Unit 2: Univariate Statistics (Resistant)

Unit 2 Post Hole:

Use exploratory data analytic techniques to describe the distribution of a variable.

Unit 2 Technical Memo and School Board Memo:

Conduct three univariate exploratory data analyses (with your variables from Memo 1).

Unit 2 (And Unit 3) Reading:

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

Unit 2: 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 2: 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 2: 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 1: 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 (mean, sd) — Unit 2 (quantiles)

```
> 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 (Estimate) — Unit 8 (Std. Error) — Unit 6 (t value, Pr(>|t|)) — Unit 5 (Adjusted R-squared) — Unit 9 (F-statistic)

```
> 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 2: 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

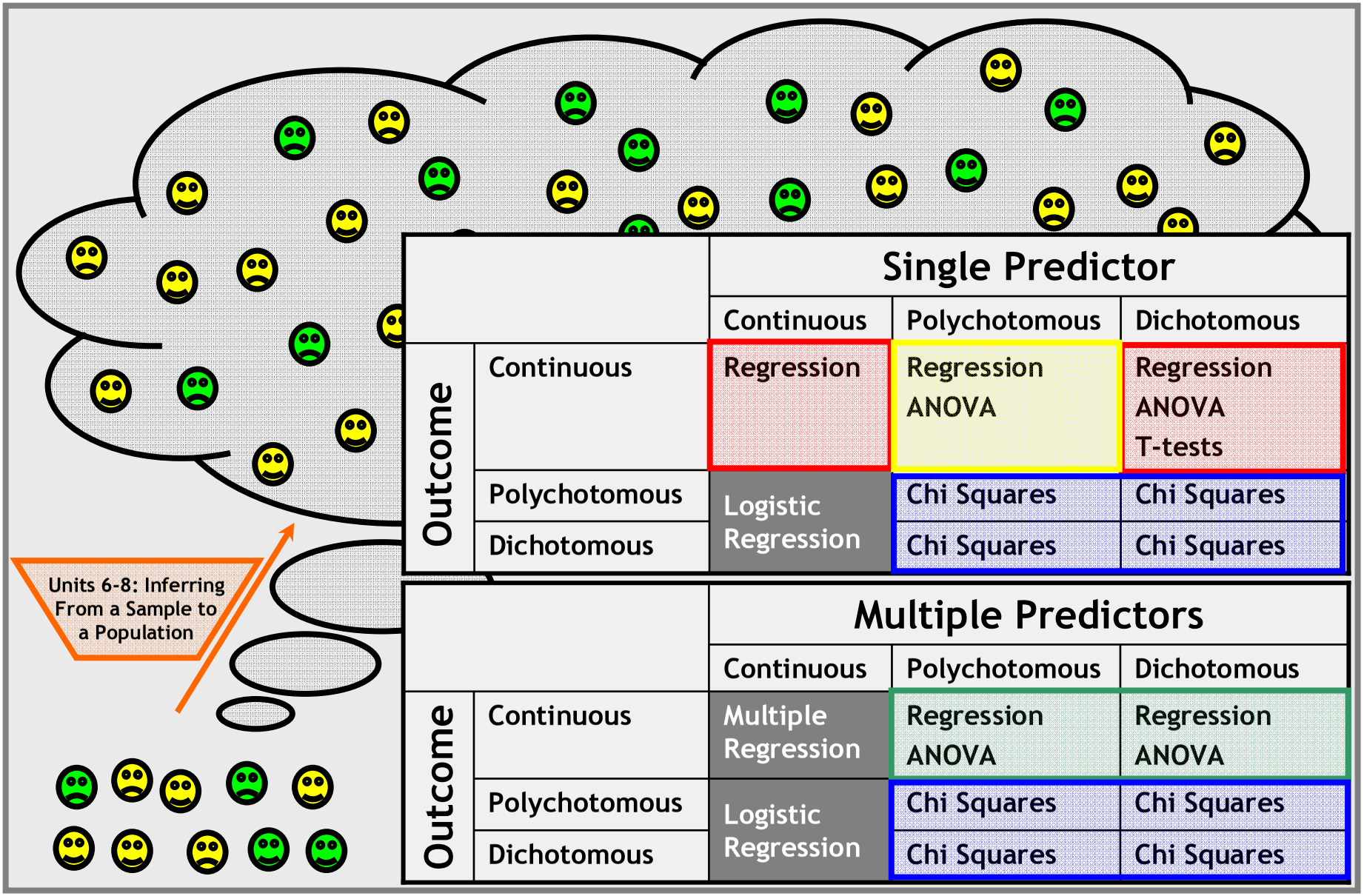
Coefficients[a]

