Unit 3: Univariate Statistics (Sensitive)

Unit 3 Post Hole:

Conduct a z-score transformation by hand from a small data set.

Unit 3 Technical Memo and School Board Memo:

Produce an appropriate table, and discuss the descriptive statistics for four variables (from Memos 1 and 2, plus an additional continuous or dichotomous predictor of your choice).

Unit 3 Reading:

http://onlinestatbook.com/
Chapter 1, Introduction
Chapter 2, Graphing Distributions
Chapter 3, Summarizing Distributions

Unit 3: Technical Memo and School Board Memo

Work Products (Part I of II):

I. Technical Memo: Have one section per biviariate analysis. For each section, follow this outline. (2 Sections)

- A. Introduction
  - i. State a theory (or perhaps hunch) for the relationship—think causally, be creative. (1 Sentence)
  - ii. State a research question for each theory (or hunch)—think correlationally, be formal. Now that you know the statistical machinery that justifies an inference from a sample to a population, begin each research question, "In the population,…" (1 Sentence)
  - iii. List the two variables, and label them "outcome" and "predictor," respectively.
  - iv. Include your theoretical model.
- B. Univariate Statistics. Describe your variables, using descriptive statistics. What do they represent or measure?
  - i. Describe the data set. (1 Sentence)
  - ii. Describe your variables. (1 Short Paragraph Each)
    - a. Define the variable (parenthetically noting the mean and s.d. as descriptive statistics).
    - b. Interpret the mean and standard deviation in such a way that your audience begins to form a picture of the way the world is. Never lose sight of the substantive meaning of the numbers.
    - c. Polish off the interpretation by discussing whether the mean and standard deviation can be misleading, referencing the median, outliers and/or skew as appropriate.
- C. Correlations. Provide an overview of the relationships between your variables using descriptive statistics.
  - i. Interpret all the correlations with your outcome variable. Compare and contrast the correlations in order to ground your analysis in substance. (1 Paragraph)
  - ii. Interpret the correlations among your predictors. Discuss the implications for your theory. As much as possible, tell a coherent story. (1 Paragraph)
  - iii. As you narrate, note any concerns regarding assumptions (e.g., outliers or non-linearity), and, if a correlation is uninterpretable because of an assumption violation, then do not interpret it.

Unit 3: Technical Memo and School Board Memo

Work Products (Part II of II):

I. Technical Memo (continued)

D. Regression Analysis. Answer your research question using inferential statistics. (1 Paragraph)

i. **Include your fitted model.**

ii. Use the R^2 statistic to convey the goodness of fit for the model (i.e., strength).

iii. To determine statistical significance, test the null hypothesis that the magnitude in the population is zero, reject (or not) the null hypothesis, and draw a conclusion (or not) from the sample to the population.

iv. Describe the direction and magnitude of the relationship in your sample, preferably with illustrative examples. Draw out the substance of your findings through your narrative.

v. Use confidence intervals to describe the precision of your magnitude estimates so that you can discuss the magnitude in the population.

vi. If simple linear regression is inappropriate, then say so, briefly explain why, and forego any misleading analysis.

X. Exploratory Data Analysis. Explore your data using outlier resistant statistics.

i. **For each variable, use a coherent narrative to convey the results of your exploratory univariate analysis of the data. Don't lose sight of the substantive meaning of the numbers. (1 Paragraph Each)**

ii. **For the relationship between your outcome and predictor, use a coherent narrative to convey the results of your exploratory bivariate analysis of the data. (1 Paragraph)**

II. School Board Memo: Concisely, precisely and plainly convey your key findings to a lay audience. Note that, whereas you are building on the technical memo for most of the semester, your school board memo is fresh each week. (Max 200 Words)

III. Memo Metacognitive

Unit 3: Road Map (VERBAL)

Nationally Representative Sample of 7,800 8th Graders Surveyed in 1988 (NELS 88).
Outcome Variable (aka Dependent Variable):

READING, a continuous variable, test score, mean = 47 and standard deviation = 9

Predictor Variables (aka Independent Variables):

FREELUNCH, a dichotomous variable, 1 = Eligible for Free/Reduced Lunch and 0 = Not
RACE, a polychotomous variable, 1 = Asian, 2 = Latino, 3 = Black and 4 = White

- Unit 1: In our sample, is there a relationship between reading achievement and free lunch?
- Unit 2: In our sample, what does reading achievement look like (from an outlier resistant perspective)?
- Unit 3: In our sample, what does reading achievement look like (from an outlier sensitive perspective)?
- Unit 4: In our sample, how strong is the relationship between reading achievement and free lunch?
- Unit 5: In our sample, free lunch predicts what proportion of variation in reading achievement?
- Unit 6: In the population, is there a relationship between reading achievement and free lunch?
- Unit 7: In the population, what is the magnitude of the relationship between reading and free lunch?
- Unit 8: What assumptions underlie our inference from the sample to the population?
- Unit 9: In the population, is there a relationship between reading and race?
- Unit 10: In the population, is there a relationship between reading and race controlling for free lunch?
- Appendix A: In the population, is there a relationship between race and free lunch?

Unit 3: Roadmap (R Output)

```
> load("E:/User/Folder/RoadmapData.rda")
> library(abind, pos=4)
> numSummary(RoadmapData[,c("FREELUNCH", "READING")],
+   statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
                mean       sd    0%   25%   50%   75%  100%    n
FREELUNCH  0.3353846 0.472155  0.00  0.00  0.00  1.00  1.00 7800
READING   47.4940397 8.569440 23.96 41.24 47.43 53.93 63.49 7800
```

Unit 3

Unit 2

```
> RegModel.1 <- lm(READING~FREELUNCH, data=RoadmapData)
> summary(RegModel.1, cor=FALSE)
Call:
lm(formula = READING ~ FREELUNCH, data = RoadmapData)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.1176     0.1147  428.17   <2e-16 ***
FREELUNCH    -4.8409     0.1981  -24.44   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.26 on 7798 degrees of freedom
Multiple R-squared: 0.07114, Adjusted R-squared: 0.07102
F-statistic: 597.3 on 1 and 7798 DF,  p-value: < 2.2e-16
```

Unit 1 Unit 8 Unit 6

Unit 5

Unit 9

```
> library(MASS, pos=4)
> Confint(RegModel.1, level=.95)
             Estimate     2.5 %     97.5 %
(Intercept) 49.117616 48.892742 49.342489
FREELUNCH   -4.840938 -5.229237 -4.452638
```

Unit 7

```
> cor(RoadmapData[,c("FREELUNCH","READING")])
            FREELUNCH    READING
FREELUNCH   1.0000000 -0.2667237
READING    -0.2667237  1.0000000
```

Unit 4

Unit 3: Roadmap (SPSS Output)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.267[a]	.071	.071	8.25952

a. Predictors: (Constant), FREELUNCH

Unit 5

Statistics

		READING	FREELUNCH
N	Valid	7800	7800
	Missing	0	0
Mean		47.4940	.3354
Std. Deviation		8.56944	.47216
Minimum		23.96	.00
Maximum		63.49	1.00
Percentiles	25	41.2400	.0000
	50	47.4300	.0000
	75	53.9300	1.0000

Unit 3

Unit 2

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	40744.322	1	40744.322	597.251	.000[a]
	Residual	531977.541	7798	68.220		
	Total	572721.864	7799			

a. Predictors: (Constant), FREELUNCH

b. Dependent Variable: READING

Unit 9

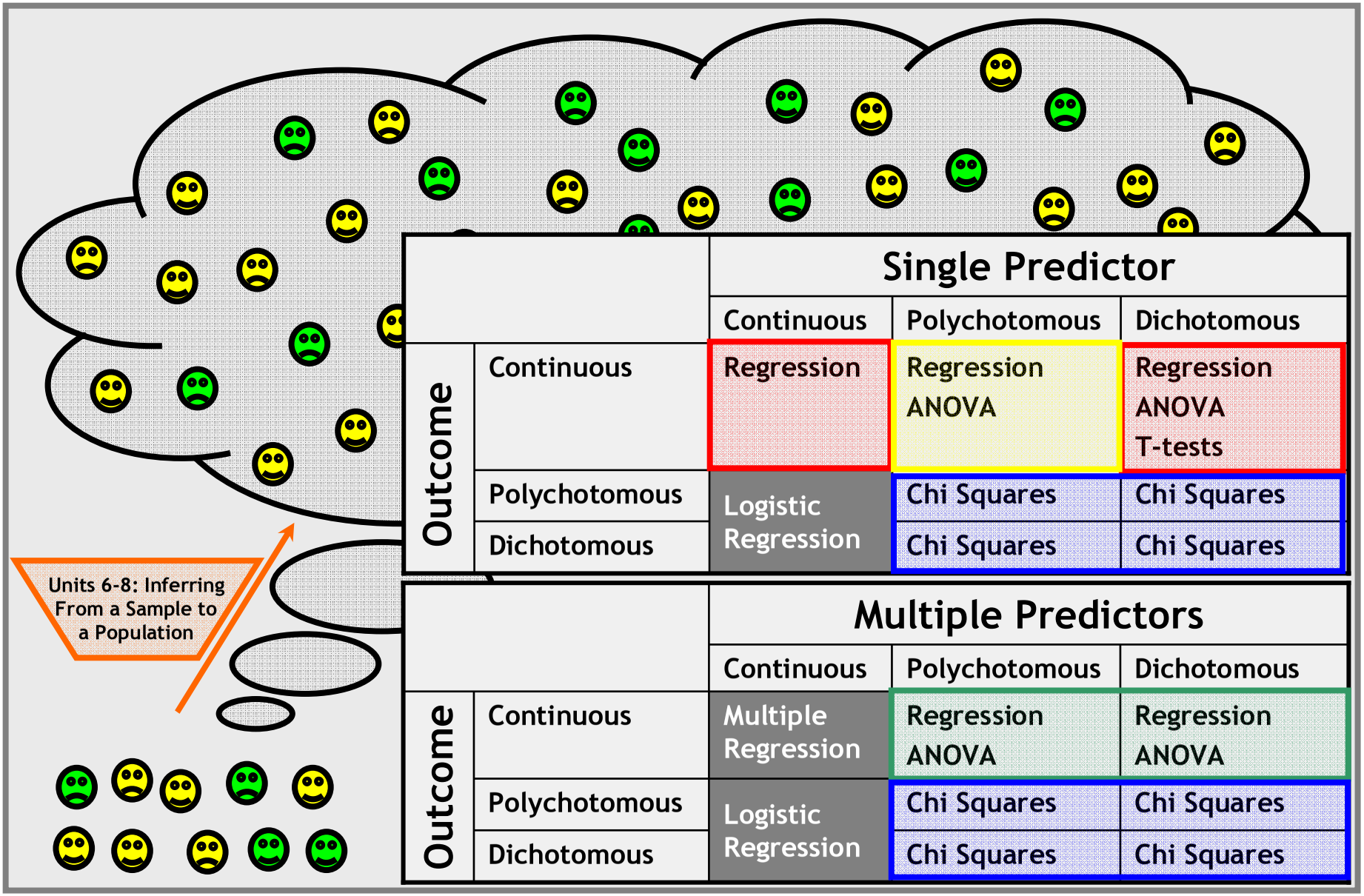
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	49.118	.115		428.169	.000	48.893	49.342
	FREELUNCH	-4.841	.198	-.267	-24.439	.000	-5.229	-4.453

a. Dependent Variable: READING

Unit 1 Unit 8 Unit 4 Unit 6 Unit 7

Unit 3: Road Map (Schematic)



Single Predictor

Outcome	Continuous	Polychotomous	Dichotomous
Continuous	Regression	Regression ANOVA	Regression ANOVA T-tests
Polychotomous	Logistic Regression	Chi Squares	Chi Squares
Dichotomous		Chi Squares	Chi Squares

Multiple Predictors

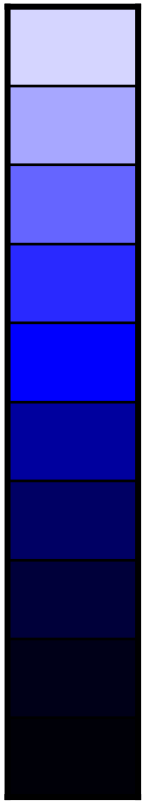
Outcome	Continuous	Polychotomous	Dichotomous
Continuous	Multiple Regression	Regression ANOVA	Regression ANOVA
Polychotomous	Logistic Regression	Chi Squares	Chi Squares
Dichotomous		Chi Squares	Chi Squares

Epistemological Minute

Nelson Goodman (http://plato.stanford.edu/entries/goodman-aesthetics) argues that, for purposes of referring to things, we have two primary tools: labeling and exemplifying. I'm thinking of a color, and if I want to refer to it, I can label it or exemplify it. Furthermore, I can do my labeling and/or exemplifying either literally or metaphorically.

	Literal	Metaphorical
Label	I can say, "I'm thinking of blue."	I can say, "I'm thinking of cool."
Example	I can point to a color swatch.	I can play some Miles Davis for you.

http://www.youtube.com/watch?v=PoPL7BExSQU



If an English-language learner asks me, "What is 'blue'?" Perhaps, I can refer to blue by labeling it in a language that she understands. If that resource is not available to me, however, I can always refer to blue by exemplifying it. Ideally, I would show her a whole spectrum of blue or at least a good sampling of blue hues and shades. What if I could only show her one swatch of blue, but I had a choice of the hue and shade. Which swatch should I show her? Does it matter? I think that, if I could only show her one swatch, I would show her something in the middle range of hue and shade.

In data analysis, if a researcher asked me to summarize a variable's distribution of values, ideally I would show her the whole distribution, perhaps by way of a histogram. What if I could only give her one number? Should I give her a value from the distribution? Does it matter what value? I think that, if I could only give her one value from the distribution, I would give her a value in the middle range of the distribution, probably the median. The median value may be the most reasonable way to literally exemplify all the values in the distribution (IF we are restricted to one value from the distribution). Yet, why restrict ourselves to literal exemplification when we can metaphorically exemplify? Perhaps there is a value that is not literally in the distribution but that, metaphorically, is at the center of the distribution? Note that the mean is not a value in the distribution, yet it exemplifies the values in the distribution. I wonder if this exemplification is metaphorical.

You might be asking, "Are not means sometimes a value in the distribution?" and I would reply, "Not if we go out enough decimal places." The mean is it's own abstract thing, but it can help us see an important feature of concrete distributions.

Unit 3: Research Question

Theory: Students who go to smaller schools will have better math achievement scores, because smaller schools form tighter communities, and consequently struggling and gifted students are less likely to fall through the cracks.

Research Question: Are students' math achievement scores negatively correlated with their school population size?

Data Set: (NELS88Math.sav)

Variables:

Outcome—Math Achievement Score (*MATHACH*)

Predictor—Number of Students in Student's School (*SchoolPop*)

Model: $MathAch = \beta_0 + \beta_1 SchoolPop + \varepsilon$

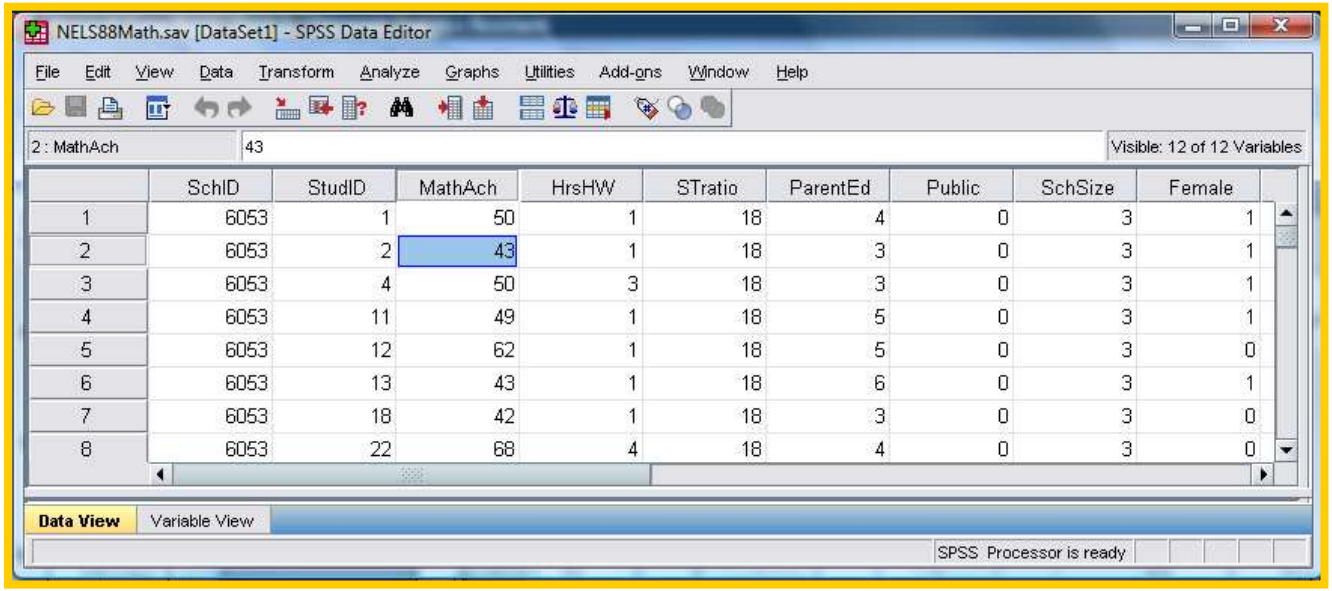
NELS88Math.Sav Codebook

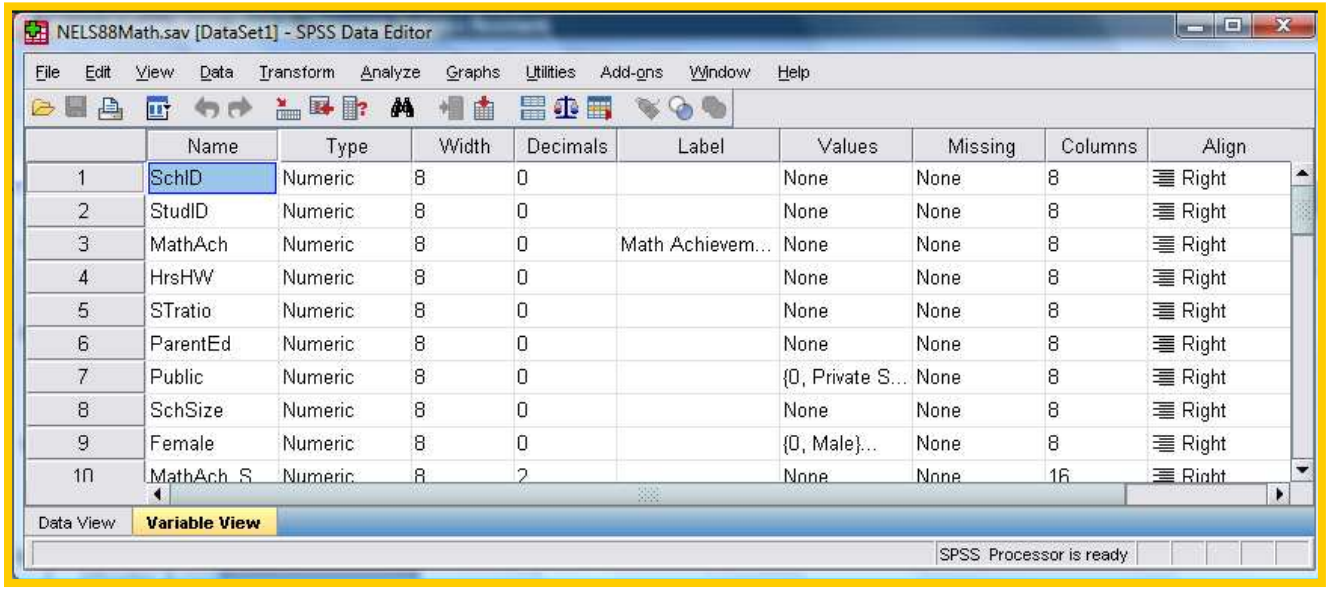
Dataset	NELS88Math.txt
Overview	Multilevel dataset on the mathematics achievement of 519 students in 23 schools, as a function of the number of hours of mathematics homework they complete each week and the student teacher ratio in their school, by selected controls.
Source	Kreft, I.G., & de Leeuw, J. *Introducing Multilevel Modeling*. Thousand Oaks, CA: Sage Publications, 1998, pp. 23-24. Data are a sub-sample from NELS-88 , which contains information on educational processes and outcomes for a nationally representative sample of eighth-graders first surveyed in 1988, and then again in 1990, 1992, 1994, and 2000. Students reported data on school, work, neighborhood, and home experiences; educational resources available to them; educational and occupational aspirations; substance abuse; and the education levels of parents and peers;. The reading, social studies, mathematics and science achievement of students were measured while they were in school. Background information was provided by teachers, parents, and school administrators. The public use dataset is available on CD-ROM and is free from NCES.
Sample size	23 schools, 519 students
Last updated	October 8, 2003

NELS88Math.Sav Codebook

Structure of Dataset			
Col. #	**Variable Name**	**Variable Description**	**Variable Metric/Labels**
1	SCHID	School identification code	Integer
2	STUID	Student identification code	Integer
3	MATHACH	Mathematics number-right achievement score	Continuous variable ranging from 30 to 71.
4	HOURSHW	Number of hours of mathematics homework completed each week	Ordinal variable: 0 = none 1 = less than 1 hour 2 = 1 hour 3 = 2 hours 4 = 3 hours 5 = 4 to 6 hours 6 = 7 to 9 hours 7 = 10 hours or more
5	STRATIO	Student/teacher ratio in the school	Continuous variable ranging from 10 to 28: 10 = 10 or less 11 = 11. etc.
6	PARENTED	Highest educational level attained by either parent.	Ordinal variable: 1 = Did not finish HS 2 = HS Grad/GED 3 = >HS & <4yr degree 4 = College grad. 5 = MA, or equiv. 6 = Ph.D., M.D., or equiv.
7	PUBLIC	Is the school in the public sector?	Dichotomous variable: 0 = no 1 = yes
8	SCHSIZE	Total school enrollment	Ordinal variable: 1 = 1–199 students 2 = 200-399 students 3 = 400-599 students 4 = 600-799 students 5 = 800-999 students 6 = 1000-1199 students 7 = 1200+ students
9	FEMALE	Is the student female?	Dichotomous variable: 0 = no 1 = yes

NELS88Math.sav

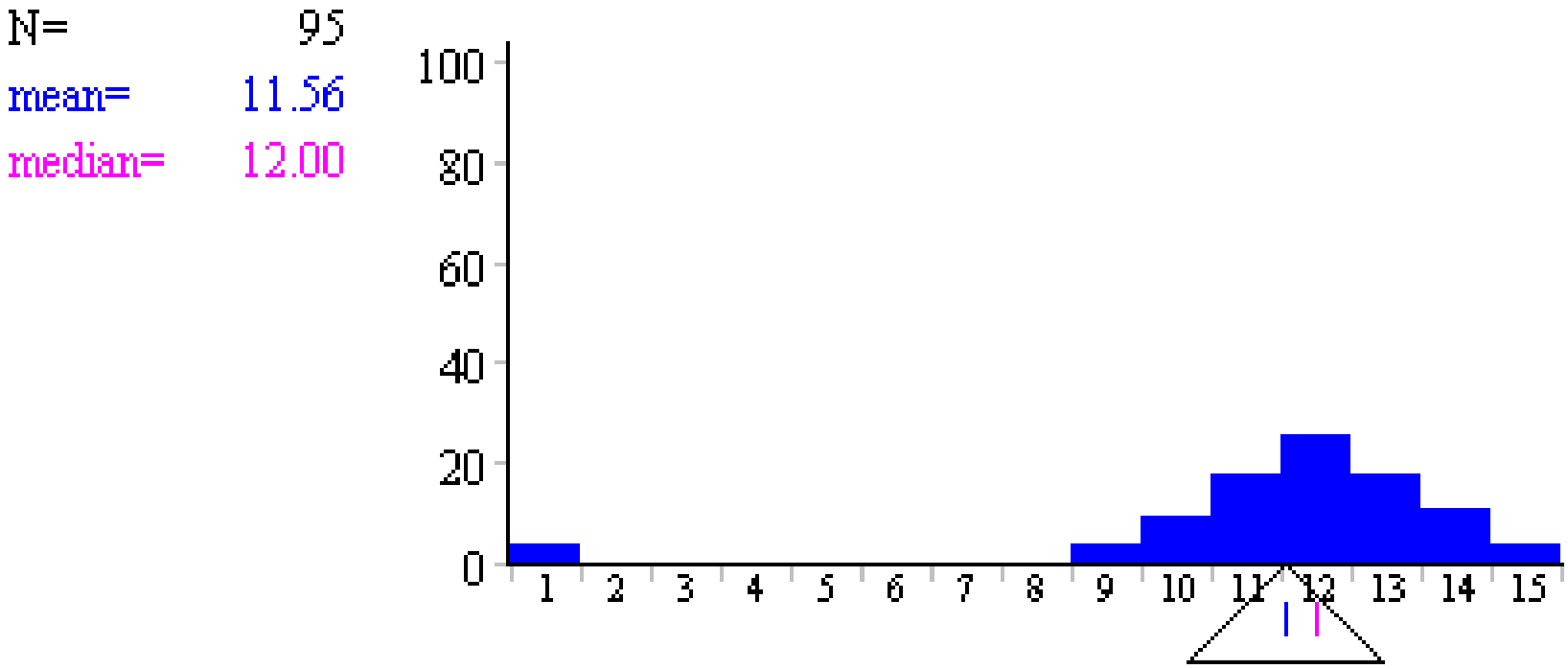




Two Families of Statistics

	Outlier Resistant Based on Rank Order	Outlier Sensitive Based on Intervals
	*Road Race: All that matters is who came in 1st place, 2nd place, 3rd place, 4th place etc.	*Road Race: It's not enough to finish 1st place; it's important to finish by the widest margin possible—spacing counts.
Location	***Median** ***50th Percentile** *The Value For The Person Who Ranked In The Middle *Line everybody up in order, and single out the first person who ranks above 50% of the others.	***Mean** ***The Balancing Point** *An Abstract Value That Amalgamates All Values *Add up all the values and divide by the number of values. Beware the "Bill Gates Effect."
Spread	***Midspread** *The midspread is the interval that covers 50% of the values.	***Standard Deviation** *The standard deviation measures how wrong, on average, the mean is as a prediction for individuals.

