net/textbook/ch2pt2.html

Gerard Dallal's Text

http://www.jerrydallal.com/LHSP/summary.htm