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 2: 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

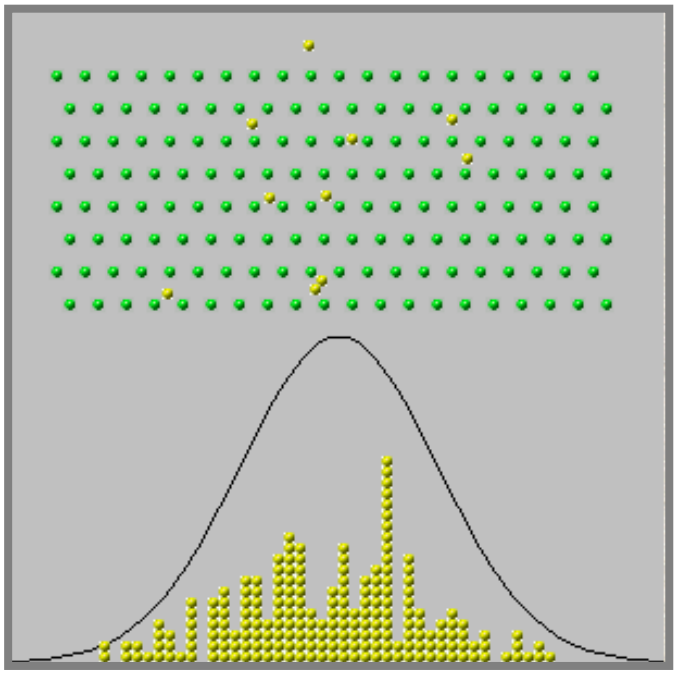
Epistemological Minute

From an epistemological perspective, randomness can be defined in terms of prediction. (There are many definitions of "random.")

Random: Something is truly random if and only if it is unpredictable in principle. Even a super-computing, hyper-reasonable cyborg angel could not predict a future random event from past events.

Pseudo-Random: Something is pseudo-random if and only if it is unpredictable by all relevant predictors. For example, the choices generated from children's choosing songs (e.g., *One Potato, Two Potato*) are pseudo-random (at least until the children are old enough to figure out the pattern).

Is a coin flip random or pseudo-random?



http://www.ms.uky.edu/~mai/java/stat/GaltonMachine.html
http://www.mathsisfun.com/data/quincunx.html

Normal distributions are the predictable product of a common kind of randomness, Gaussian randomness.

Emergence: Wholes are made up of parts. An emergent property is a property of the whole that is not a property of any of its parts.

Puzzle: When we talk about group averages are we necessarily talking about individuals?

A univariate distribution is a collection (a whole) made up of individual observations (parts). If the individual observations are random (or pseudo-random) but the distribution is predictable, then that predictability is emergent.

Unit 2: 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

This is a pain in the butt, but it illustrates a point relevant to Unit 2, which I explicitly note below in the third paragraph:

I recoded *SCHSIZE* into a new variable, *SCHPOP*. I hate that they collected ***ordinal*** data for data that is truly ***interval*** (or to be more precise, ***ratio***). In my recoding, I did a sort of "rounding." For *SCHSIZE* = 1, I saw that it has between 1 and 199 students, so I took the "middle," and I called it SCHPOP = 100 students. For *SCHSIZE* = 2, I saw that it has between 200 and 399 students, so I took the "middle," and I called it SCHPOP = 300 students. And, so on. For *SCHSIZE* = 7, I saw that it has 1200+ students, so I held my nose (or rather held my nose harder), and I called it SCHPOP = 1300 students. Do you see how I am really guessing? I am guessing that a SCSSIZE = 1 is a SCHPOP = 300, when it could be SCHPOP = 201 or SCHPOP = 398. Because I have no further information within the range, I'm guessing the "middle." Now, I wish the data collectors just asked for the exact school size/population, but they didn't, so I'm stuck with this data. I call this data "binny" because the researchers metaphorically sorted schools into 7 bins. I hate binny data because we lose information (unless the data is naturally binny!)