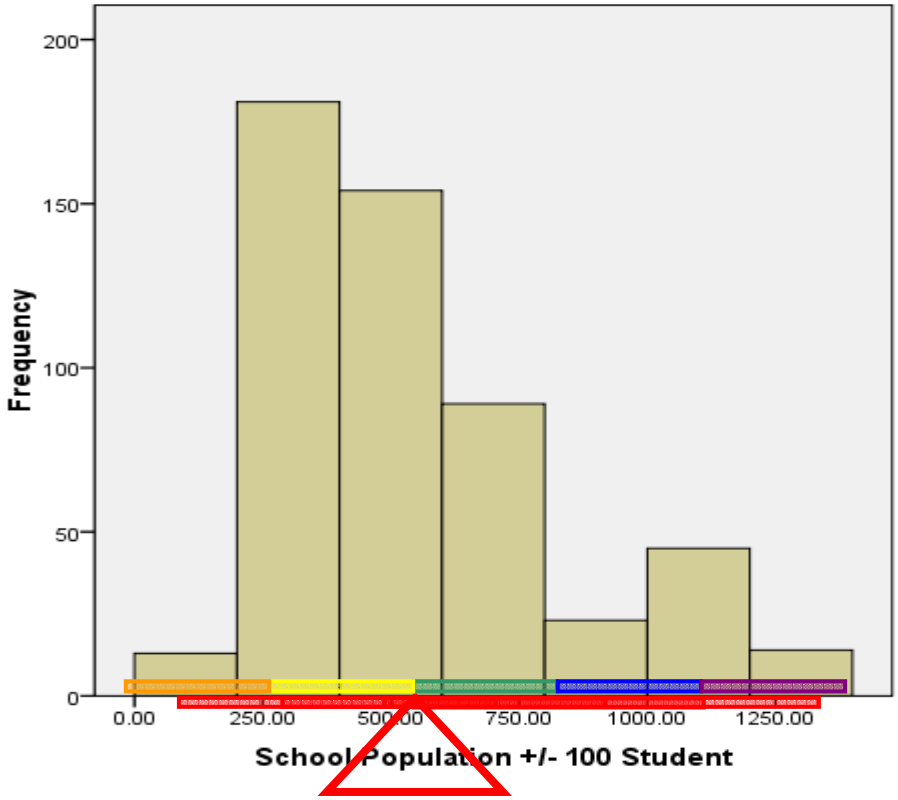
Outlier Resistant vs. Outlier Sensitive Statistics



http://onlinestatbook.com/simulations/balance/balance_sim.html

Describing Math Achievement and School Size

Figure 3.1. Histogram and univariate statistics for students' school population sizes (n = 519).



I invite you to think of the mean as the most reasonable* prediction for individuals in the absence of further information. That is, if being close matters, otherwise we would use the most common value, the mode.

*Not necessarily very reasonable.

School Population +/- 100 Student

N	Valid	519.0000
	Missing	.0000
Mean		545.8574
Std. Deviation		280.0600
Minimum		100.0000
Maximum		1300.0000
Percentiles	25	300.0000
	50	500.0000
	75	700.0000

We can ask of each individual, how many standard deviations from the mean are you?

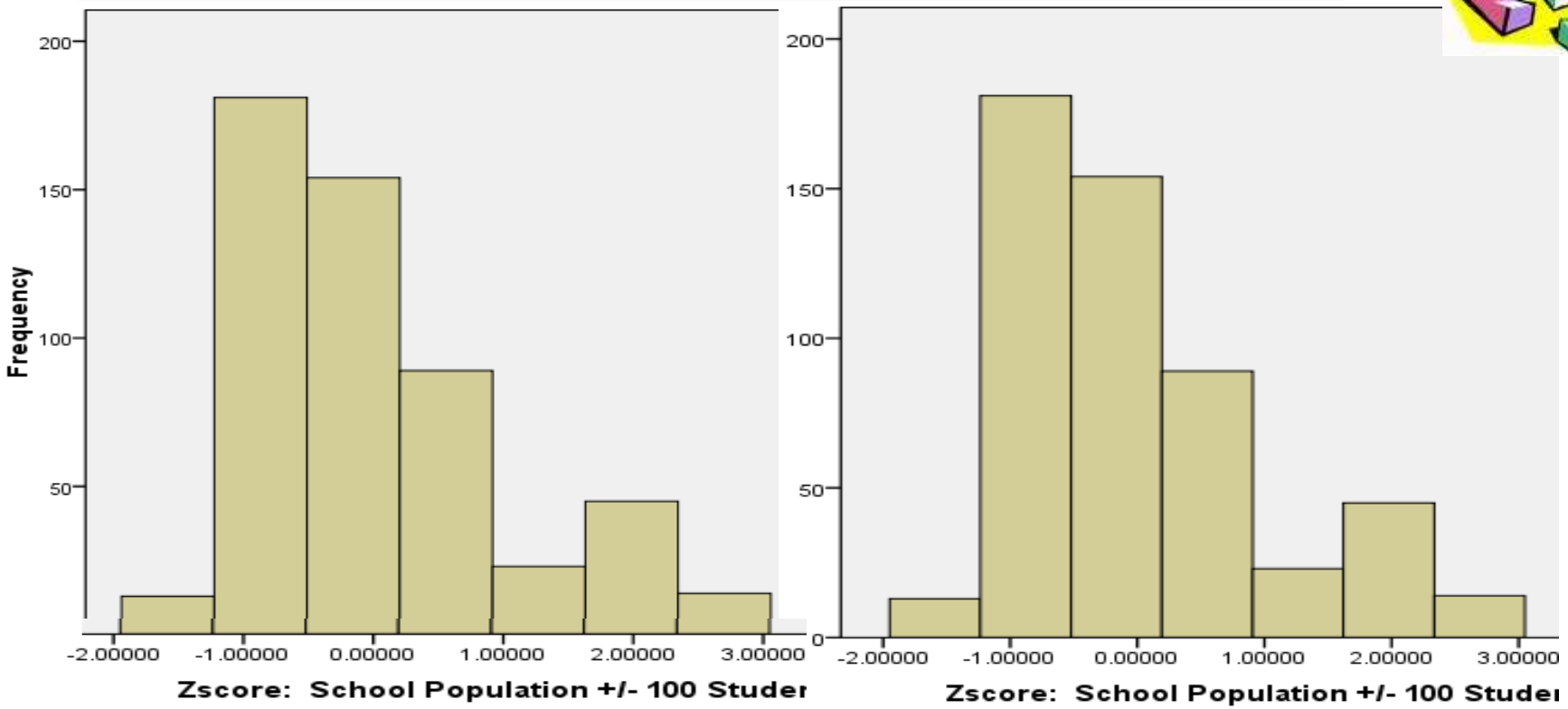
The standard deviation measures how wrong, on average, the mean is as a prediction for individuals.

Note: I use "average" as a general term for location (or measure of central tendency), so for example means, medians, and modes are all averages in my book.

"Deviation" is distance from the mean. "Standard" is the average.

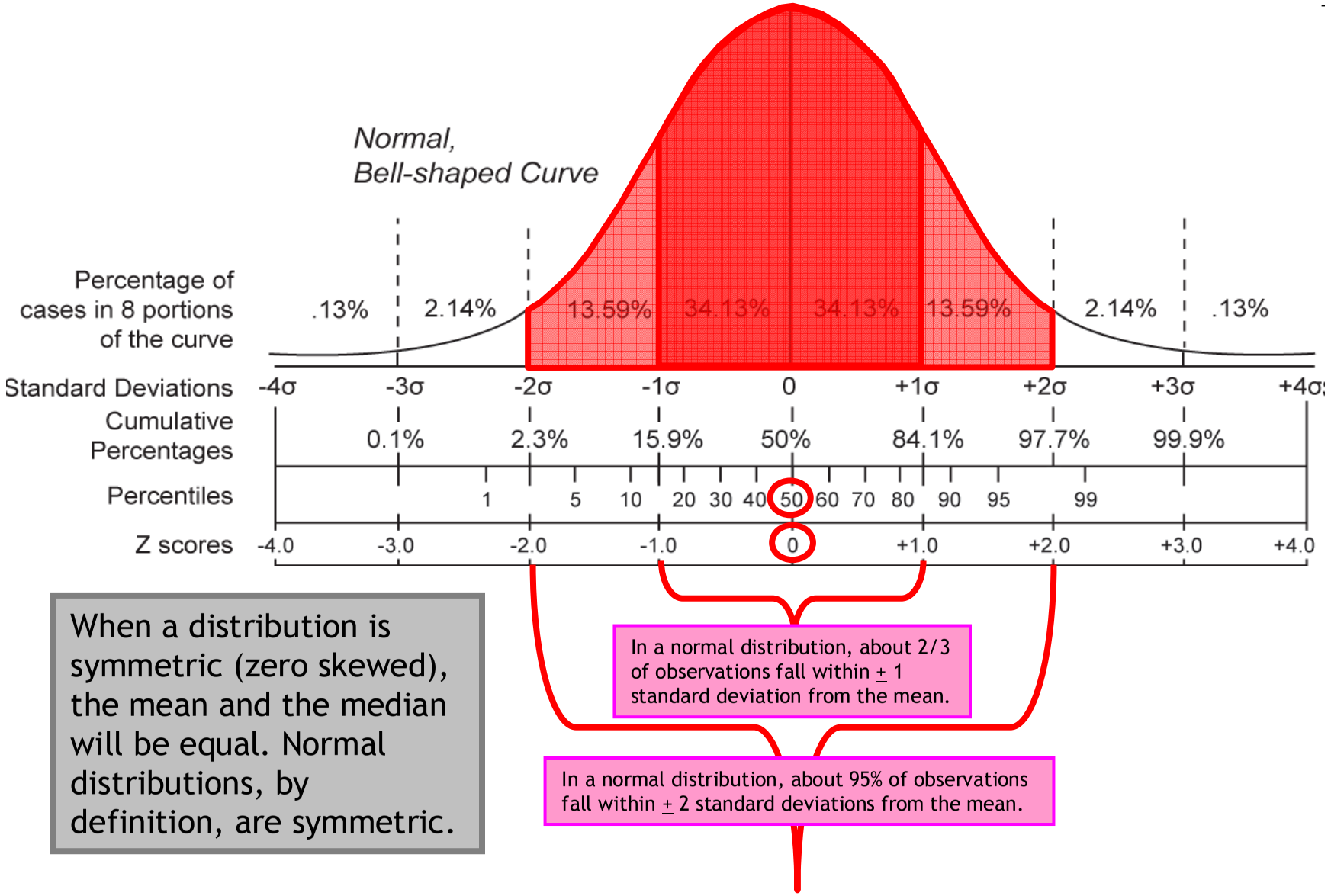
Describing Math Achievement and School Size

Figure 3.1. Histogram and univariate statistics for students' school population sizes (n = 519).



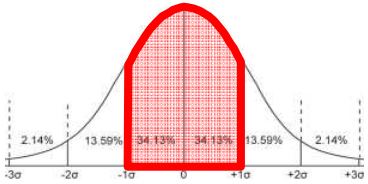
A z-score (or standardized score) is a linear transformation of the raw score. From each raw score, we subtract the mean and divide by the standard deviation. Because we are only adding/subtracting and multiplying/dividing, we do not change the shape of the distribution (hence, *linear* transformation). In essence, we call the mean "zero" and we assign a value to everybody based on how many standard deviations they are from the mean.

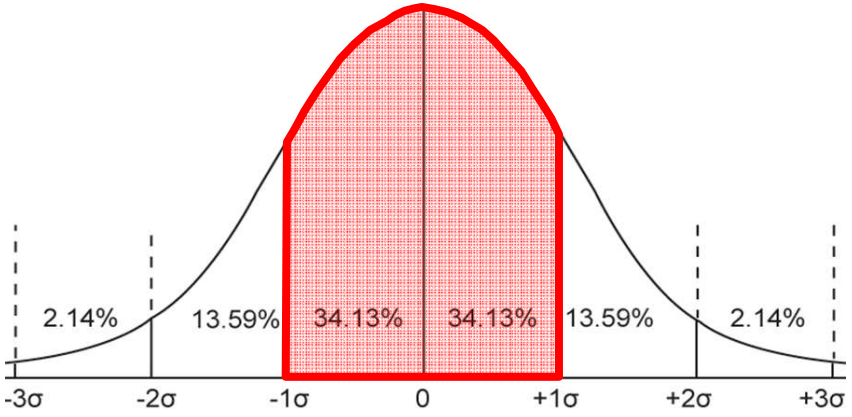
Standard Deviations and Normal Curves



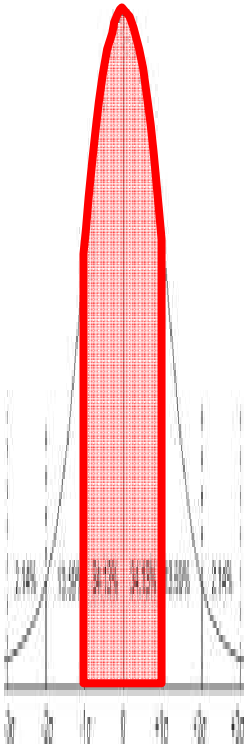
When a distribution is symmetric (zero skewed), the mean and the median will be equal. Normal distributions, by definition, are symmetric.

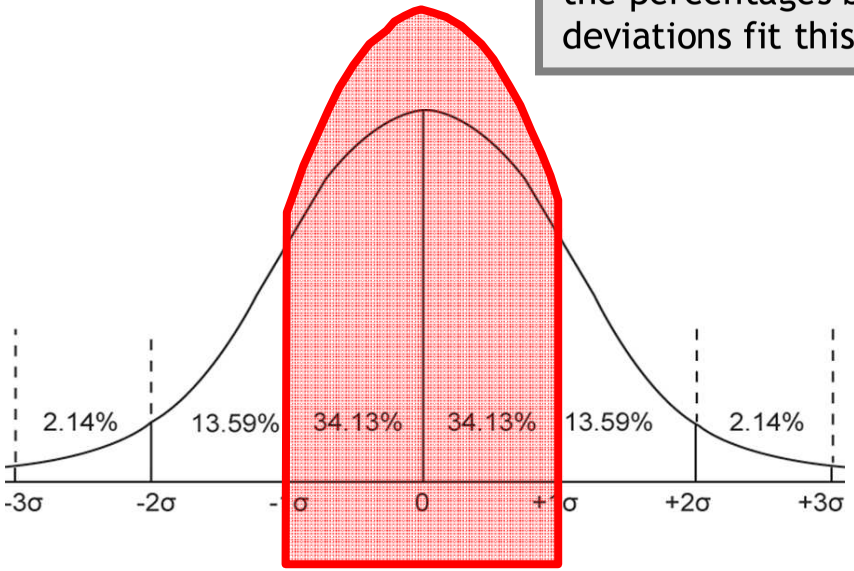
The Area Under a Normal Curve is Definitional

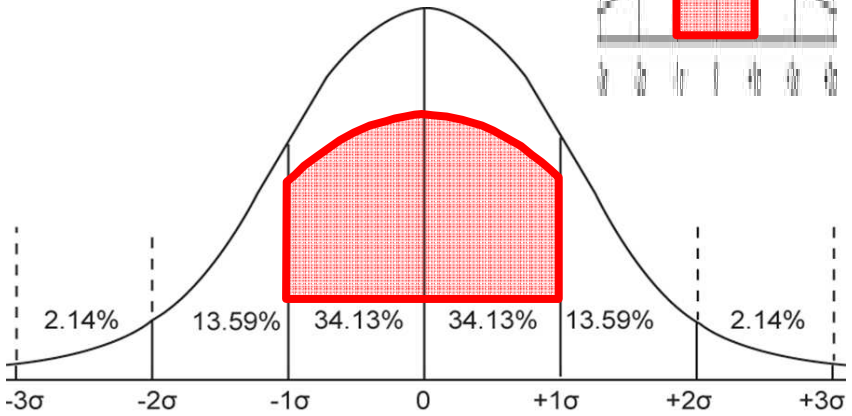


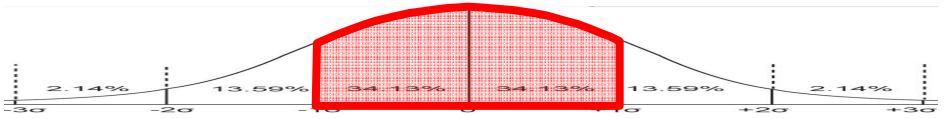


What makes a distribution normal is that the percentages between the standard deviations fit this exact pattern.



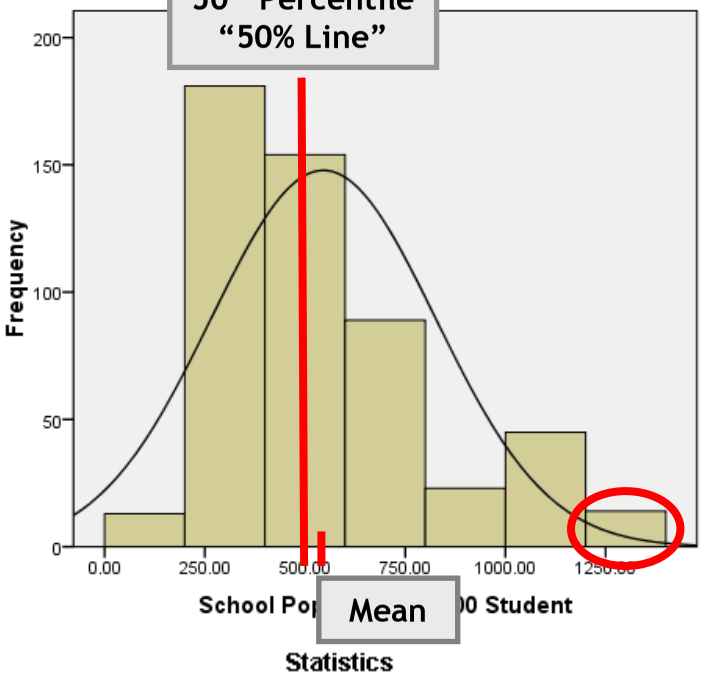


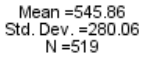


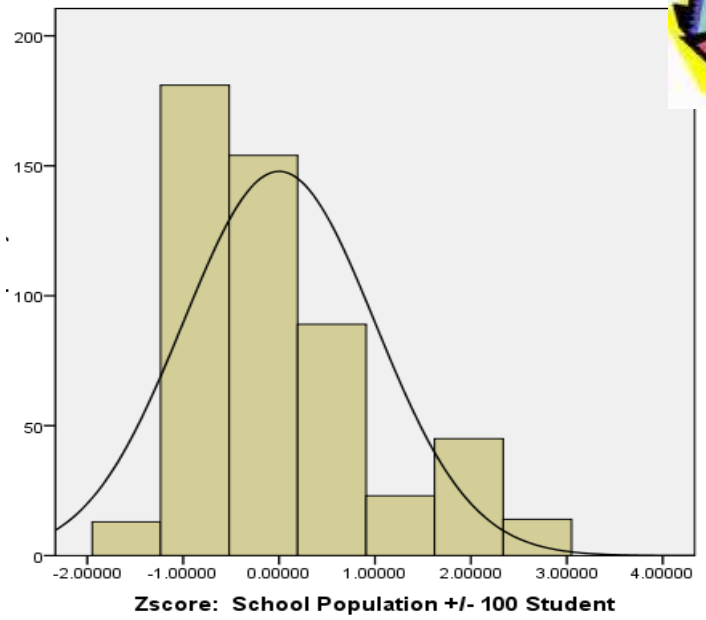


Describing Math Achievement and School Size

Figu… and univariate statistics for students' school population sizes (n = 519).







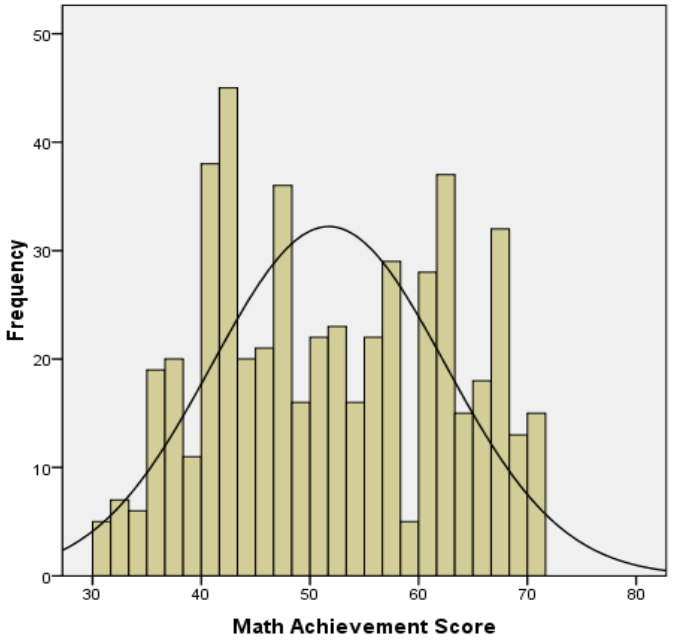
School Population +/- 100 Student

N	Valid	519.0000
	Missing	.0000
Mean		545.8574
Std. Deviation		280.0600
Minimum		100.0000
Maximum		1300.0000
Percentiles	25	300.0000
	50	500.0000
	75	700.0000

Students in our sample go to schools of different sizes, the average student goes to a school of about 546 students (m = 546, sd = 280). The preponderance of students go to schools of between 266 and 826 students (±1 standard deviation from the mean). The distribution is positively skewed, so the few students from the largest schools, schools of approximately 1300 students, are exerting unreciprocated leverage on the mean, pulling the mean away from the median. (We may need to explain to our audience how more than half the students can go to smaller than average schools.)

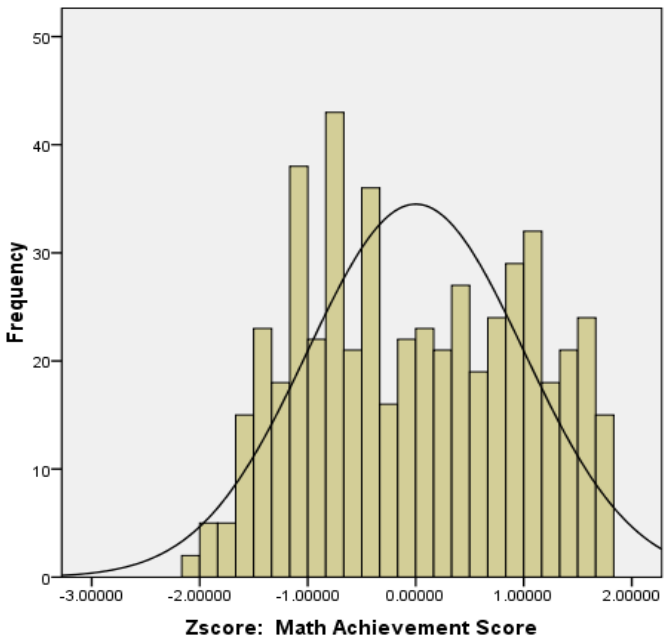
Describing Math Achievement and School Size

Figure 3.2. Histogram and univariate statistics for math achievement scores (n = 519).



Mean =51.72
Std. Dev. =10.709
N =519

Notice that the shapes are slightly different; this is strictly an artifact of binning. The shapes should be identical!



Statistics

Math Achievement Score

N	Valid	519.00
	Missing	.00
Mean		51.72
Std. Deviation		10.71
Minimum		30.00
Maximum		71.00
Percentiles	25	43.00
	50	51.00
	75	62.00

The 519 students in our sample took a math achievement test, with a mean score of 52 and a standard deviation of 11. All but two students fall within ± 2 standard deviations from the mean with scores between 30 and 74, and the two exceptions fall just outside -2 standard deviations from the mean. The distribution is symmetric as evidenced by the nearly equal mean and median, but the distribution is bimodal, so it does not appear the students are clustering around one score, but rather two scores.

Formal Table of Univariate Descriptive Statistics

Figure 3.3. Table of select descriptive statistics from the National Education Longitudinal Study (NELS) dataset (n = 519).

	n	Mean	Standard Deviation	Min.	Max.
Math Achievement Score	**519**	**51.72**	**10.71**	**30**	**71**
School Population	**519**	**545.86**	**280.06**	**100**	**1300**
Student/Teacher Ratio	**519**	**16.76**	**4.93**	**10**	**28**
Female = 1 Male = 0	**519**	**0.52**	**0.50**	**0**	**1**
Public = 1 Private = 0	**519**	**0.61**	**0.49**	**0**	**1**

- Notice the well written caption.
- Notice that there are only a few horizontal lines and no vertical lines. This is a throwback to old typesetting restrictions. Many journals still adhere to this convention. In Word, go to Format > Borders and Shading…. You can select only certain rows for the sake of determining borders.
- Notice that the decimals are vertically aligned. You can simply right justify if all your values in a given column have the same number of decimal places. Otherwise, Word has a trick.
- It is probably too redundant to have "n = 519" so many times. I'm thinking about getting rid of the column (or the parenthetical note in the caption).

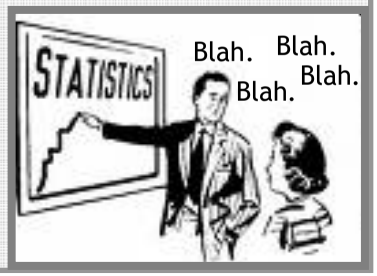
- I am not a stickler for any particular table formatting convention (e.g., APA). Some researchers, however, are sticklers. Please be patient with them, especially if they sign your checks. Every journal has its own rules. This exemplar will get you close to most. For the purposes of this class, make your tables look good. This table looks good to me, but so would several other variations.

Notice that the mean of dichotomous (1/0) variables is the proportion of 1s. If the proportion of 1s is 0.50, doesn't it make sense that the standard deviation would also be 0.50? In our sample, 52% of our subjects are female, and 61% percent of our subjects attend public school.

Discussing Univariate Descriptive Statistics

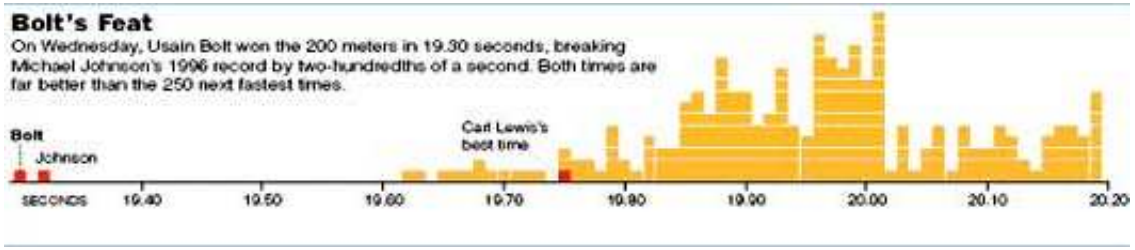
When discussing univariate descriptive statistics (or any statistics for that matter), make sure the audience has enough information to draw the right conclusion (or at least enough information to not draw the wrong conclusion!).

i. Define the variable noting the mean and standard deviation (perhaps in parentheses).
ii. Interpret the mean and standard deviation in such a way that your audience begins to form a picture of the way the world is.
iii. Polish off the interpretation by discussing whether the mean and standard deviation can be misleading, referencing the median, outliers and/or skew as appropriate.



Never lose sight of the substantive meaning of the numbers.

Usually, this is the post hole, but the Unit 3 Post Hole is next…



http://freakonomics.blogs.nytimes.com/2008/08/21/usain-bolt-its-just-not-normal/

Steps For Conducting a Z Transformation by Hand

1) Create a stem and leaf plot to get an initial handle on the distribution.
2) Calculate the mean.
3) Calculate the standard deviation of the sample.
    1) Calculate the deviations from the mean.
    2) Square the mean deviations.
    3) Sum the squared mean deviations.
    4) Divide the sum of squared mean deviations by the sample size (less one).*

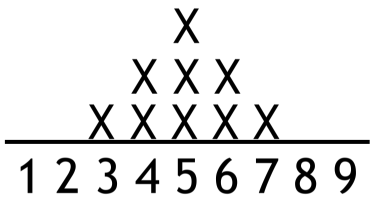
        *You've just calculated the variance, the average squared deviation.
    5) Take the square root of the variance.
4) Standardize (or z transform) each value.
    1) For each value, calculate the deviation from the mean.*

        *This was your first step in calculating the standard deviation.
    2) Divide each mean deviation by the standard deviation.

Conducting a Z Transformation by Hand (Steps 1 & 2)

Our Sample Sample: 6, 7, 5, 4, 6, 5, 3, 5, 4

1) Create a stem and leaf plot to get an initial handle on the distribution.



2) Calculate the mean.

$$mean = \bar{x} = \frac{6+7+5+4+6+5+3+5+4}{9} = \frac{45}{9} = 5$$

Conducting a Z Transformation by Hand (Step 3)

3) Calculate the standard deviation of the sample.

1) Calculate the deviations from the mean.
2) Square the mean deviations.
3) Sum the squared mean deviations.
4) Divide the sum of squared mean deviations by the sample size (less one).
    1) You've just calculated the variance, the average squared deviation.
5) Take the square root of the variance.

This is one reason why statisticians love squares. Squares are always positive, so you can sum them and take means.

Raw	Mean	Mean Deviation	(Mean Deviation)^2
3	5	-2	4
4	5	-1	1
4	5	-1	1
5	5	0	0
5	5	0	0
5	5	0	0
6	5	+1	1
6	5	+1	1
7	5	+2	4
		sum=0	sum=12

Variance:

$$s^2 = \frac{12}{9-1} = \frac{12}{8} = \frac{3}{2} = 1.5$$

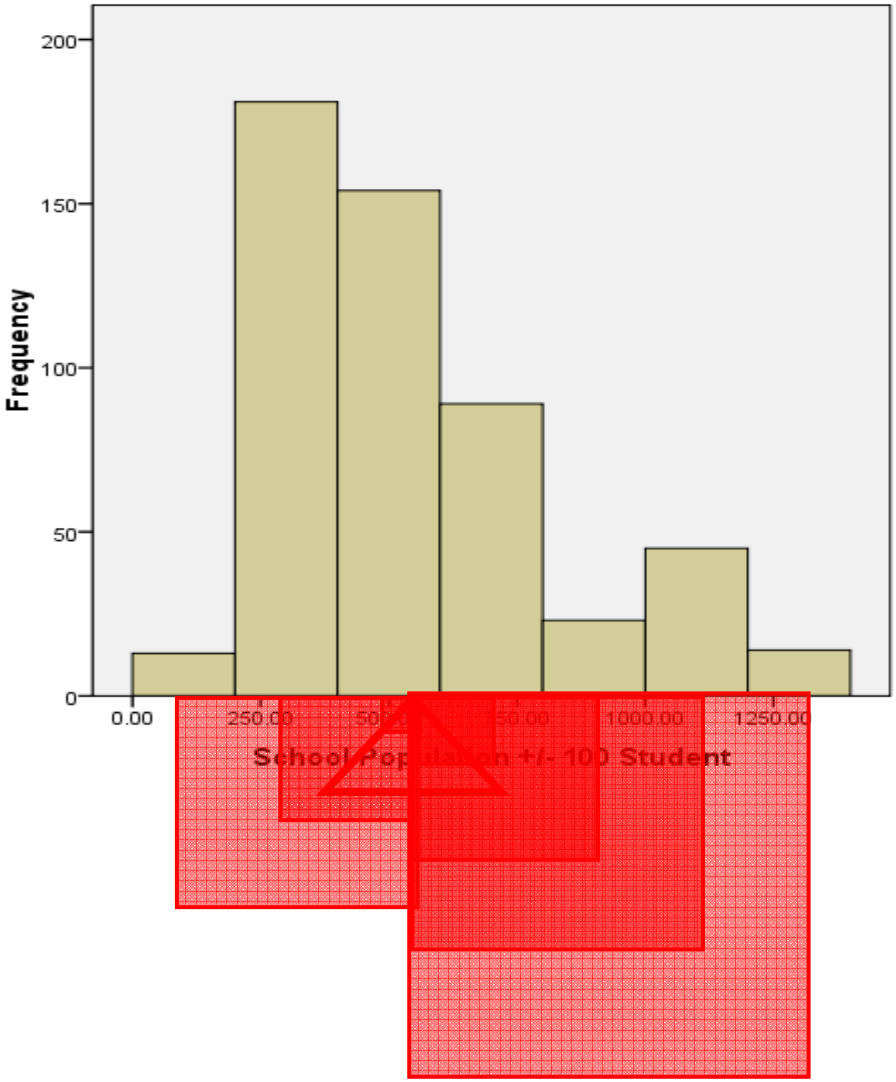
Standard Deviation:

$$s = \sqrt{1.5} = 1.22$$

When we take the average squared mean deviation why do we divide by n - 1 instead of just n? Technically, we divide by the degrees of freedom, not the sample size. The degrees of freedom is (roughly speaking) the "effective sample size." It has to do with unbiased inferences from the sample to the population. But, we're not yet thinking in terms of samples vs. populations, so for now, just take my word for it. In the Unit 7 Math Appendix, I give an intuitive explanation why.

Variance (A Step Toward Calculating Standard Deviation)

Figure 2.5. Histogram and univariate statistics for students' school population sizes (n = 519).



Mean =545.86
Std. Dev. =280.06
N =519

The variance is the average square. It is the mean square if you ignore the -1 to the n.

School Population +/- 100 Student

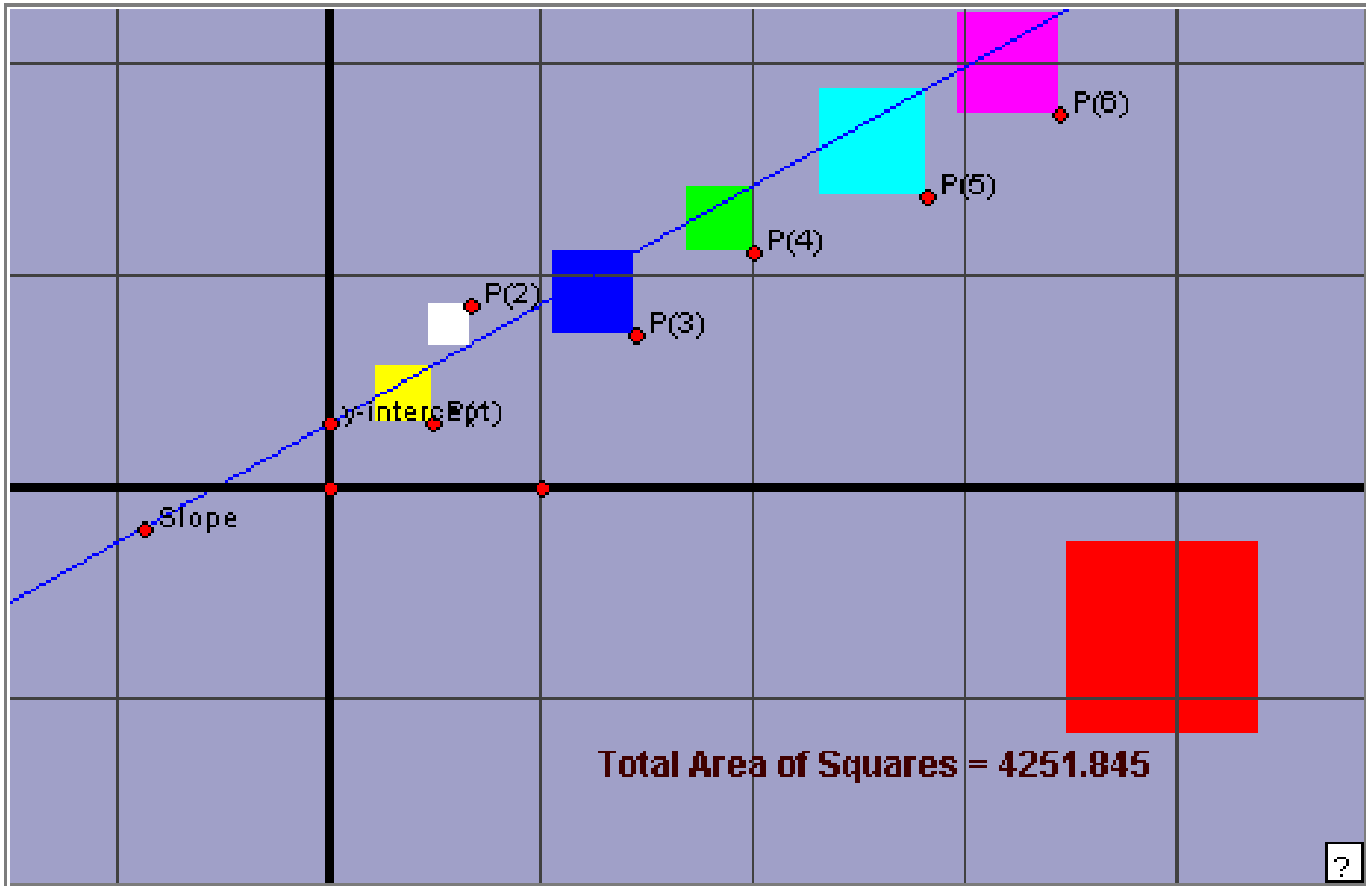
N	Valid	519.0000
	Missing	.0000
Mean		545.8574
Std. Deviation		280.0600
Minimum		100.0000
Maximum		1300.0000
Percentiles	25	300.0000
	50	500.0000
	75	700.0000

The standard deviation measures how wrong, on average, the mean is as a prediction for individuals.

The mean is sensitive to outliers, and the standard deviation is more so. Not just doubly, but squarely so! (Also, sometimes the mean gets balanced out by two opposite extremes, but not the standard deviation.)

Ordinary Least Squares (OLS) Regression

How does SPSS fit the line? The Method of Ordinary Least Squares



http://www.dynamicgeometry.com/JavaSketchpad/Gallery/Other_Explorations_and_Amusements/Least_Squares.html

Conducting a Z Transformation by Hand (Step 4)

4) Z transform each value.

1) For each value, calculate the deviation from the mean.
    1) This was your first step in calculating the standard deviation.
2) Divide each mean deviation by the standard deviation.

Raw Score	Mean	Mean Deviation	(Mean Deviation)^2	Z Score
3	5	-2	4	-1.6
4	5	-1	1	-0.8
4	5	-1	1	-0.8
5	5	0	0	0
5	5	0	0	0
5	5	0	0	0
6	5	+1	1	0.8
6	5	+1	1	0.8
7	5	+2	4	1.6
		sum=0	sum=12	

Variance:

$$s^2 = \frac{12}{9-1} = \frac{12}{8} = \frac{3}{2} = 1.5$$

Standard Deviation:

$$s = \sqrt{1.5} = 1.22$$

E.g., take the raw score of 3:

A raw score of 3 has a mean deviation of -2 (3-5). I.e., 3 is two units below the mean of 5. But how many standard deviations is it below the mean? The standard deviation is 1.22, so being 2 points below the mean is more than one standard deviation from the mean but less than two standard deviations from the mean. Let's divide the mean deviation (-2) by the standard deviation (1.22) to get an exact answer: -1.6.

Dig the Post Hole

Unit 3 Post Hole:

Conduct a z-score transformation by hand from a small data set.

Evidentiary materials: a small data set.

Math Scores: 60 72 64 53 51 60 44 59 62 65
Please show your work:

Please note the **mean** of the raw distribution:	
Please note the **sum of squared mean deviations**:	
Please note the **variance** of the raw distribution:	
Please note the **standard deviation** of the raw distribution:	

Tip 1: Use the boxes to guide you if you get lost.
Tip 2: You can use a calculator or spreadsheet software, but show your steps.
Tip 3: Check with your stem-and-leaf plot to make sure that the mean and s.d. make sense.

- **On scrap paper, jot down a stem-and-leaf plot.**
- **Calculate the mean.**
- **Calculate the mean deviations.**
- **Calculate the squared mean deviations.**
- **Calculate the sum of squared mean deviations.**
- **Calculate the variance (dividing by n - 1).**
- **Calculate the standard deviation.**
- **Calculate the z-scores.**

Here is my shot:

Math Scores: 60 72 64 53 51 60 44 59 62 65
Please show your work:

Raw	Mean	Mean Deviation	Square Mean Deviation	Z-Score
60	59	1	1	0.126103
72	59	13	169	1.639344
64	59	5	25	0.630517
53	59	-6	36	-0.75662
51	59	-8	64	-1.00883
60	59	1	1	0.126103
44	59	-15	225	-1.89155
59	59	0	0	0
62	59	3	9	0.37831
65	59	6	36	0.75662

Please note the **mean** of the raw distribution:	**59**
Please note the **sum of squared mean deviations**:	**566**
Please note the **variance** of the raw distribution:	**62.9**
Please note the **standard deviation** of the raw distribution:	**7.9**

The Mean and Standard Deviation For Dichotomies: Example I

Let's say our sample is 50/50 males/females:

FEMALE: 0 1 0 0 1 1 1 0 0 1 1 0

Please show your work:

Raw	Mean	Mean Deviation	Square Mean Deviation	Z-Score
0	.5	-.5	.25	-.96
1	.5	.5	.25	.96
0	.5	-.5	.25	-.96
0	.5	-.5	.25	-.96
1	.5	.5	.25	.96
1	.5	.5	.25	.96
1	.5	.5	.25	.96
0	.5	-.5	.25	-.96
0	.5	-.5	.25	-.96
1	.5	.5	.25	.96
1	.5	.5	.25	.96
0	.5	-.5	.25	-.96

Please note the **mean** of the raw distribution:	.5
Please note the **sum of squared mean deviations**:	3
Please note the **variance** of the raw distribution:	.27
Please note the **standard deviation** of the raw distribution:	.52

Conceptually, when the mean of a dichotomous variable is .5, the standard deviation should be .5 and the z-scores should be 1 or -1, but the pesky degrees of freedom (n-1) fouls that up when n is small.

Why (conceptually) should the standard deviation be .5? Recall that the standard deviation is how wrong the mean is on average. Well, if for all the zeroes the mean is wrong by .5, and if for all the ones the mean is wrong by .5, and we only have zeroes and ones, then on average the mean should be wrong by .5.

How do the pesky degrees of freedom (n-1) foul things up? When n is large (e.g., 500), it hardly matters whether you divide the sum of squared deviations by 500 or 499 to get the variance. However, when n is small (e.g., 12), it makes a difference whether you divide the sum of squared mean deviations by 12 or 11 to get the average squared mean deviation. If we divided by 12, the sample size, then the standard deviation would work out to be .5 as expected, but instead we divide by 11, the degrees of freedom, so the standard deviation is a little larger than .5.

The Mean and Standard Deviation For Dichotomies: Example II

Let's say our sample is 1/3 students eligible for free lunch:

FREELUNCH: 1 0 0 1 0 0 1 0 1 0 0 0

Please show your work:

Raw	Mean	Mean Deviation	Square Mean Deviation	Z-Score
1	.33	.67	.449	1.37
0	.33	-.33	.109	.67
0	.33	-.33	.109	.67
1	.33	.67	.449	1.37
0	.33	-.33	.109	.67
0	.33	-.33	.109	.67
1	.33	.67	.449	1.37
0	.33	-.33	.109	.67
1	.33	.67	.449	1.37
0	.33	-.33	.109	.67
0	.33	-.33	.109	.67
0	.33	-.33	.109	.67

Please note the **mean** of the raw distribution:	**.5**
Please note the **sum of squared mean deviations**:	**2.67**
Please note the **variance** of the raw distribution:	**.24**
Please note the **standard deviation** of the raw distribution:	**.49**

Why is the mean the proportion of ones?

In our sample of 12, we have 4 students who are eligible for free lunch. Each of those 4 students gets a 1 for *FREELUNCH*, and everybody else gets a 0. When we add up the values for *FREELUNCH*, we are actually counting the number of students eligible for free lunch. (That is the beauty of 0/1 coding for dichotomies, also unfortunately known as "dummy coding.") When we take the average, we are dividing the total number of eligible students by the total number of students in our sample, thus we get the proportion of eligible students in our sample: .33, or 1/3, or 33%.