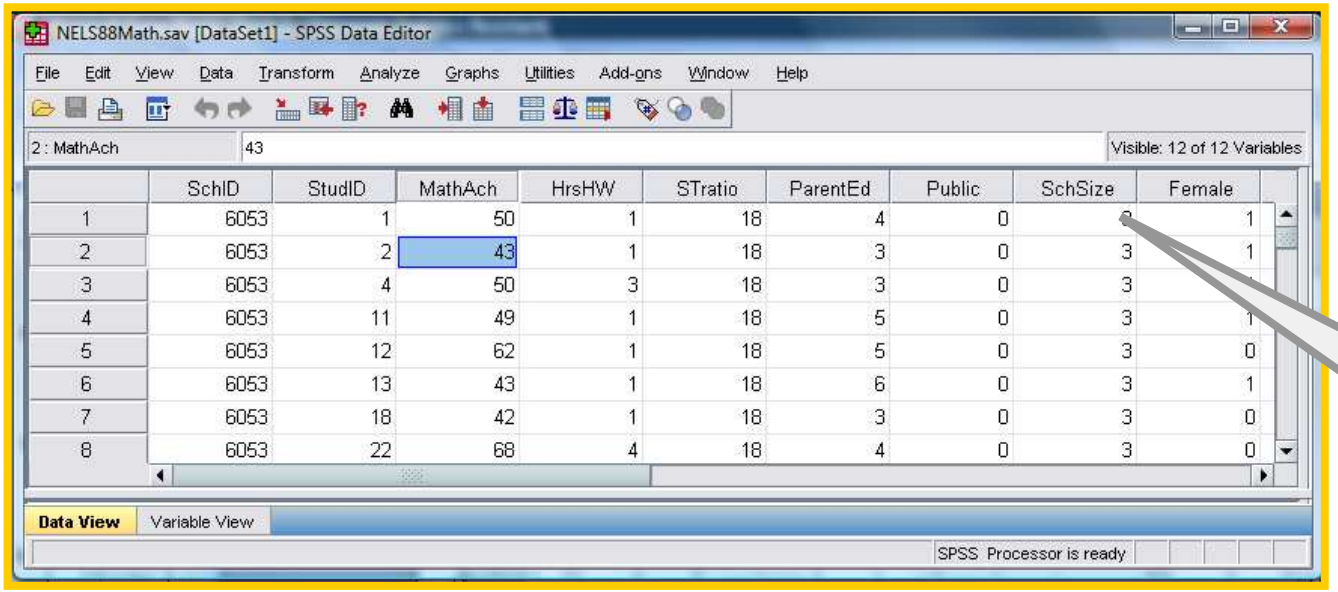
Why did I do this recoding? I want to talk about the relationship between math achievement and number of students. I don't want the number of students to be in arbitrary terms of 1 through 7. I want it to be in natural terms of, well, number of students. As data analysts, we want our numbers to be meaningful!

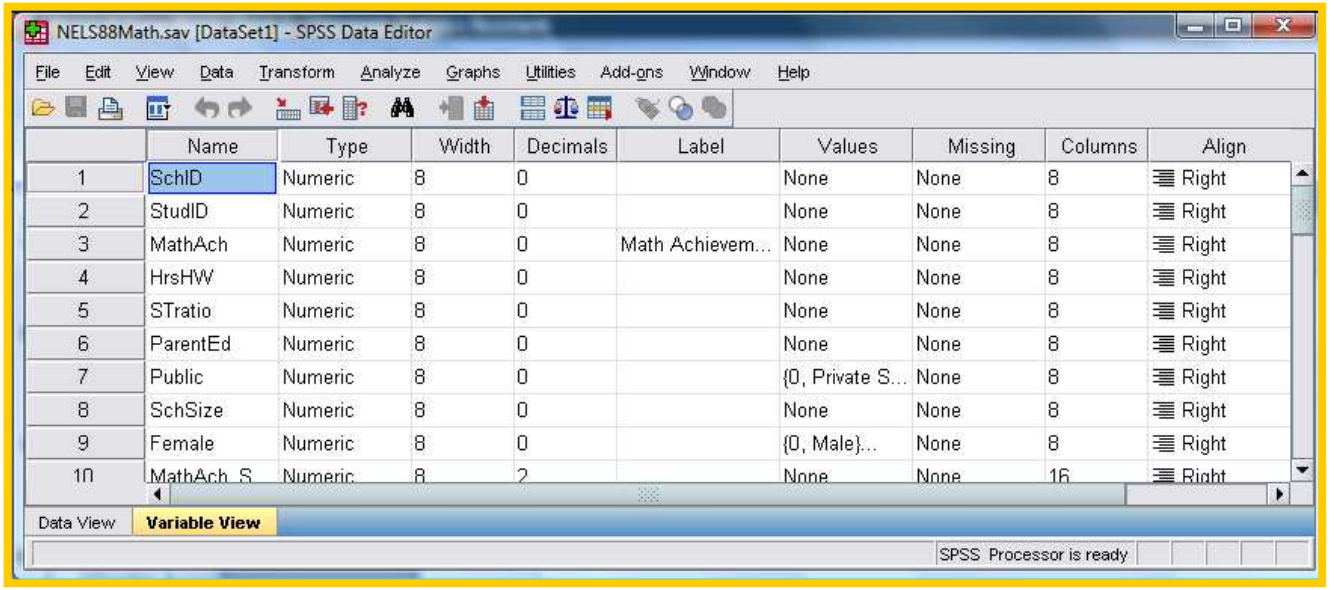
Point relevant to Unit 2: In Unit 2, and throughout this course, we are going to be taking the "middle" as our best guess in the absence of further information. I hope it's intuitive to you that the "middle" is reasonable as a best guess, noting that "best guess" does not necessarily mean "good guess." In this course, your best is good enough, but in data analysis our best is often not good enough! In Unit 2, we are going to look at one way to define the "middle," and in Unit 3 we will look at another.