The trick to naming dummy variables:

You can name variables anything you want, but there is an especially helpful naming convention for dummy variables. You should name the variable after the thing that gets a 1, so that the 1 stands for "Yes" and the 0 stands for "No."

Good Practice:
MALE, a variable where 1 = male and 0 = female
FEMALE, a variable where 1 = female and 0 = male

Bad Practice:
GENDER, a variable where 1 = male and 0 = female

Frequency Tables (A Side Note)

When you have dummy variables, the mean is an easy way to get a handle on the frequency (or count), but you can always get a direct handle using your statistical software.

SPSS Syntax and Output

```
FREQUENCIES VARIABLES=FREELUNCH
  /ORDER=ANALYSIS.
```

Statistics

FREELUNCH

N	Valid	12
	Missing	0

FREELUNCH

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	8	66.7	66.7	66.7
	1	4	33.3	33.3	100.0
	Total	12	100.0	100.0	

In general, don't feel like you have to understand every single number that your package spits out. (I know I don't.) SPSS especially fires out strange statistics. Never, ever, ivitty ever use a statistic that you don't understand just because your computer barfed it.

R Syntax and Output

```
table(FREELUNCH, data=MYDATASET)
```

```
FREELUNCH
0 1
8 4
```

As usual, the SPSS output is prettier, but filled with extraneous information.

FYI: In SPSS, the difference between "Percent" and "Valid Percent" has to do with whether missing data are counted in the total. Because there is no missing data here, the Percent and Valid Percent are identical.

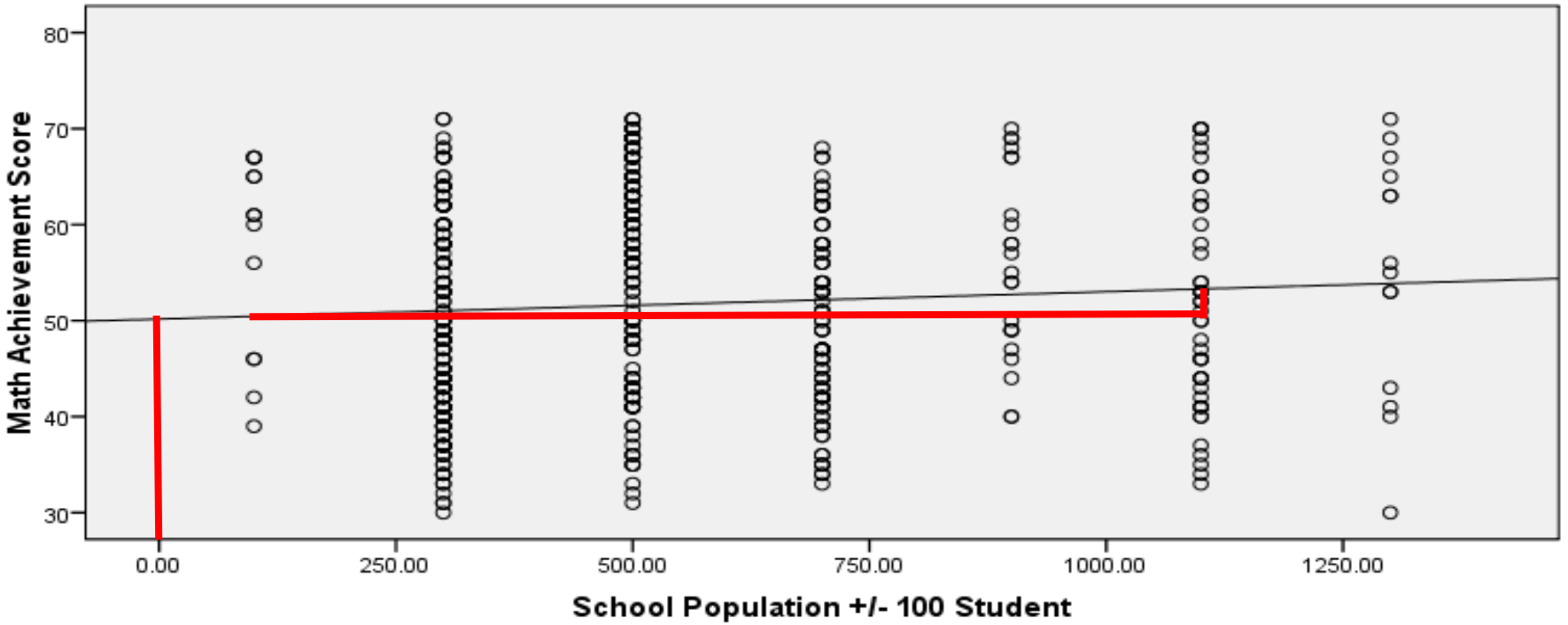
Re-Examining Math Achievement and School Size

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	50.167	1.029		48.767	.000
	School Population +/- 100 Student	.003	.002	.075	1.700	.090

a. Dependent Variable: Math Achievement Score

$$\hat{MathAch} = 50.2 + 0.003 SchoolPop$$



Examining Standardized Math Achievement and Standardized School Size

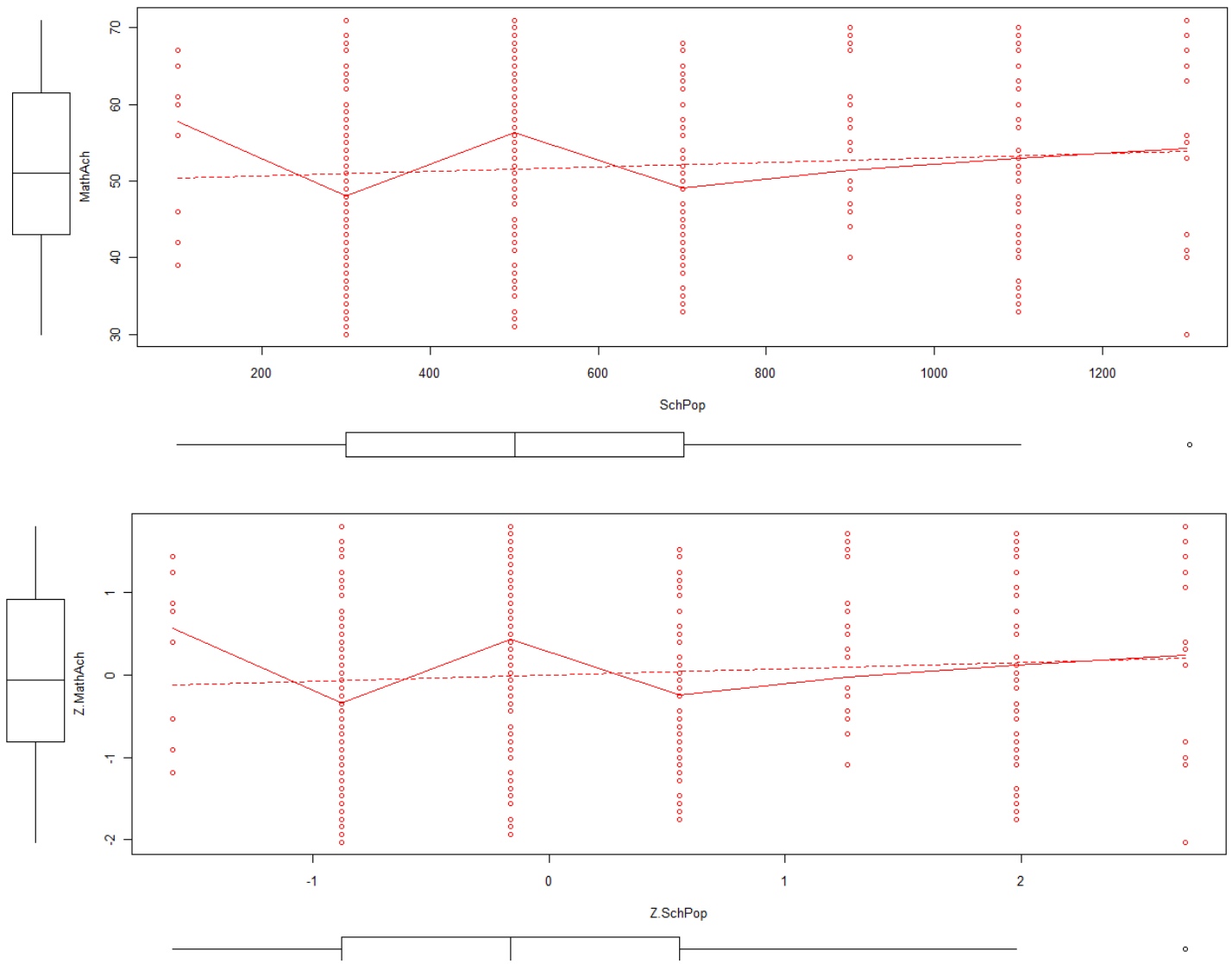
When you see an E in a number, that's a flag that scientific notation is at play. -4.943E-17 is really -0.00000000000000004943.

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-4.943E-17	.044		.000	1.000
	Zscore: School Population +/- 100 Student	.075	.044	.075	1.700	.090

a. Dependent Variable: Zscore: Math Achievement

$$\hat{ZMathAch} = 0.0 + 0.075 ZSchoolPop$$

Examining R Scatterplots



Notice that the plots have the same shape even though one uses standardized variables.

R Commander includes marginal boxplots as a default, and we can see that they remain the same.

Only the scales differ.

R Commander, as a default, includes a LOESS line (locally estimated scatterplot smoothing line). A LOESS line just finds the conditional means and connects the dot.

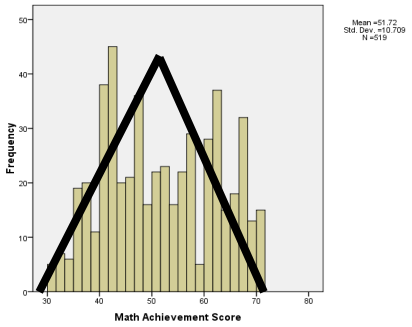
R Commander, as a default, also includes a now familiar OLS regression line.

Nonlinear Transformation Examples (Subtle)

Raw

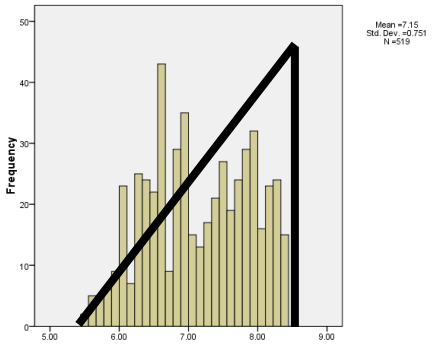
Note that the black lines are only to guide the eye.

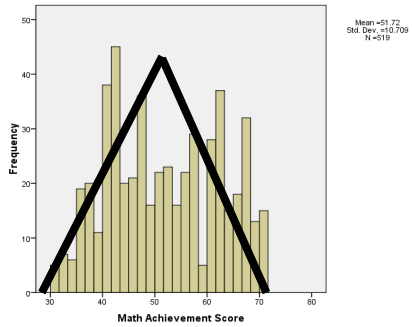
Transformed



Square root transformations expand the lower tail and contract the upper tail.

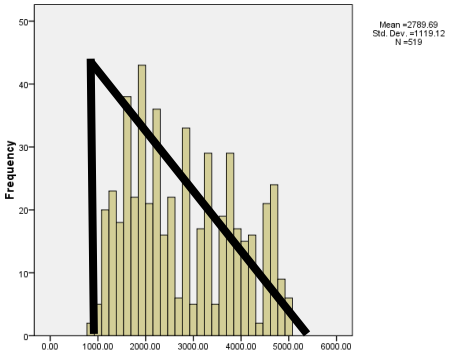
```
Compute MathAchSQRT = MathAch**(1/2).
```

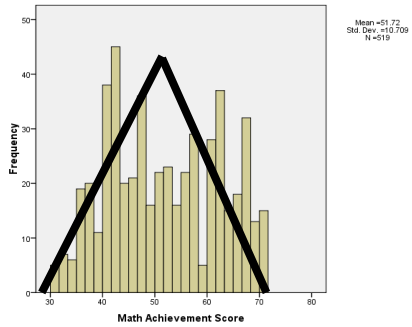




Square transformations contract the lower tail and expand the upper tail.

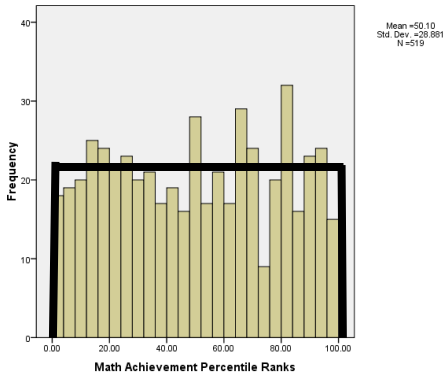
```
Compute MathAchSQ = MathAch**(2).
```





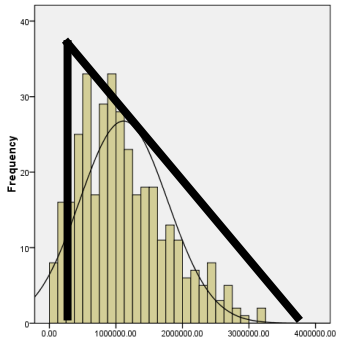
Percentile rank transformations flatten the distribution.

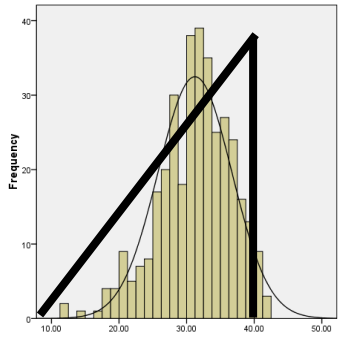
```
RANK VARIABLES=MathAch (A)
  /PERCENT
  /TIES=MEAN.
```



Nonlinear Transformation Examples (Obvious)

Raw





Note that the black lines are only to guide the eye.

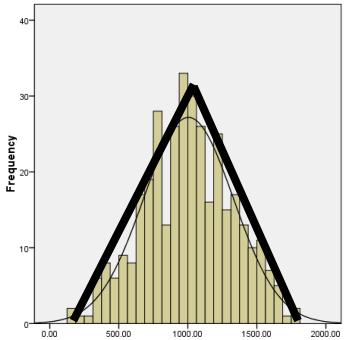
Square root transformations expand the lower tail and contract the upper tail.

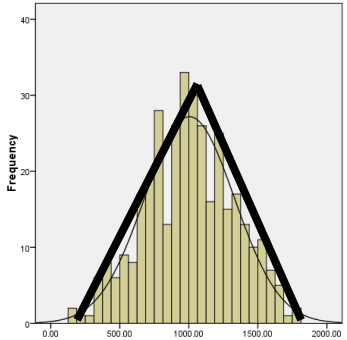
Compute VARSQRT = VAR(1/2).**

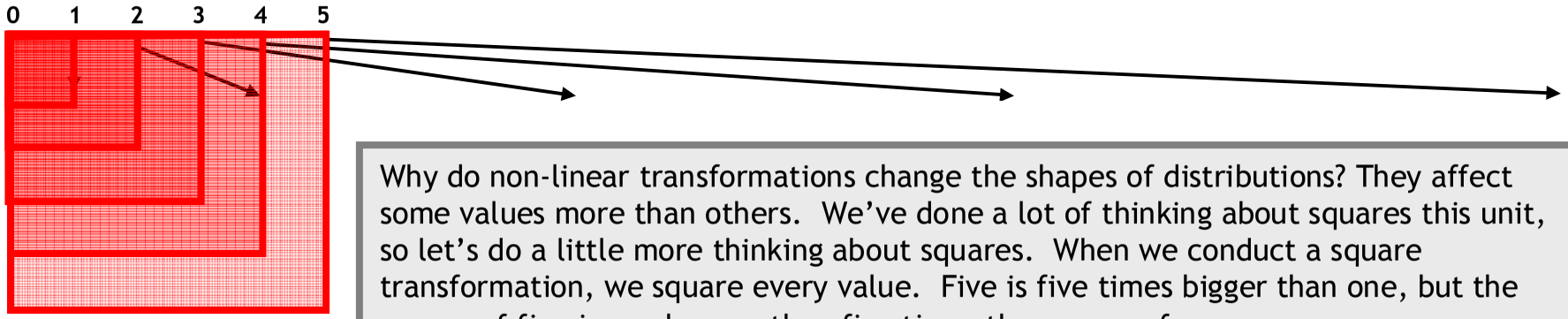
Square transformations contract the lower tail and expand the upper tail.

Compute VARSQ = VAR(2).**

Transformed



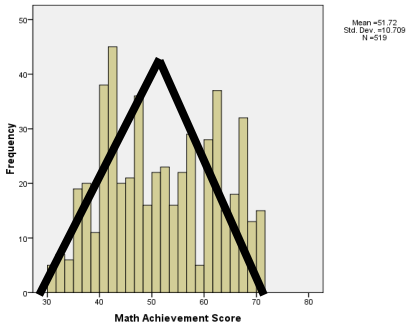




Why do non-linear transformations change the shapes of distributions? They affect some values more than others. We've done a lot of thinking about squares this unit, so let's do a little more thinking about squares. When we conduct a square transformation, we square every value. Five is five times bigger than one, but the square of five is much more than five times the square of one.

Linear Transformation Examples

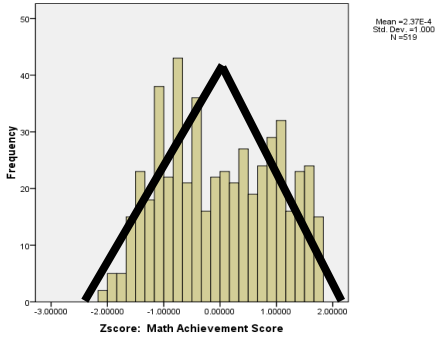
Raw

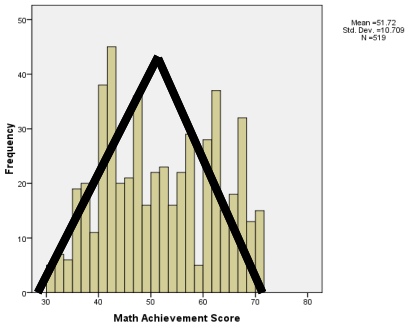


Z transforming (aka standardizing) does not change the shape of the distribution.

Compute ZMathAch = (MathAch-51.72)/10.71.

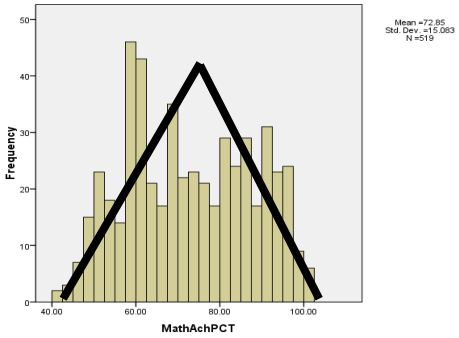
Transformed





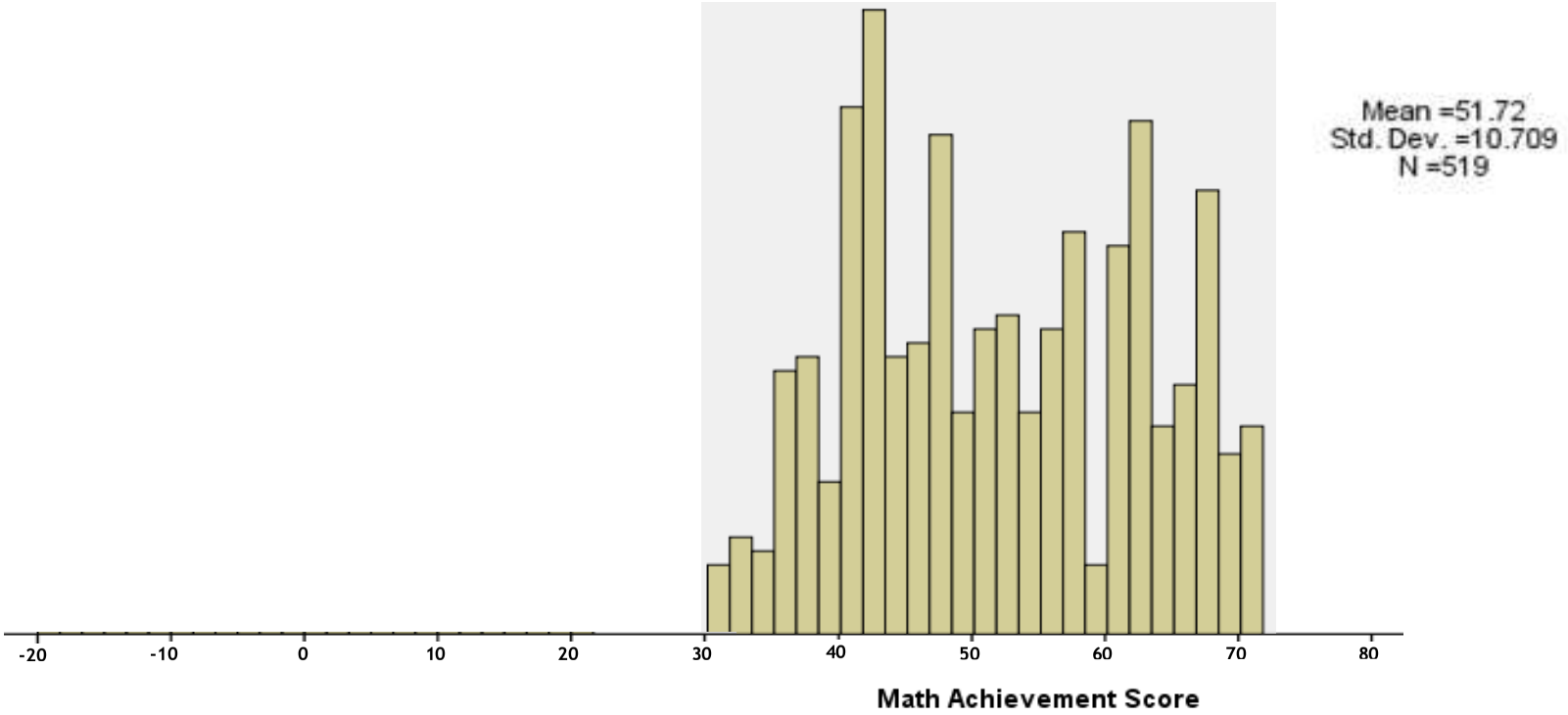
Transforming into percentages does not change the shape of the distribution.

Compute MathAchPCT = (MathAch/71)*100.



A linear transformation is one in which the only mathematical operations are addition/subtraction and multiplication/division. Notice that the linear equation, $y=mx+b$, uses only those basic operations. The addition/subtraction adjusts the mean of the distribution. The multiplication/division adjusts the standard deviation of the distribution. All the while, the shape of the distribution remains the same (although it may appear different due to rounding and binning).

Linear Transformations Are Shape Preserving



When we conduct a z transformation, we add/subtract something (the mean), and we multiply/divide by something (the standard deviation).

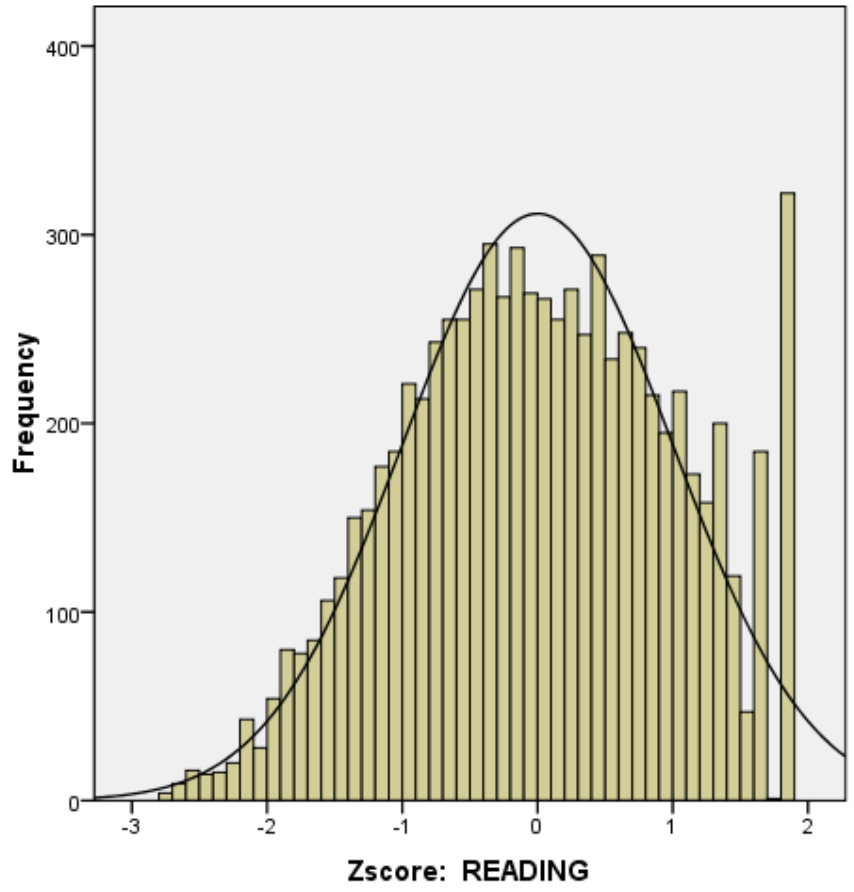
When we add/subtract something to/from every score, we simply move the whole distribution to the right or to the left, respectively. We change the ***location***!

When we multiply/divide every score by something, we just grow or shrink, respectively, the whole distribution. We change the ***spread***!

Linear transformations treat every score the same, so the distribution looks the same when we are done adding/subtracting/multiplying/dividing. We do not change the ***shape***!

Answering our Roadmap Question

Unit 3: In our sample, what does reading achievement look like (from an outlier sensitive perspective)?



Statistics

		READING	FREELUNCH
N	Valid	7800	7800
	Missing	0	0
Mean		47.4940	.3354
Std. Deviation		8.56944	.47216
Minimum		23.96	.00
Maximum		63.49	1.00
Percentiles	25	41.2400	.0000
	50	47.4300	.0000
	75	53.9300	1.0000

In our sample of 7,800 students, the distribution of reading scores has a mean of 47.49 and a standard deviation of 8.57. The distribution is approximately normal as we would expect from a purposefully designed standardized test. Because of a ceiling effect, however, no scores are more than two standard deviations above the mean, whereas scores below the mean tail off at close to negative three standard deviations. Despite this lack of symmetry, the distribution is generally symmetrical, and this is evidenced by a nearly identical mean and median, 47.49 and 47.43, respectively.

Unit 3 Appendix: Key Concepts

- The standard deviation measures how wrong, on average, the mean is as a prediction for individuals.
  - "Deviation" is distance from the mean.
  - "Standard" is the average.
- In a normal distribution, about 2/3 of observations fall within ± 1 standard deviation from the mean, and about 95% of observations fall within ± 2 standard deviations from the mean.
- The mean of a 0/1 dichotomous variable is the proportion of 1s. Also, for every such mean, there is only one possible standard deviation.
- A z-score (or standardized score) is a linear transformation of the raw score. From each raw score, we subtract the mean and divide by the standard deviation. Because we are only adding/subtracting and multiplying/dividing, we do not change the shape of the distribution (hence, *linear* transformation). In essence, we call the mean "zero" and we assign a value to everybody based on how many standard deviations they are from the mean.
- The mean is sensitive to outliers, and the standard deviation is more so. Not just doubly, but squarely so! (Also, sometimes the mean gets balanced out by two opposite extremes, but not the standard deviation.)
- Be reasonable with rounding. As data analysts, we are interested in meaningful differences. If there is no meaningful difference between 2.007 and 2, go with 2. Don't trust data analysts who don't round reasonably.

Unit 3 Appendix: Key Interpretations

Students in our sample go to schools of different sizes, the average student goes to a school of about 546 students (m = 546, sd = 280). The preponderance of students go to schools of between 266 and 826 students ($\pm$1 standard deviation from the mean). The distribution is positively skewed, so the few students from the largest schools, schools of approximately 1300 students, are exerting unreciprocated leverage on the mean, pulling the mean away from the median. (We may need to explain to our audience how more than half the students can go to smaller than average schools.)

The 519 students in our sample took a math achievement test, with a mean score of 52 and a standard deviation of 11. All but two students fall within $\pm$ 2 standard deviations from the mean with scores between 30 and 74, and the two exceptions fall just outside -2 standard deviations from the mean. The distribution is symmetric as evidenced by the nearly equal mean and median, but the distribution is bimodal, so it does not appear the students are clustering around one score, but rather two scores.

In our sample of 7,800 students, the distribution of reading scores has a mean of 47.49 and a standard deviation of 8.57. The distribution is approximately normal as we would expect from a purposefully designed standardized test. Because of a ceiling effect, however, no scores are more than two standard deviations above the mean, whereas scores below the mean tail off at close to negative three standard deviations. Despite this lack of symmetry, the distribution is generally symmetrical, and this is evidenced by a nearly identical mean and median, 47.49 and 47.43, respectively.

Unit 3 Appendix: Key Terminology

Note: I use "average" as a general term for location (or measure of central tendency), so for example means, medians, and modes are all averages in my book.

- Mean: The mean is a type of average. It is a measure of central tendency for a distribution. The mean is very common, but very abstract. It is very possible that the mean of a distribution is not even a value of the distribution. The mean is sensitive to outliers.
- Variance: The variance is the average squared mean deviation. When you take the average square, remember to divide by *n*-1, not *n* (i.e., divide by the degrees of freedom).
- Standard Deviation: Intuitively, the standard deviation is the average mean deviation. We know that it's a bit more complex.
- Z Transformation: A z transformation is a consistent manipulation of the distributional values that sets the mean to zero and the standard deviation to one. We add/subtract the same number to every value, and we multiply/divide the same number to every values. Because we are only adding/subtracting/ multiplying/dividing, we do not change the shape of the distribution; in other words, our transformation is linear.
- Z Score or Standardized Score: A z-score (or standardized score) is the score for an individual once the distribution has been z transformed. The average (i.e., mean) z score is, by definition, zero.
- Outlier Sensitive: Outlier sensitive statistics give outliers a lot of influence based on their distance(s) from the center, sometime based on their squared distance(s) from the center.
- Outlier Resistant: Outlier resistant statistics refuse to give outliers inordinate influence. They usually think in terms of rank orders instead of distances.

Unit 3 Appendix: Math

Mean of a Sample:

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

Mean of a Population:

$$\mu = \frac{\sum x}{N}$$

Standard Deviation of a Sample:

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

Standard Deviation of a Population:

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

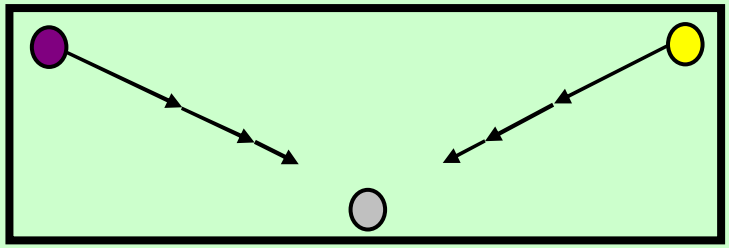
Z Transformation:

$$z = \frac{x - \bar{x}}{s}$$

Notes on Notation	
x denotes an observation, $\sum$, the summation sign, tells us to "add 'em up," so $\sum x$ tells us to add up all the observations.	
n = sample size	N = population size
We often use Greek letters to denote population statistics.	
$\bar{x}$ ="x bar" = sample mean	μ = mu ="mew"= population mean
s = sample standard deviation	σ = sigma = population s.d.

Unit 3 Appendix: Hilarious Hijinks

Once upon a time there was a distracting love triangle between a data analyst, a mathematician and a philosopher. The data analyst and the mathematician were both in love with the philosopher. The philosopher acted decisively. The philosopher gathered everybody together in a large, empty room. The philosopher put the data analyst and the mathematician in separate but adjacent corners. The philosopher then took a place across the room from both of them, centered on the opposite wall.



Moral of the Story: Be reasonable with rounding. As data analysts, we are interested in meaningful differences. If there is no meaningful difference between 2.007 and 2, go with 2. Don't trust data analysts who don't round reasonably.

The philosopher said, "Each of you, give me reasons that we should be together, and, for each of your reasons, cross half the distance between us. The first of you to reach me is mine, and I am yours." The mathematician said a nice thing, "I am a curve, and you are my integral." The data analyst said a nice thing, "Alone, I am a skewed univariate distribution. Alone, you are a skewed univariate distribution. Together, we have a perfect bivariate relationship (r= 1.00)." This continued until they were both within arms' reach of the philosopher. The data analyst lovingly embraced the philosopher, and the mathematician laughed at the data analyst, "You fool! If, for each reason that you provide, you can only traverse half the distance to your goal, you cannot reach your goal unless you provide an infinite number of reasons." Finally, having come down from the height of ecstasy, the mathematician found that the data analyst and philosopher had gone off together.

(Adapted by SP, Source Unknown)

Unit 3 Appendix: SPSS and R Syntax

```
*************************************************************************.
*Here is the SPSS syntax for standardization.
*The key here is to have the mean and standard deviation ready to hand.
*Then it's just a matter of naming a new variable and telling SPSS how to manipulate an old variable
to get the
transformation that you desire.
*************************************************************************.
COMPUTE ZTOTAL=(TOTAL-965.92)/74.821.
EXECUTE.

* Here is the SPSS shortcut.
* Analyze > Descriptive Statistics > Desciptives… click "Save standardized values as variables".
DESCRIPTIVES VARIABLES=PctAdv Schmariable1 Schmariable 2
  /SAVE.

# Here is the R syntax for standardization.
# Through R Commander, it is very easy: Data > Manage Variables in Active Data Set > Standardize
Variables.
library(foreign, pos=4)
Dataset <-
  read.spss("E:/CD140 2010/Data Sets/NELS Math Achievement/NELS88Math.sav",
  use.value.labels=TRUE, max.value.labels=Inf, to.data.frame=TRUE)
.Z <- scale(Dataset[,c("MathAch")])
Dataset$Z.MathAch <- .Z[,1]
remove(.Z)
```

Perceived Intimacy of Adolescent Girls (Intimacy.sav)

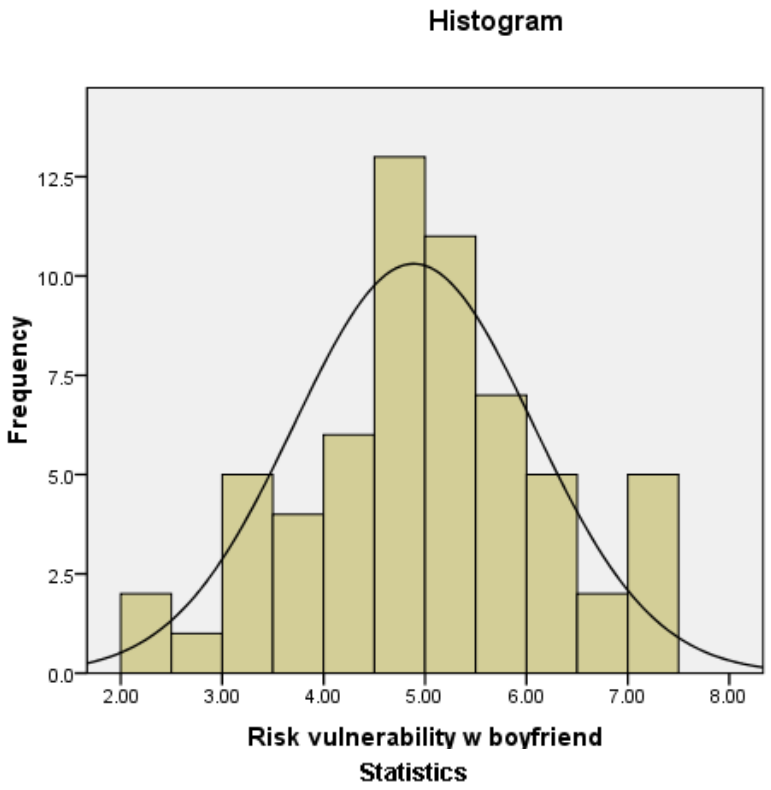
- **Overview: Dataset contains self-ratings of the intimacy that adolescent girls perceive themselves as having with: (a) their mother and (b) their boyfriend.**

- **Source: HGSE thesis by Dr. Linda Kilner entitled Intimacy in Female Adolescent's Relationships with Parents and Friends (1991). Kilner collected the ratings using the Adolescent Intimacy Scale.**
- **Sample: 64 adolescent girls in the sophomore, junior and senior classes of a local suburban public school system.**
- **Variables:**

Self Disclosure to Mother (M_Seldis)
Trusts Mother (M_Trust)
Mutual Caring with Mother (M_Care)
Risk Vulnerability with Mother (M_Vuln)
Physical Affection with Mother (M_Phys)
Resolves Conflicts with Mother (M_Cres)

Self Disclosure to Boyfriend (B_Seldis)
Trusts Boyfriend (B_Trust)
Mutual Caring with Boyfriend (B_Care)
Risk Vulnerability with Boyfriend (B_Vuln)
Physical Affection with Boyfriend (B_Phys)
Resolves Conflicts with Boyfriend (B_Cres)

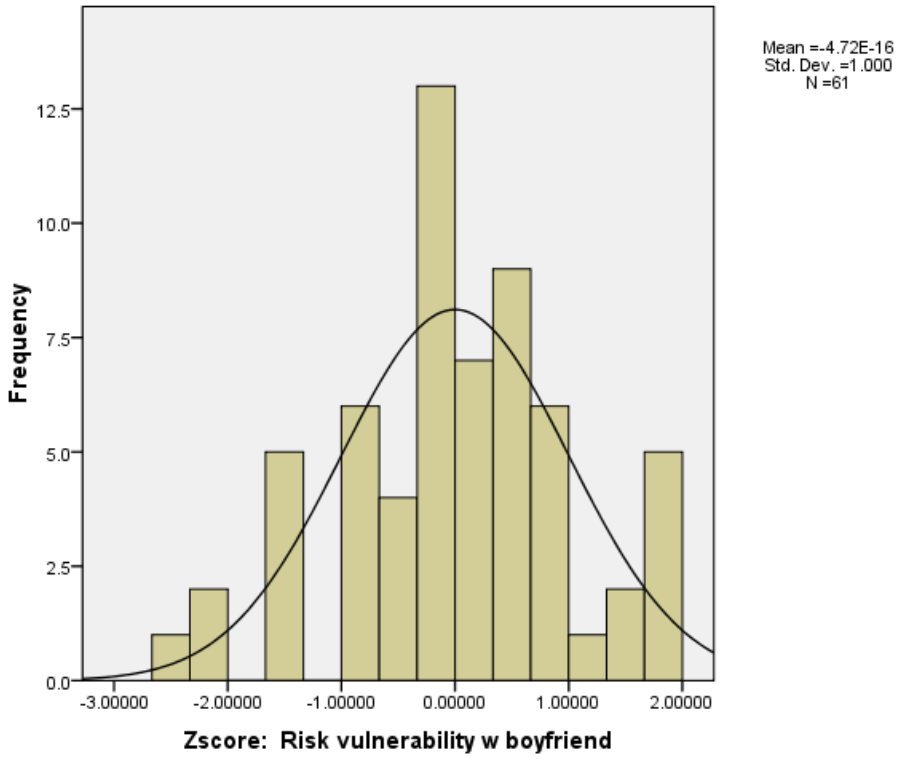
Perceived Intimacy of Adolescent Girls (Intimacy.sav)



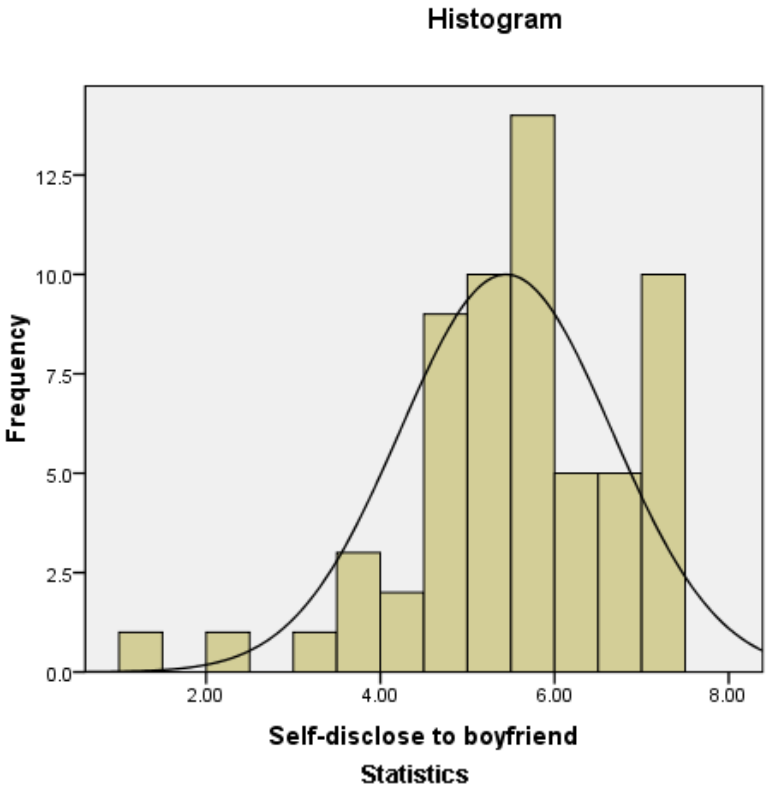
Statistics

Risk vulnerability w boyfriend

N	Valid	61.0000
	Missing	3.0000
Mean		4.8885
Std. Deviation		1.1804
Minimum		2.0000
Maximum		7.0000
Percentiles	25	4.3000
	50	4.8000
	75	5.5000



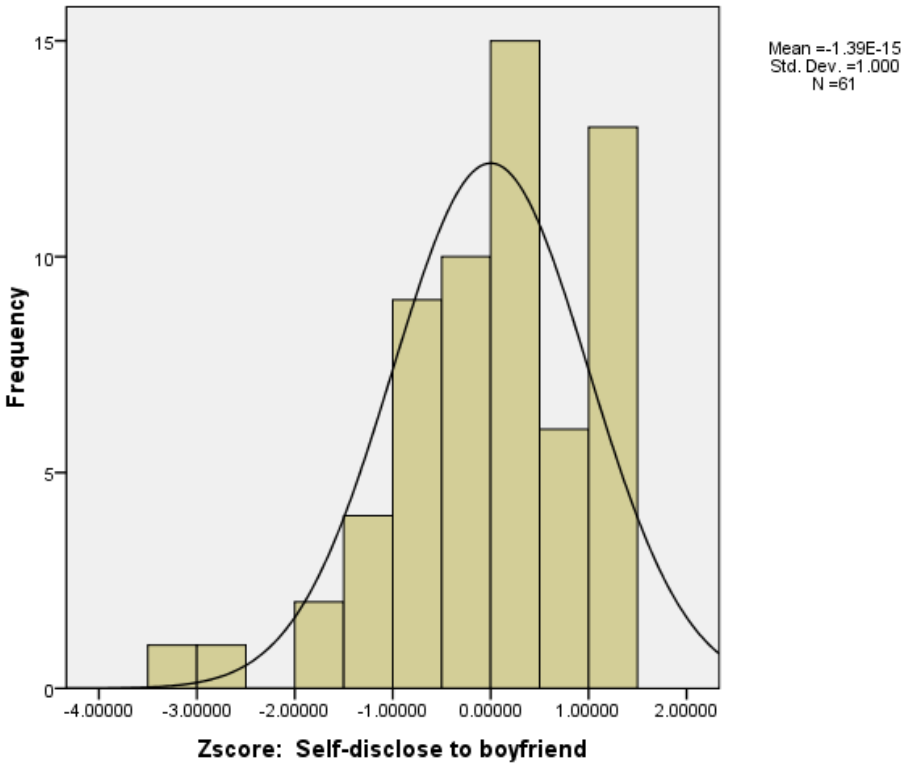
Perceived Intimacy of Adolescent Girls (Intimacy.sav)



Statistics

Self-disclose to boyfriend

N	Valid	61.0000
	Missing	3.0000
Mean		5.4426
Std. Deviation		1.2169
Minimum		1.3000
Maximum		7.0000
Percentiles	25	4.8000
	50	5.5000
	75	6.4000



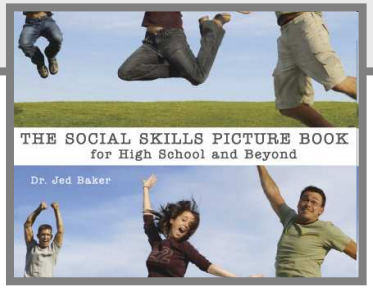
Perceived Intimacy of Adolescent Girls (Intimacy.sav)

Statistics

Self-disclose to mother

N	Valid	63.0000
	Missing	1.0000
Mean		4.0905
Std. Deviation		1.7381
Minimum		1.0000
Maximum		7.0000
Percentiles	25	2.8000
	50	4.0000
	75	5.5000

High School and Beyond (HSB.sav)



- **Overview: High School & Beyond – Subset of data focused on selected student and school characteristics as predictors of academic achievement.**
- **Source: Subset of data graciously provided by Valerie Lee, University of Michigan.**
- **Sample: This subsample has 1044 students in 205 schools. Missing data on the outcome test score and family SES were eliminated. In addition, schools with fewer than 3 students included in this subset of data were excluded.**
- **Variables:**

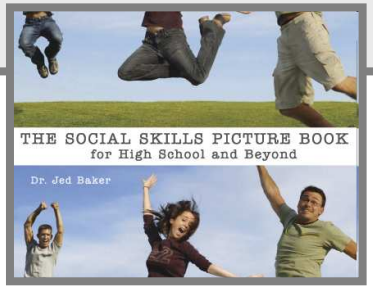
Variables about the student—

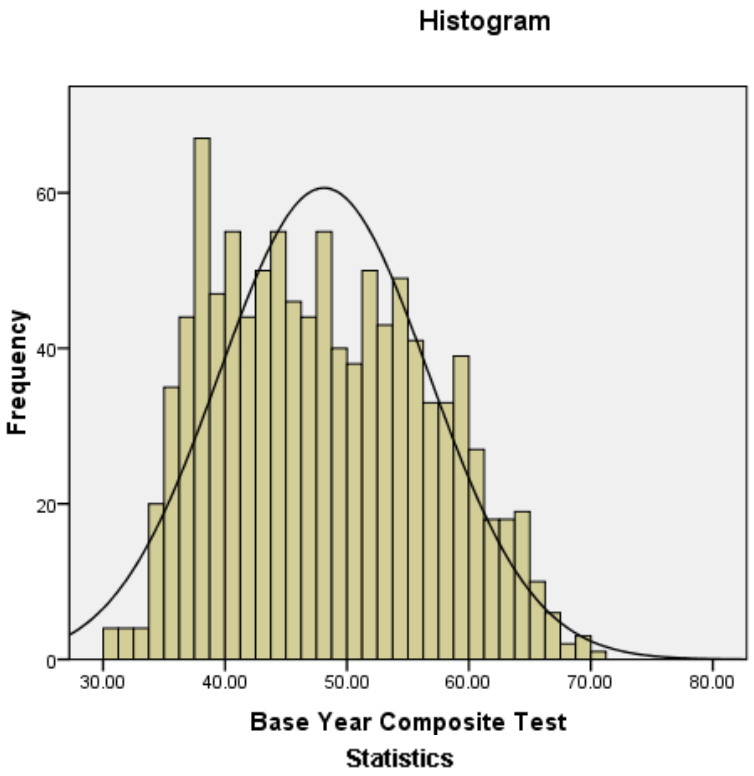
(Black) 1=Black, 0=Other
(Latin) 1=Latino/a, 0=Other
(Sex) 1=Female, 0=Male
(BYSES) Base year SES
(GPA80) HS GPA in 1980
(GPS82) HS GPA in 1982
(BYTest) Base year composite of reading and math tests
(BBConc) Base year self concept
(FEConc) First Follow-up self concept