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





Note that I took this snapshot before I created my new variable, SCHPOP, based on SCHSIZE, but in terms of (guesstimated) number of students.

Percent Terminology Etc. (Not Necessarily To Be Memorized)

Percentage: If I scored 80% the first time I took a test, and then I scored 100% on the retake, my score went up 25% (because 80% + (80%*25%) = 100%).

Do not confuse "percentage differences" with "percentage point differences."

Percentage Point: If I scored 80% the first time I took a test, and then I scored 100% on the retake, my score went up 20 percentage points.

Go ahead, confuse "percentiles" with "percentile ranks." Everybody else does!

Percentile: An n^{th} percentile is the value of your variable at which n percent of your values fall below.* Say the 25^{th} percentile of *MathAch* is 300, then 25% of the math achievement scores fall below 300.

Percentile Rank: In the example, a math achievement score of 300 has a percentile rank of 25. A percentile rank is a number from 0 to 99. A percentile is a number on the scale of the variable.

Median: The 50^{th} percentile. The middle observation when you line up the observations from lo to hi.

Upper Quartile: A marker of the top 25%, the 75^{th} percentile.

Lower Quartile: A marker of the bottom 25%, the 25^{th} percentile.

Tukey's Hinges: The 25^{th} and 75^{th} percentiles (basically).

Interquartile Range: The "distance" from the 25^{th} to the 75^{th} percentile (aka "Midspread").

Midspread: The "distance" from the 25^{th} to the 75^{th} percentile (aka "Interquartile Range").

} Synonyms

Reasonable Upper/Lower Bound For Outlier Detection (RUB/RLB): The RUB is the 75^{th} percentile plus 1.5 midspreads. The RLB is the 25^{th} percentile minus 1.5 midspreads. This is a "rule of thumb."

Rule of Thumb #1: Rules of thumb only work when they work. Use your own judgment.

*Sometimes percentiles are defined not as "below" but as "at or below." In the spirit of "soft eyes" that EDA requires, I generally don't care about the difference, and my indifference shows up in the remaining definitions.

Intro to Stem and Leaf Plots (Leading to Histograms!) Part I of IV

*NELS88Math.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1 : MathAch | 60 | Visible: 12 of 12 Variables

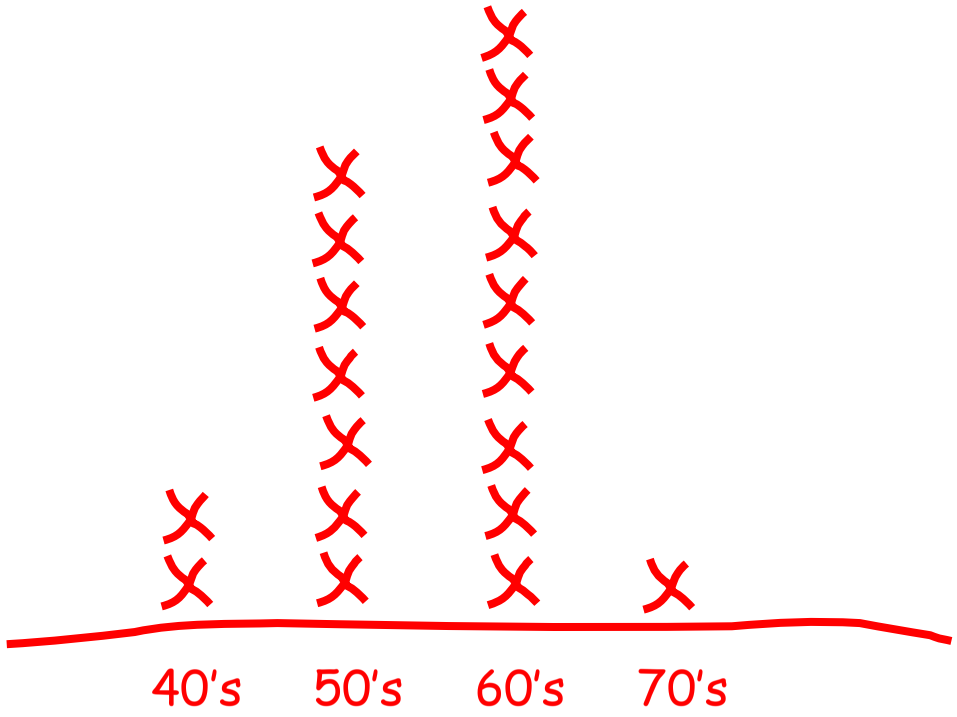
	SchID	StudID	MathAch
1	6053	39	60
2	6053	42	69
3	6053	43	44
4	6053	44	61
5	6053	45	65
6	6053	53	67
7	6053	55	57
8	6053	56	64
9	6053	57	67
10	6053	60	65
11	6053	63	70
12	6053	64	59
13	6053	67	54
14	6053	68	53
15	6053	69	58
16	6053	71	59
17	6053	76	60
18	6053	77	48
19	6053	82	59