Variables about the student's school—

(PctMin) % HS that is minority students Percentage
(HSSize) HS Size
(PctDrop) % dropouts in HS Percentage
(BYSES_S) Average SES in HS sample
(GPA80_S) Average GPA80 in HS sample
(GPA82_S) Average GPA82 in HS sample
(BYTest_S) Average test score in HS sample
(BBConc_S) Average base year self concept in HS sample
(FEConc_S) Average follow-up self concept in HS sample

High School and Beyond (HSB.sav)

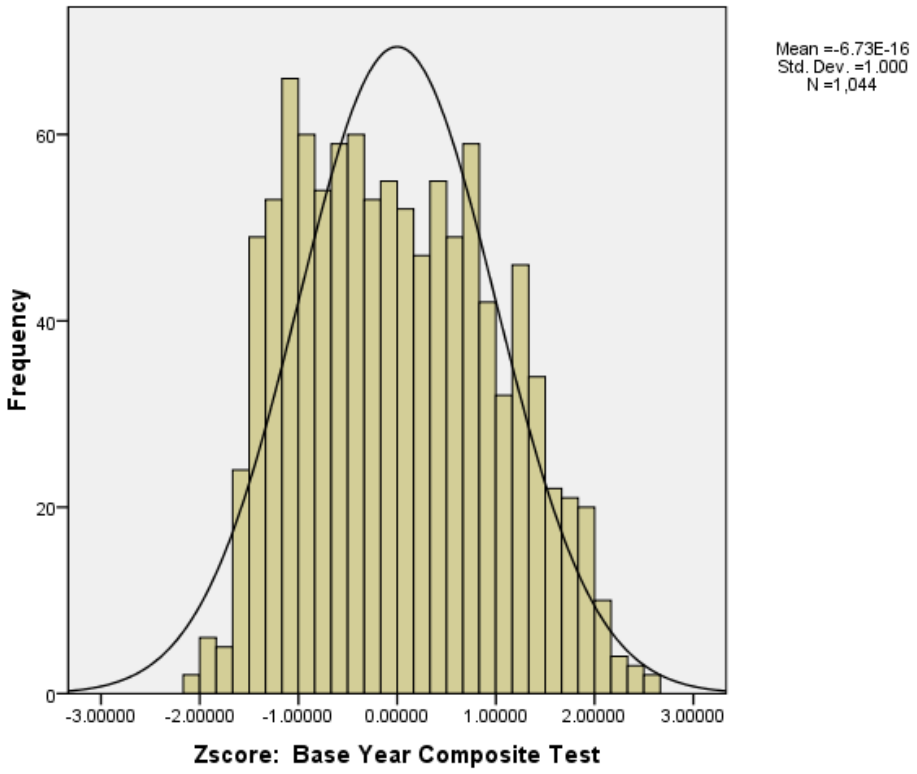




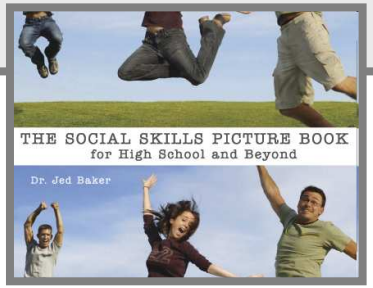
Statistics

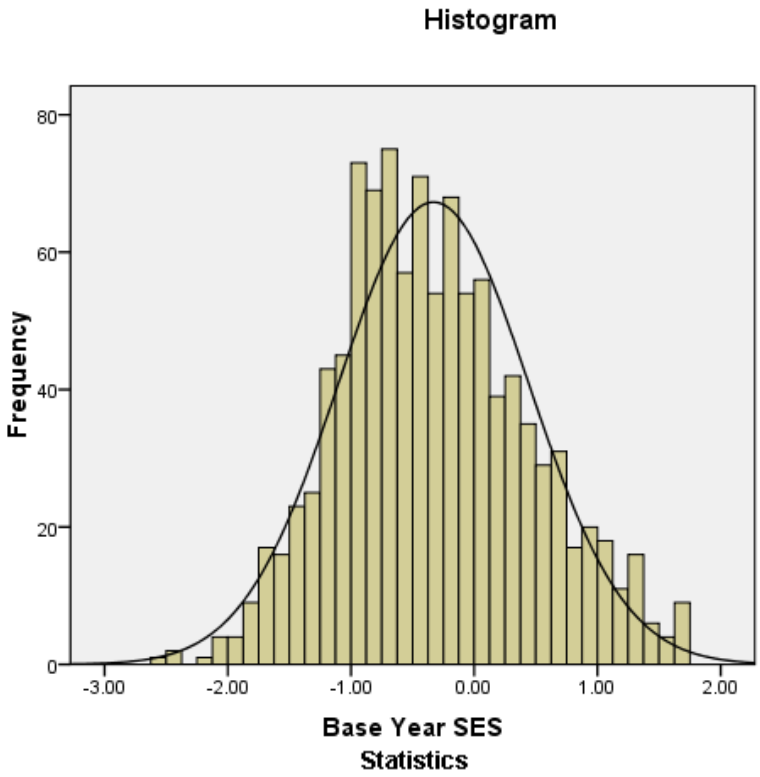
Base Year Composite Test

N	Valid	1044.0000
	Missing	.0000
Mean		48.1139
Std. Deviation		8.5876
Minimum		30.0400
Maximum		70.5600
Percentiles	25	40.7925
	50	47.5750
	75	54.8175



High School and Beyond (HSB.sav)

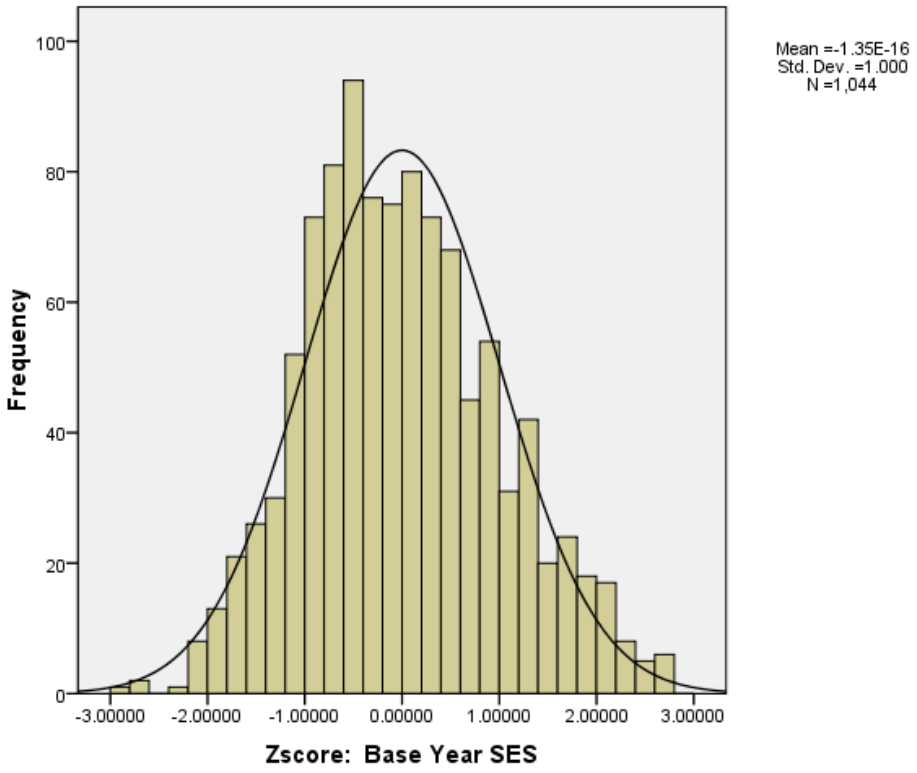




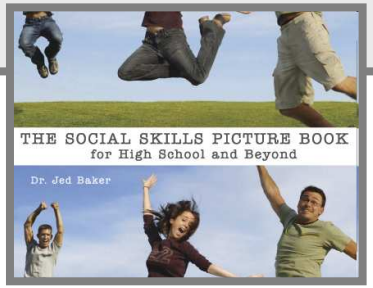
Statistics

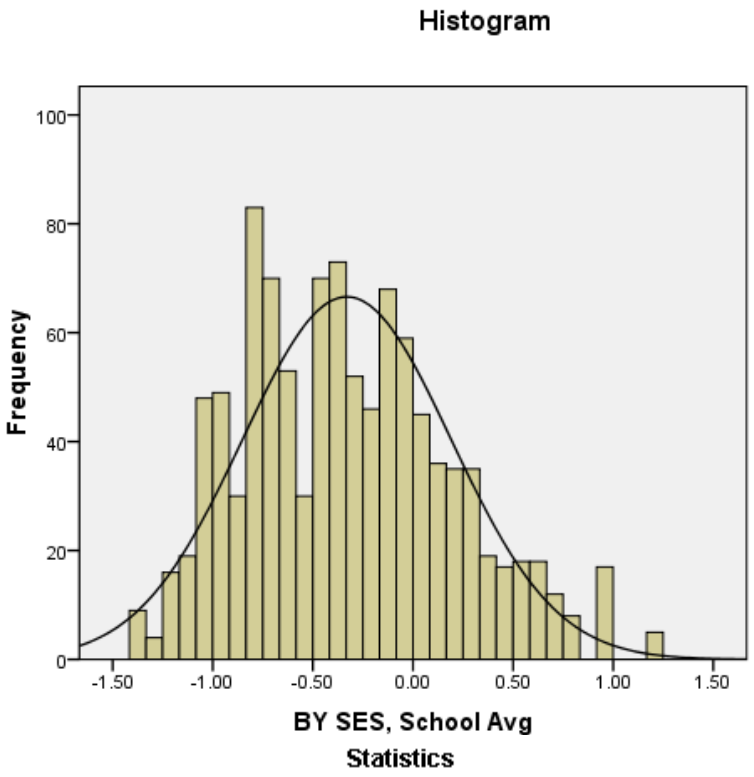
Base Year SES

N	Valid	1044.0000
	Missing	.0000
Mean		-.3304
Std. Deviation		.7736
Minimum		-2.5800
Maximum		1.6900
Percentiles	25	-.8800
	50	-.4100
	75	.1800



High School and Beyond (HSB.sav)

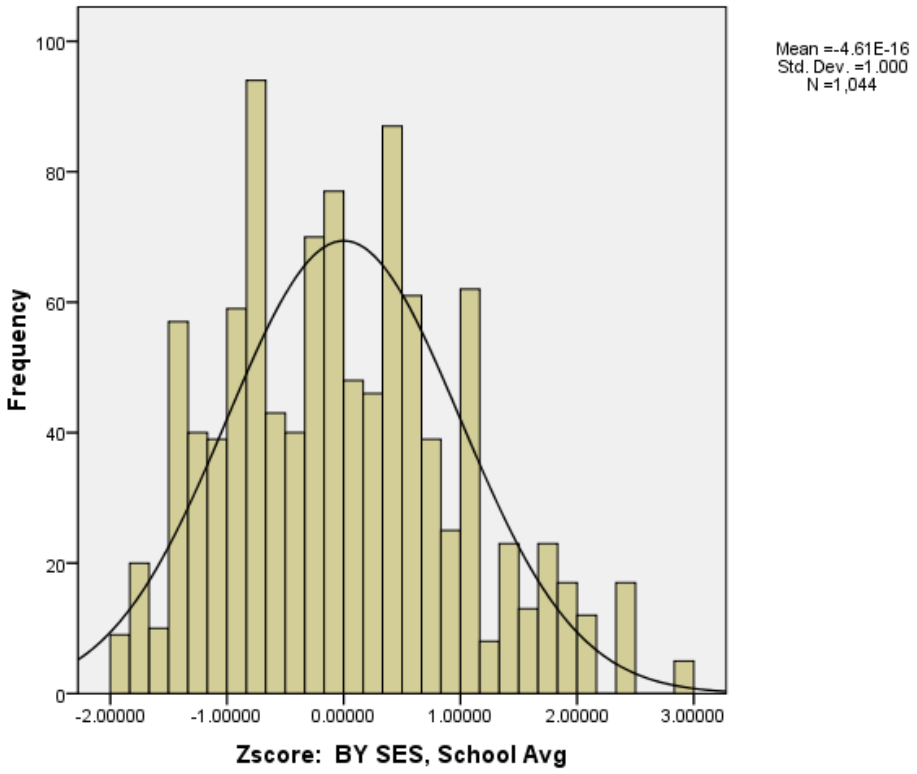




Statistics

BY SES, School Avg

N	Valid	1044.0000
	Missing	.0000
Mean		-.3304
Std. Deviation		.5211
Minimum		-1.3625
Maximum		1.2240
Percentiles	25	-.7483
	50	-.3640
	75	.0103

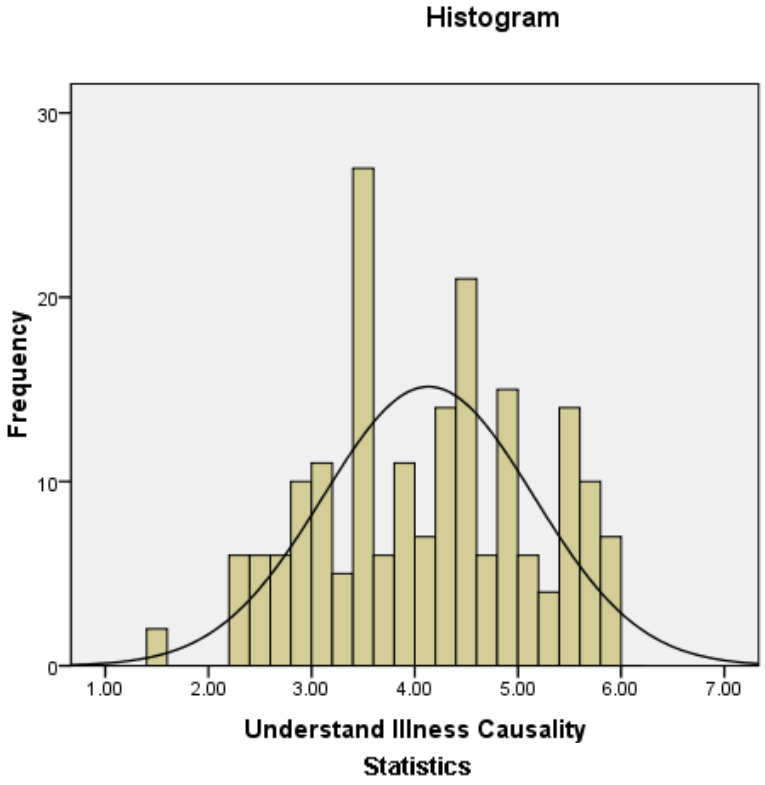


Understanding Causes of Illness (ILLCAUSE.sav)

- **Overview: Data for investigating differences in children's understanding of the causes of illness, by their health status.**
- **Source: Perrin E.C., Sayer A.G., and Willett J.B. (1991). Sticks And Stones May Break My Bones: Reasoning About Illness Causality And Body Functioning In Children Who Have A Chronic Illness, *Pediatrics*, 88(3), 608-19.**
- **Sample: 301 children, including a sub-sample of 205 who were described as asthmatic, diabetic,or healthy. After further reductions due to the *list-wise deletion* of cases with missing data on one or more variables, the analytic sub-sample used in class ends up containing: 33 diabetic children, 68 asthmatic children and 93 healthy children.**
- **Variables:**

(ILLCAUSE)	Child's Understanding of Illness Causality
(SES)	Child's SES (Note that a high score means low SES.)
(PPVT)	Child's Score on the Peabody Picture Vocabulary Test
(AGE)	Child's Age, In Months
(GENREAS)	Child's Score on a General Reasoning Test
(ChronicallyIll)	1 = Asthmatic or Diabetic, 0 = Healthy
(Asthmatic)	1 = Asthmatic, 0 = Healthy
(Diabetic)	1 = Diabetic, 0 = Healthy

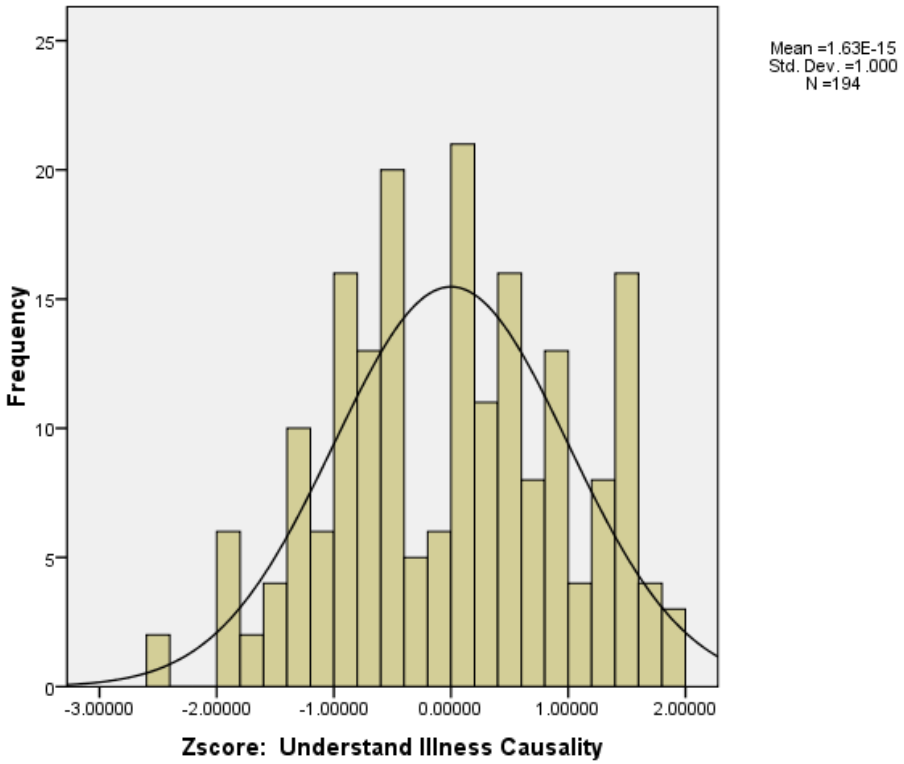
Understanding Causes of Illness (ILLCAUSE.sav)



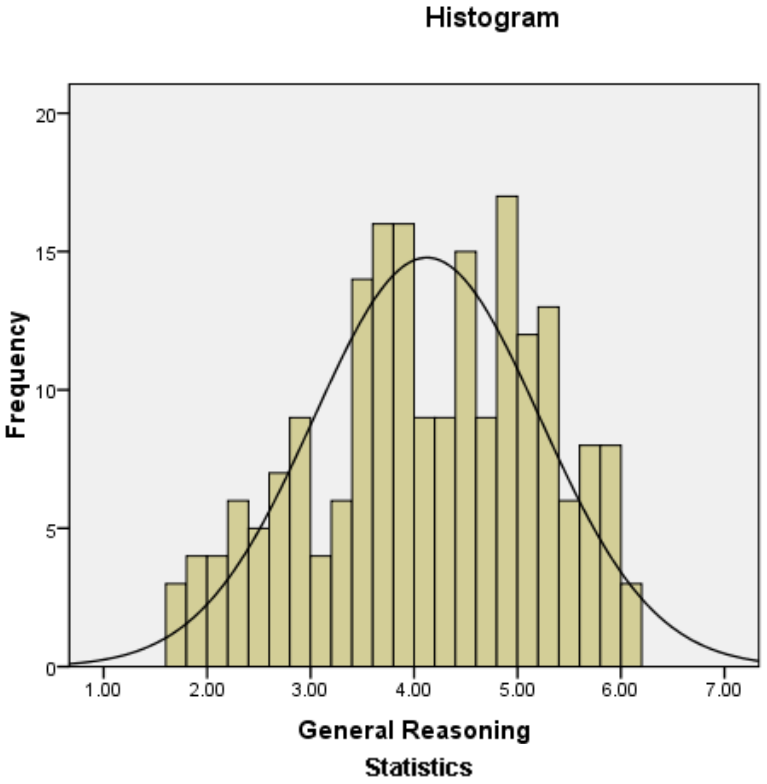
Statistics

Understand Illness Causality

N	Valid	194.0000
	Missing	11.0000
Mean		4.1333
Std. Deviation		1.0219
Minimum		1.5710
Maximum		6.0000
Percentiles	25	3.4290
	50	4.2145
	75	4.8928



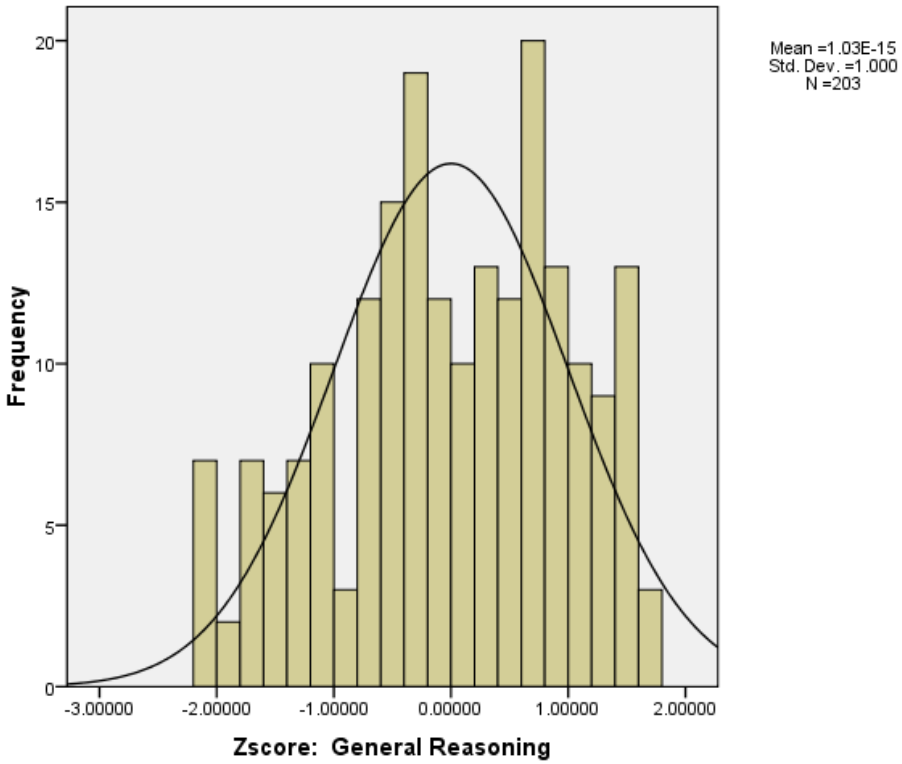
Understanding Causes of Illness (ILLCAUSE.sav)



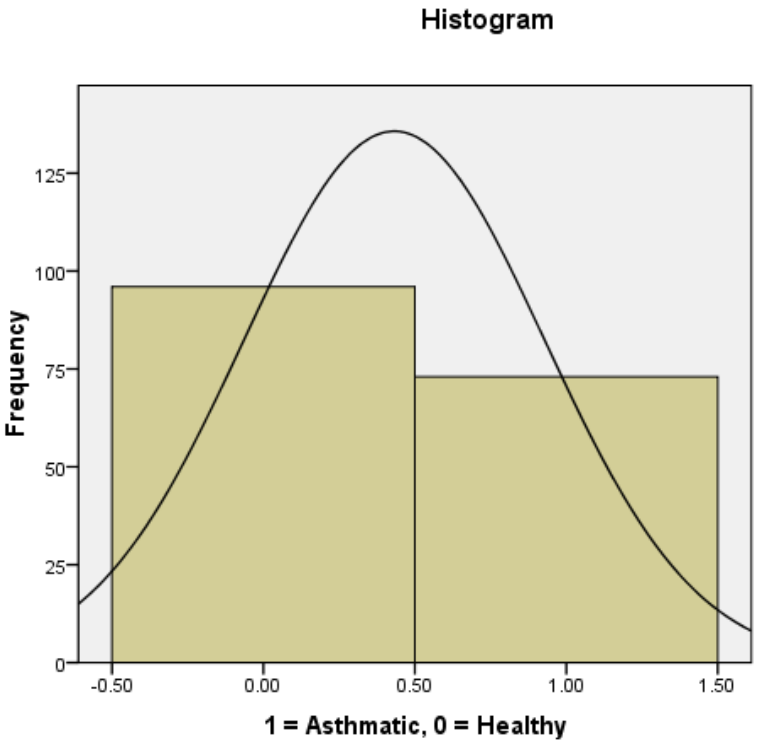
Statistics

General Reasoning

N	Valid	203.0000
	Missing	2.0000
Mean		4.1244
Std. Deviation		1.0957
Minimum		1.7500
Maximum		6.0000
Percentiles	25	3.4170
	50	4.1460
	75	4.9690



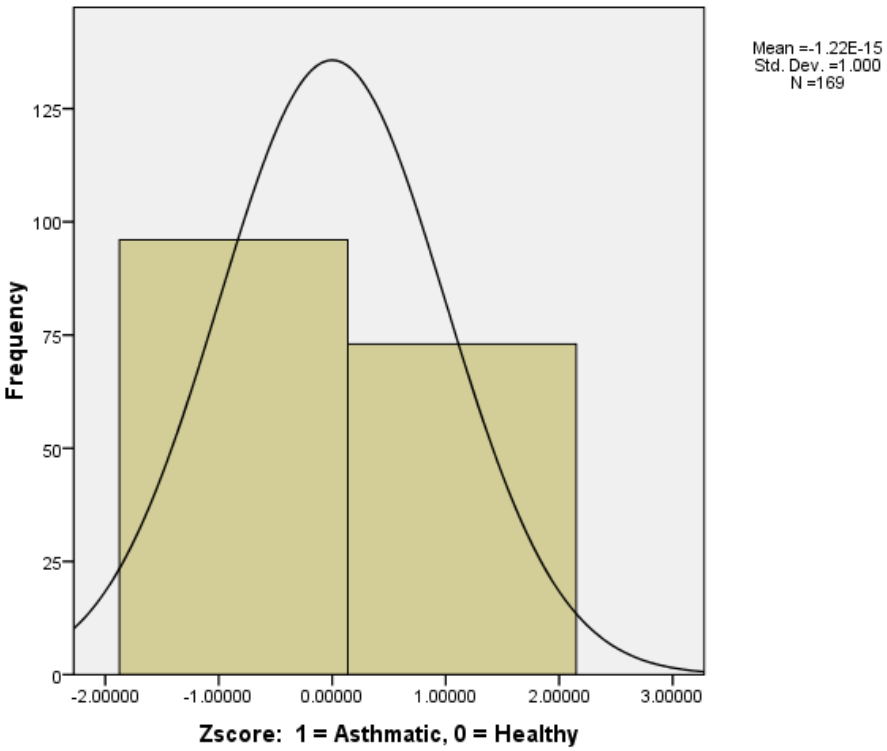
Understanding Causes of Illness (ILLCAUSE.sav)



Statistics

1 = Asthmatic, 0 = Healthy

N	Valid	169.0000
	Missing	36.0000
Mean		.4320
Std. Deviation		.4968
Minimum		.0000
Maximum		1.0000
Percentiles	25	.0000
	50	.0000
	75	1.0000



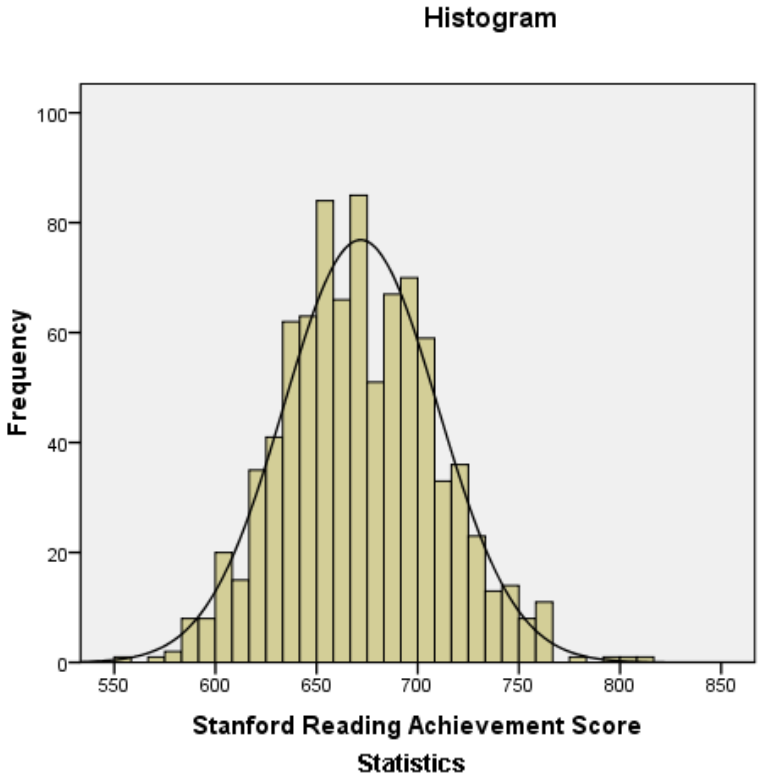
The mean of a 0/1 dichotomous variable is the proportion of 1s. Also, for every mean, there is only one possible standard deviation.

Children of Immigrants (ChildrenOfImmigrants.sav)

- **Overview: "CILS is a longitudinal study designed to study the adaptation process of the immigrant second generation which is defined broadly as U.S.-born children with at least one foreign-born parent or children born abroad but brought at an early age to the United States. The original survey was conducted with large samples of second-generation children attending the 8th and 9th grades in public and private schools in the metropolitan areas of Miami/Ft. Lauderdale in Florida and San Diego, California" (from the website description of the data set).**
- **Source: Portes, Alejandro, & Ruben G. Rumbaut (2001). *Legacies: The Story of the Immigrant SecondGeneration*. Berkeley CA: University of California Press.**
- **Sample: Random sample of 880 participants obtained through the website.**
- **Variables:**

(Reading)	Stanford Reading Achievement Score
(Freelunch)	% students in school who are eligible for free lunch program
(Male)	1=Male 0=Female
(Depress)	Depression scale (Higher score means more depressed)
(SES)	Composite family SES score

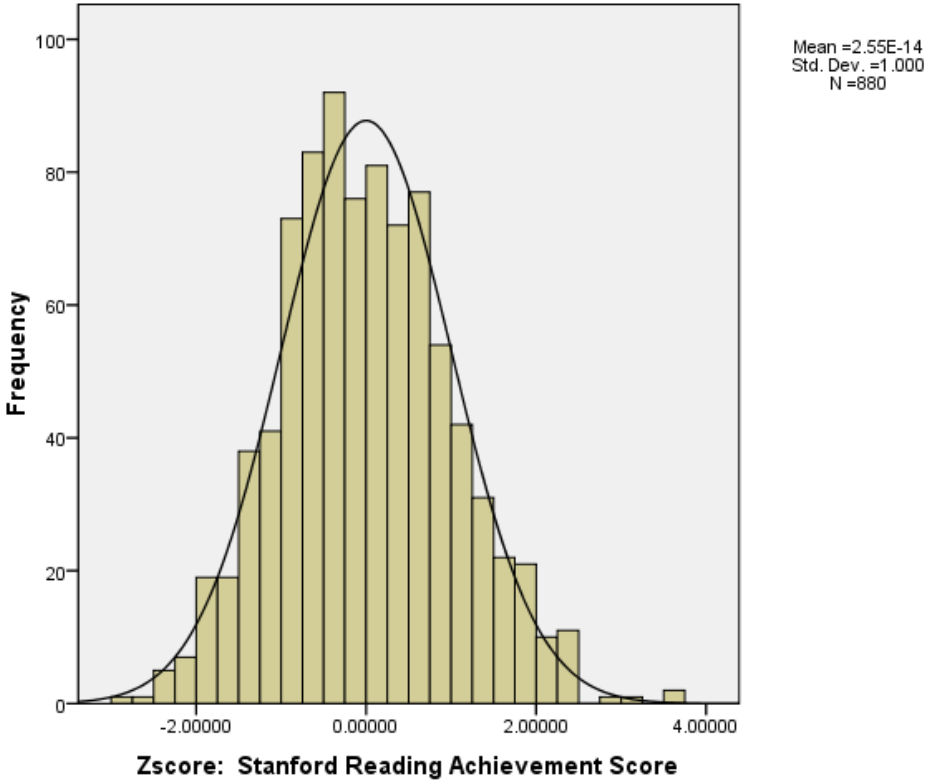
Children of Immigrants (ChildrenOfImmigrants.sav)



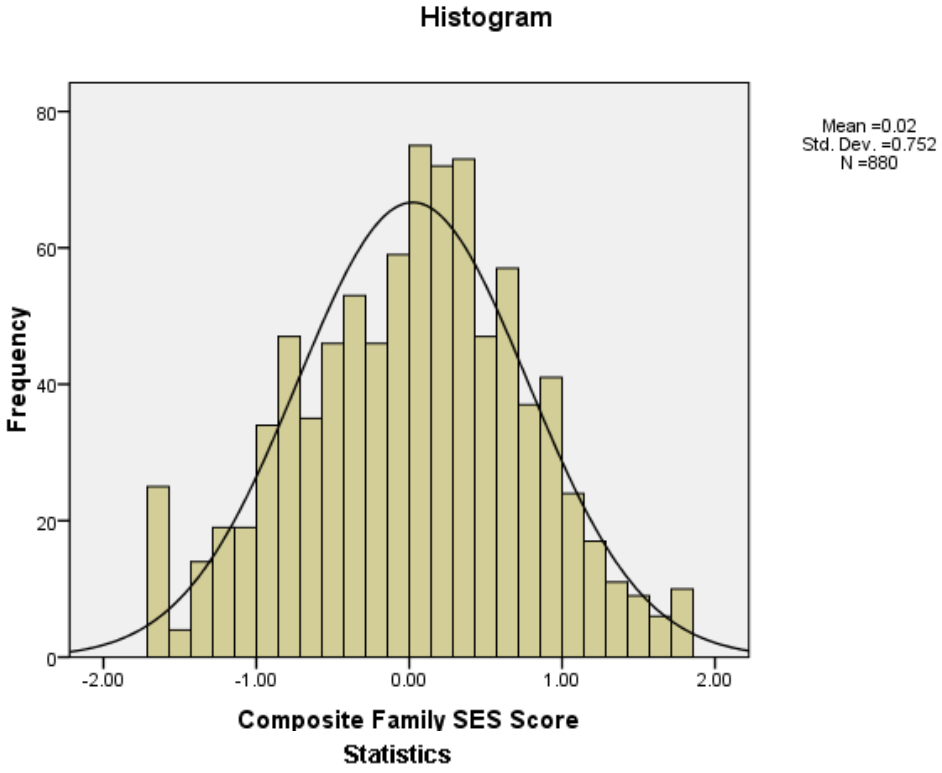
Statistics

Stanford Reading Achievement Score

N	Valid	880.00
	Missing	.00
Mean		671.82
Std. Deviation		38.05
Minimum		558.00
Maximum		813.00
Percentiles	25	646.00
	50	669.00
	75	697.00



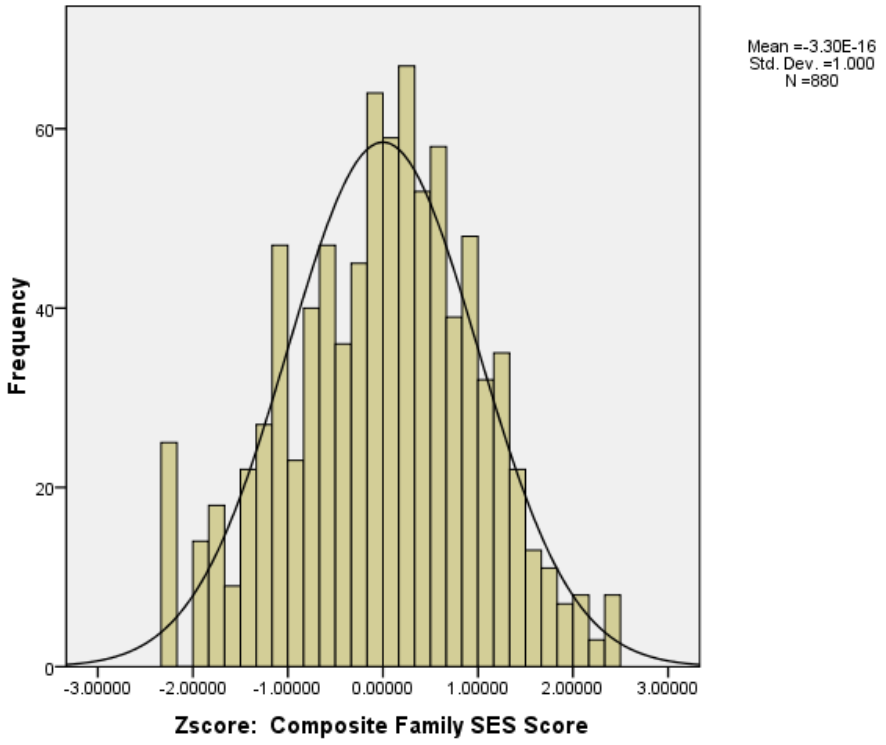
Children of Immigrants (ChildrenOfImmigrants.sav)



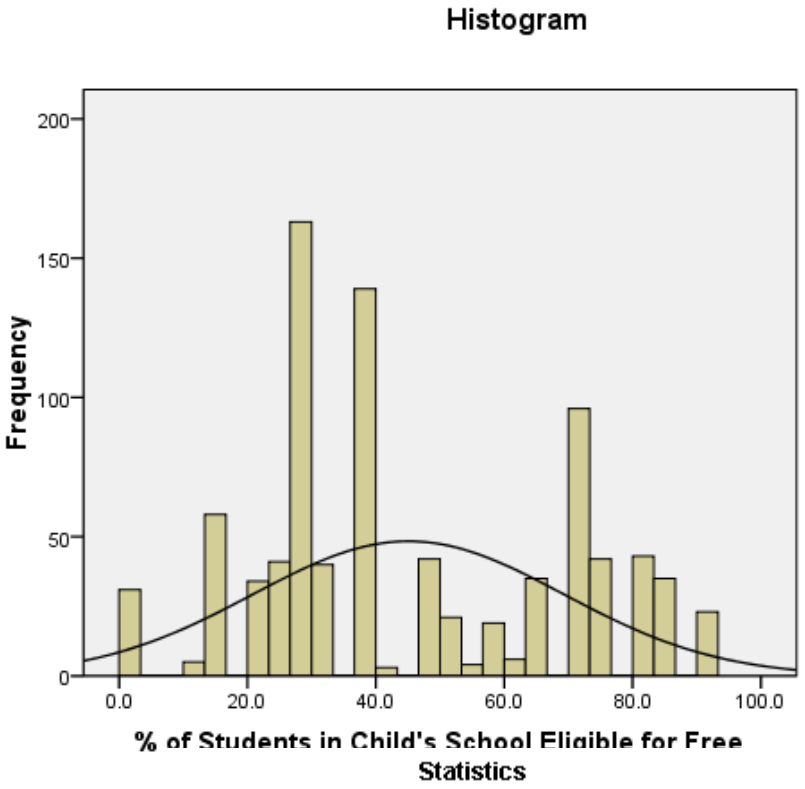
Statistics

Composite Family SES Score

N	Valid	880.0000
	Missing	.0000
Mean		.0228
Std. Deviation		.7522
Minimum		-1.6600
Maximum		1.8500
Percentiles	25	-.5100
	50	.0600
	75	.5575



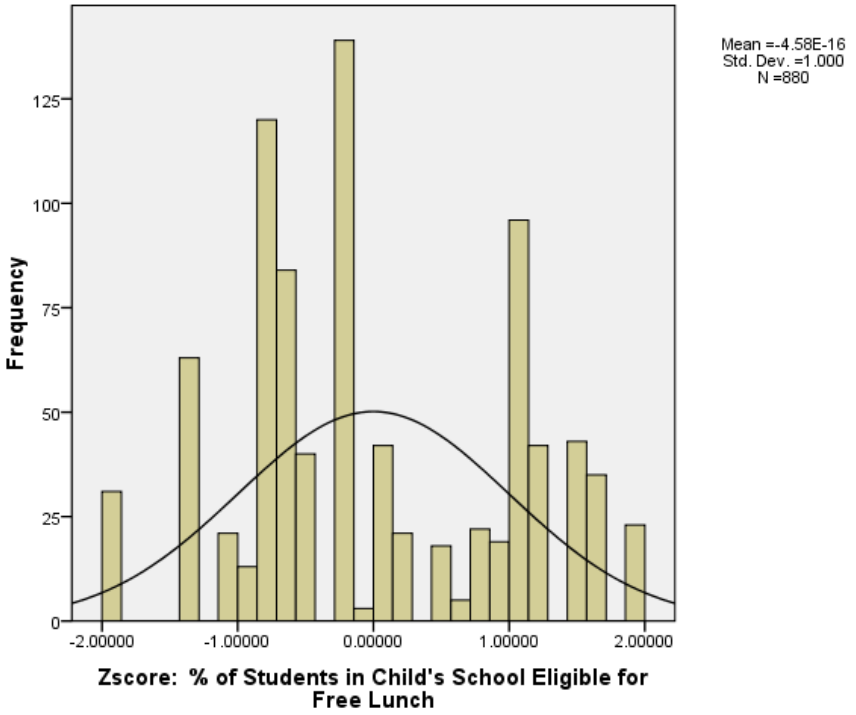
Children of Immigrants (ChildrenOfImmigrants.sav)



Statistics

% of Students in Child's School Eligible for Free Lunch

N	Valid	880.000
	Missing	.000
Mean		45.073
Std. Deviation		24.194
Minimum		.000
Maximum		92.300
Percentiles	25	27.500
	50	38.900
	75	72.200

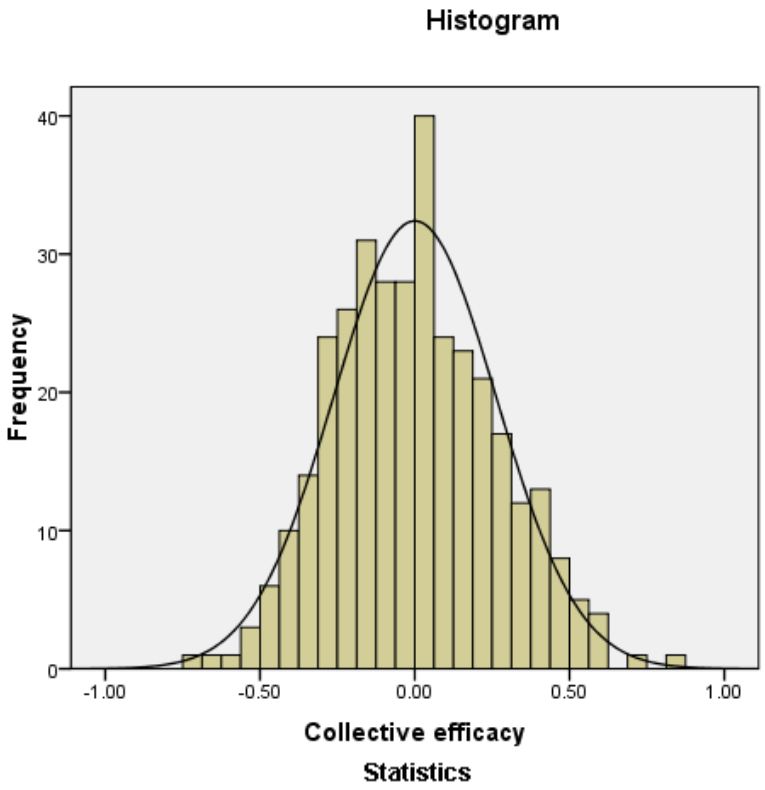


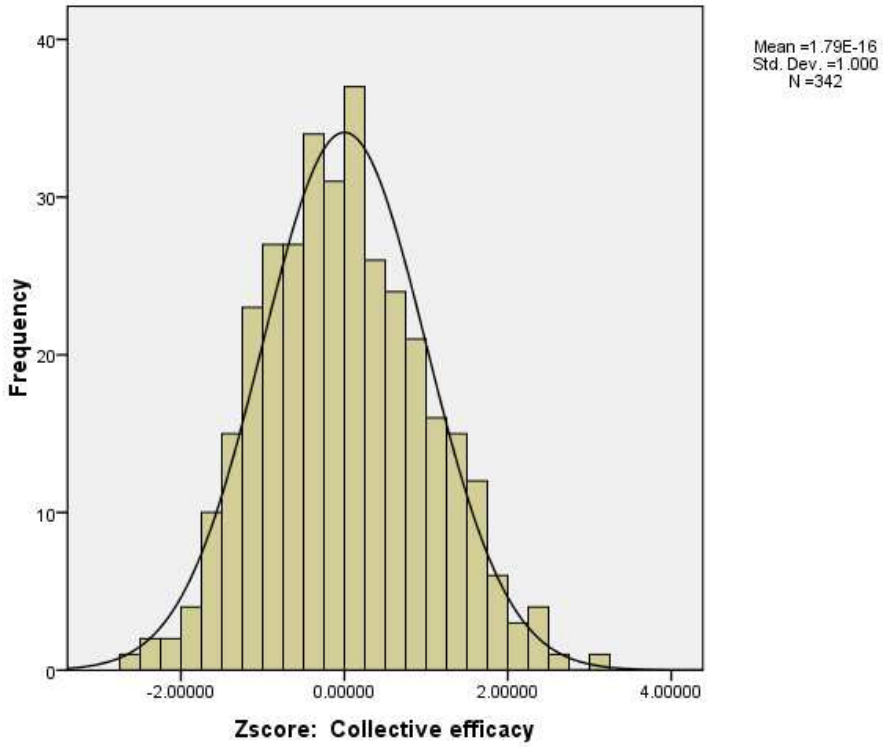
Human Development in Chicago Neighborhoods (Neighborhoods.sav)

- **These data were collected as part of the Project on Human Development in Chicago Neighborhoods in 1995.**
- **Source: Sampson, R.J., Raudenbush, S.W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, 277, 918-924.**
- **Sample: The data described here consist of information from 343 Neighborhood Clusters in Chicago Illinois. Some of the variables were obtained by project staff from the 1990 Census and city records. Other variables were obtained through questionnaire interviews with 8782 Chicago residents who were interviewed in their homes.**
- **Variables:**

(Homr90)	Homicide Rate c. 1990
(Murder95)	Homicide Rate 1995
(Disadvan)	Concentrated Disadvantage
(Imm_Conc)	Immigrant
(ResStab)	Residential Stability
(Popul)	Population in 1000s
(CollEff)	Collective Efficacy
(Victim)	% Respondents Who Were Victims of Violence
(PercViol)	% Respondents Who Perceived Violence

Human Development in Chicago Neighborhoods (Neighborhoods.sav)



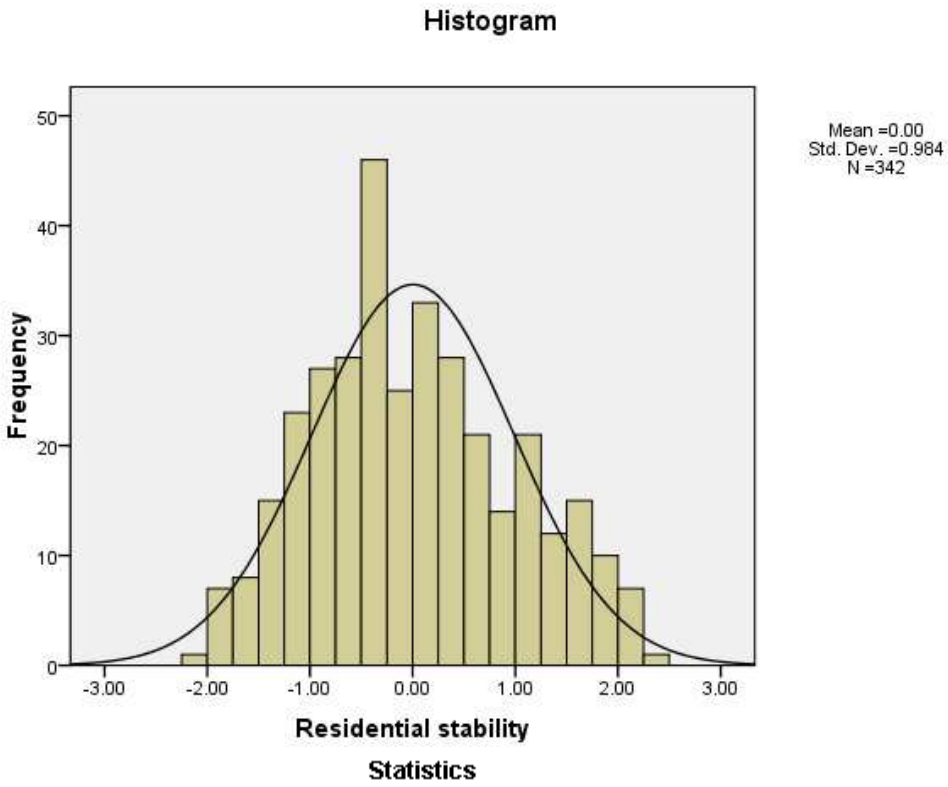


Statistics

Collective efficacy

N	Valid	342.0000
	Missing	.0000
Mean		.0002
Std. Deviation		.2631
Minimum		-.7100
Maximum		.8400
Percentiles	25	-.1900
	50	-.0100
	75	.1725

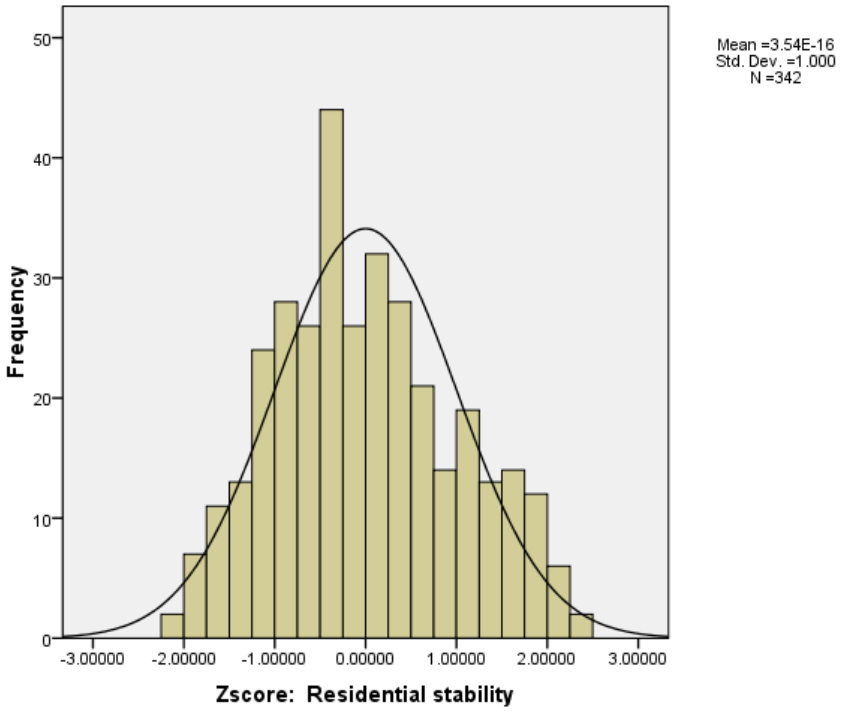
Human Development in Chicago Neighborhoods (Neighborhoods.sav)



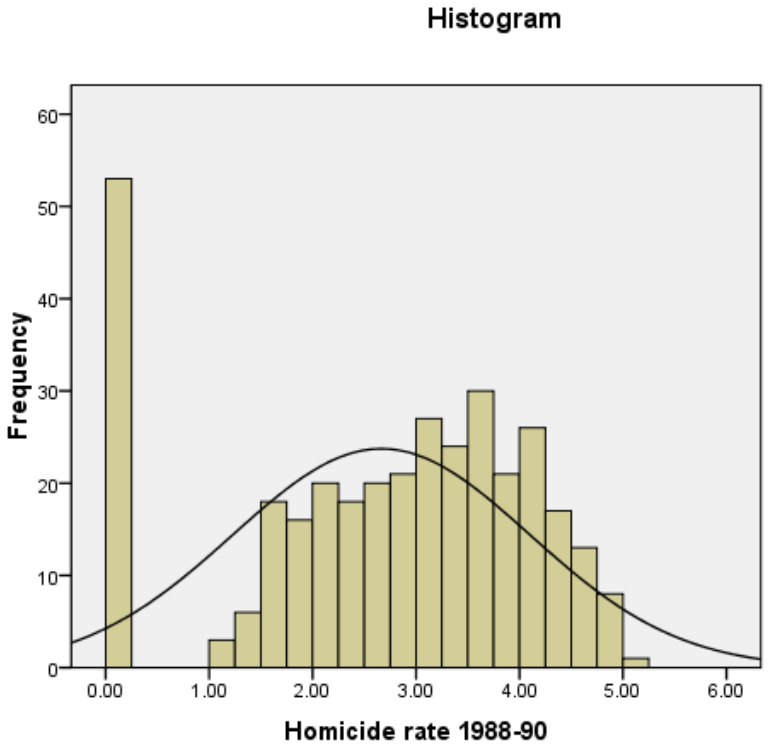
Statistics

Residential stability

N	Valid	342.0000
	Missing	.0000
Mean		.0027
Std. Deviation		.9843
Minimum		-2.1800
Maximum		2.3300
Percentiles	25	-.7325
	50	-.1050
	75	.6800



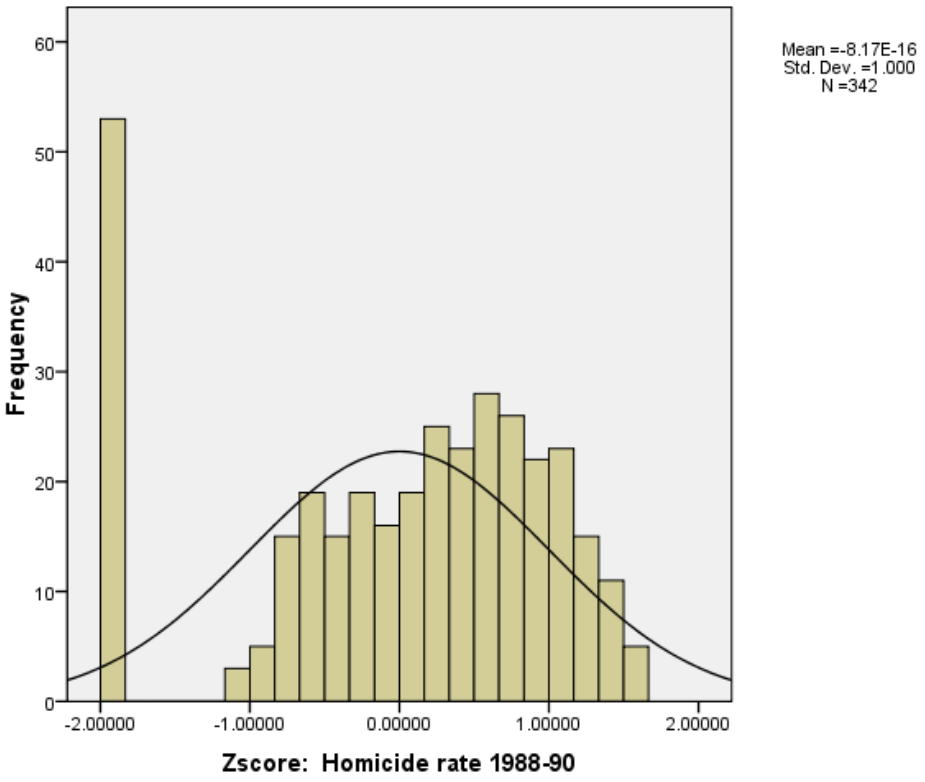
Human Development in Chicago Neighborhoods (Neighborhoods.sav)



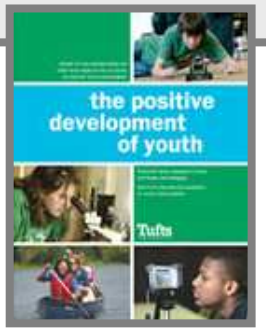
Statistics

Homicide rate 1988-90

N	Valid	342.0000
	Missing	.0000
Mean		2.6646
Std. Deviation		1.4373
Minimum		.0000
Maximum		5.0400
Percentiles	25	1.8300
	50	2.9600
	75	3.7525



4-H Study of Positive Youth Development (4H.sav)



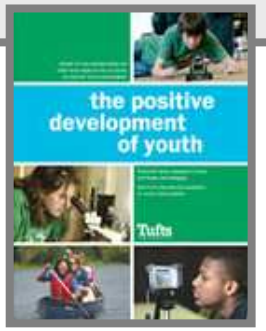
- 4-H Study of Positive Youth Development
- Source: Subset of data from IARYD, Tufts University

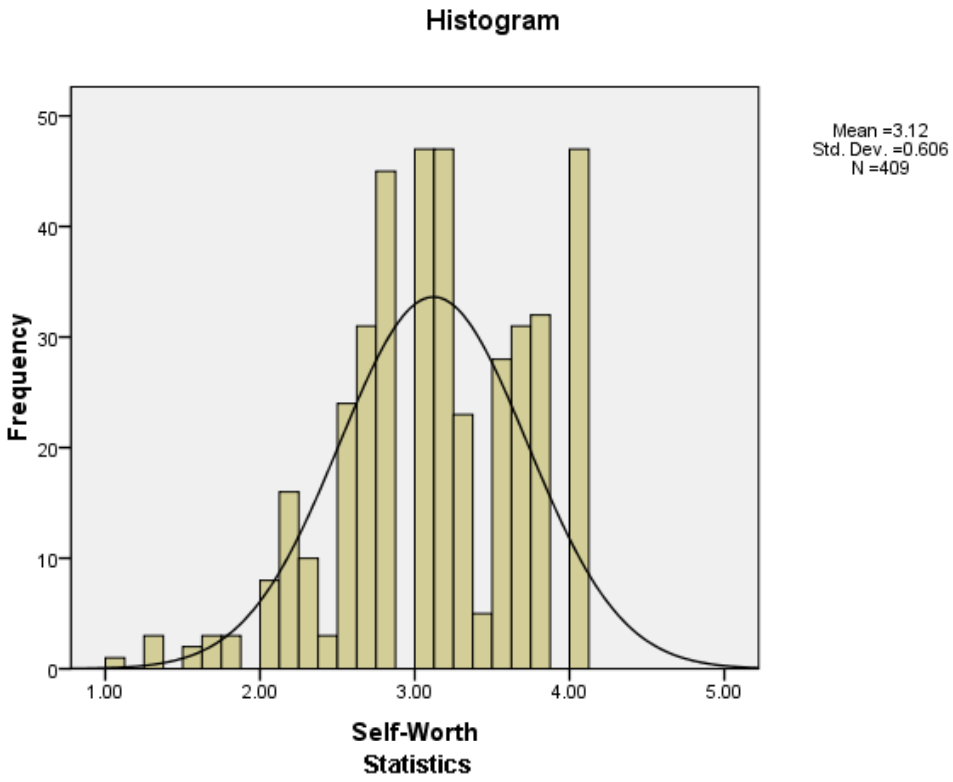
- Sample: These data consist of seventh graders who participated in Wave 3 of the 4-H Study of Positive Youth Development at Tufts University. This subfile is a substantially sampled-down version of the original file, as all the cases with any missing data on these selected variables were eliminated.
- Variables:

(SexFem)	1=Female, 0=Male
(MothEd)	Years of Mother's Education
(Grades)	Self-Reported Grades
(Depression)	Depression (Continuous)
(FrInfl)	Friends' Positive Influences
(PeerSupp)	Peer Support
(Depressed)	0 = (1-15 on Depression) 1 = Yes (16+ on Depression)

(AcadComp)	Self-Perceived Academic Competence
(SocComp)	Self-Perceived Social Competence
(PhysComp)	Self-Perceived Physical Competence
(PhysApp)	Self-Perceived Physical Appearance
(CondBeh)	Self-Perceived Conduct Behavior
(SelfWorth)	Self-Worth

4-H Study of Positive Youth Development (4H.sav)

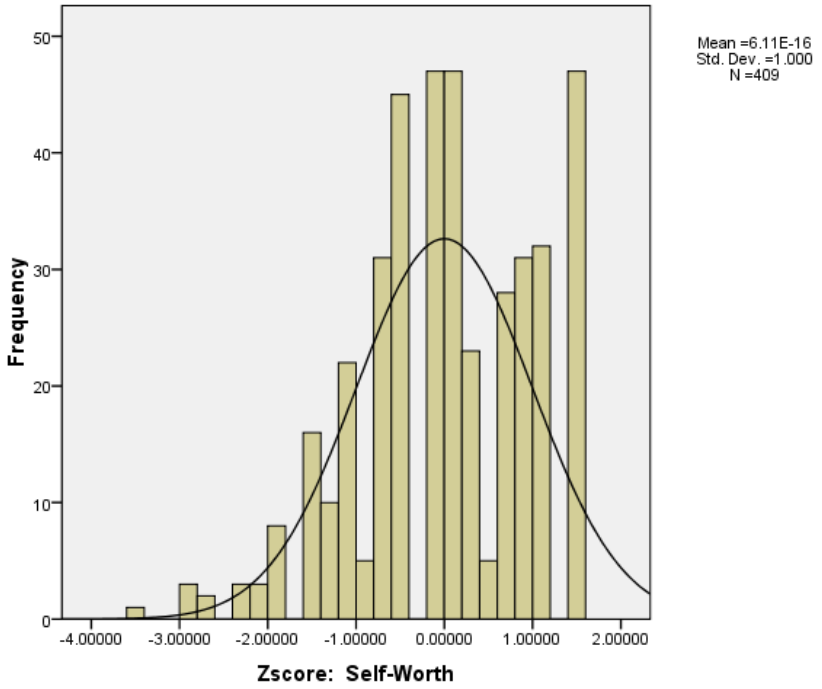




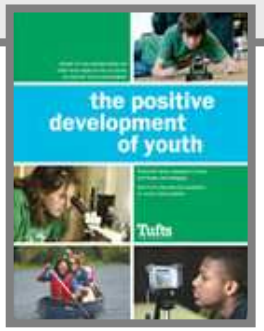
Statistics

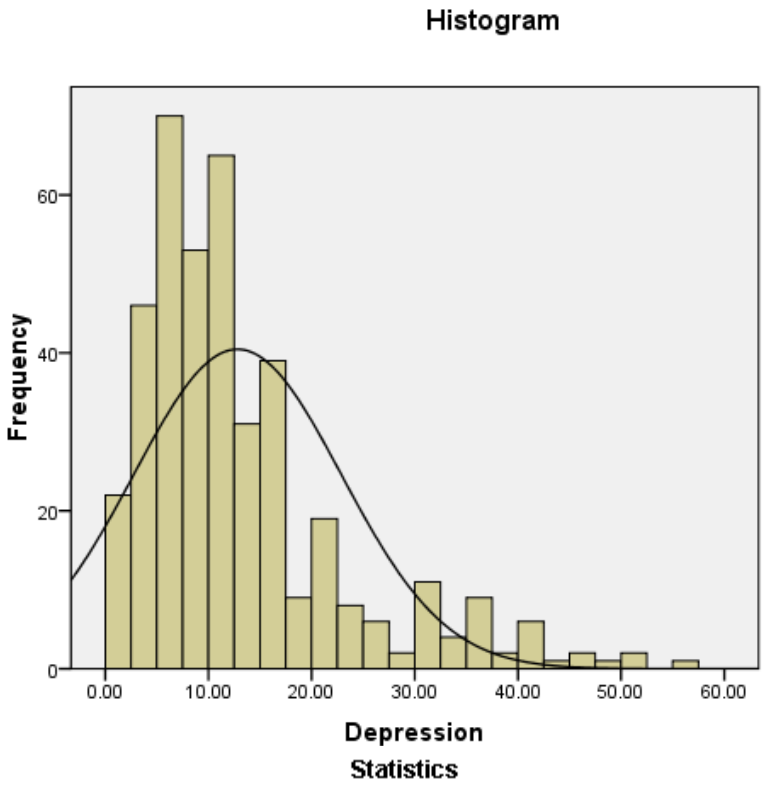
Self-Worth

N	Valid	409.0000
	Missing	.0000
Mean		3.1209
Std. Deviation		.6064
Minimum		1.0000
Maximum		4.0000
Percentiles	25	2.6667
	50	3.1667
	75	3.6667



4-H Study of Positive Youth Development (4H.sav)

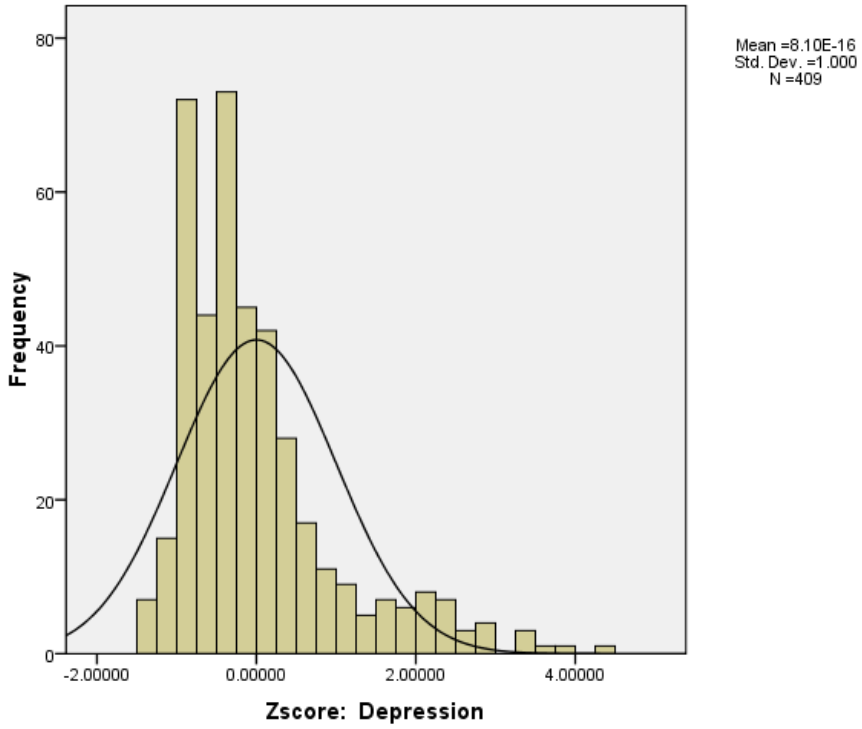




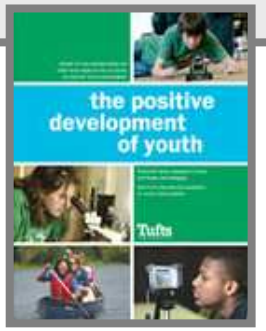
Statistics

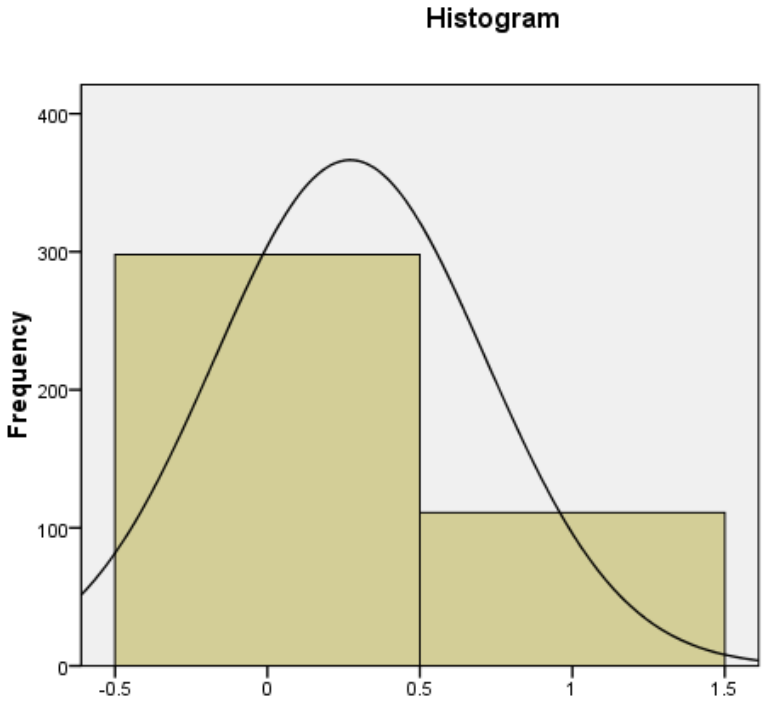
Depression

N	Valid	409.0000
	Missing	.0000
Mean		12.8193
Std. Deviation		10.0814
Minimum		.0000
Maximum		56.0000
Percentiles	25	6.0000
	50	10.0000
	75	16.0000



4-H Study of Positive Youth Development (4H.sav)





Statistics

Depressed = 1, Not Depressed = 0

N	Valid	409.00
	Missing	.00
Mean		.27
Std. Deviation		.45
Minimum		.00
Maximum		1.00
Percentiles	25	.00
	50	.00
	75	1.00