Data View | Variable View

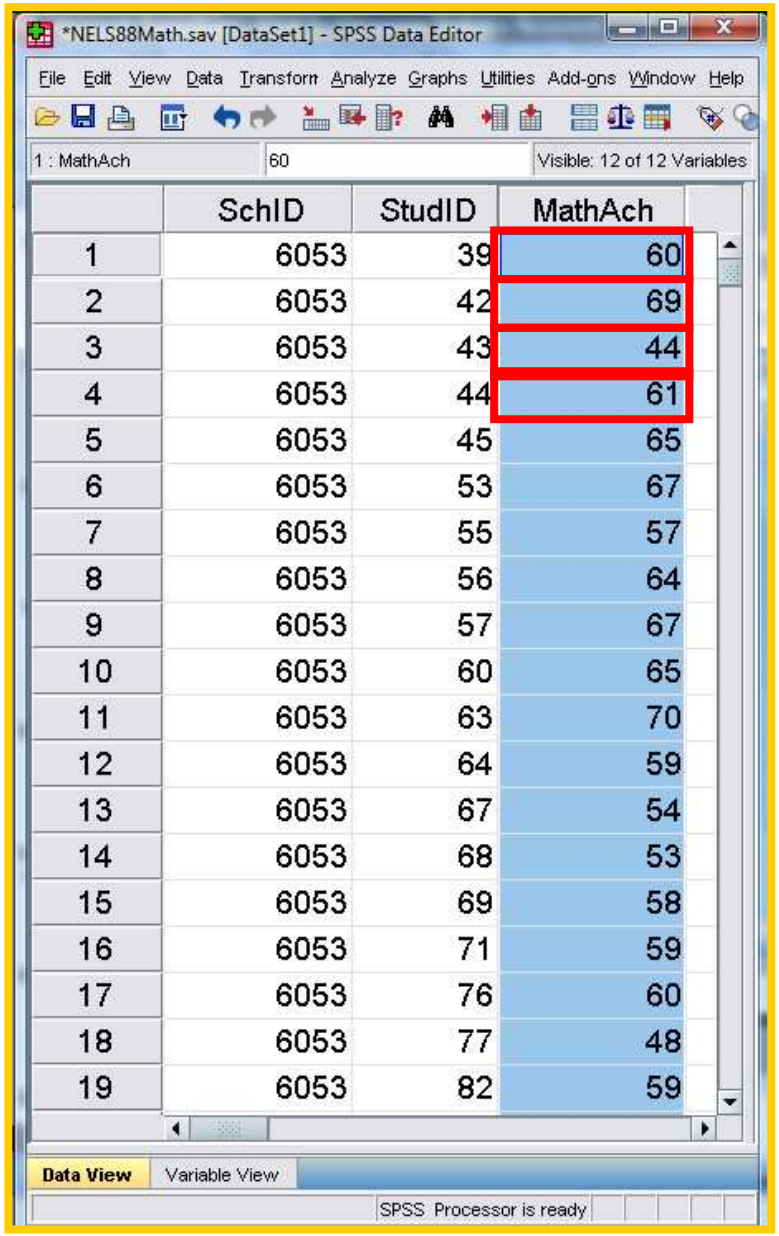
SPSS Processor is ready

With pencil and paper, we can quickly get a handle on small data sets by using stem-and-leaf plots.

What does math achievement "look" like?
Is a score of 60 high or low?



Intro to Stem and Leaf Plots (Leading to Histograms!) Part II of IV

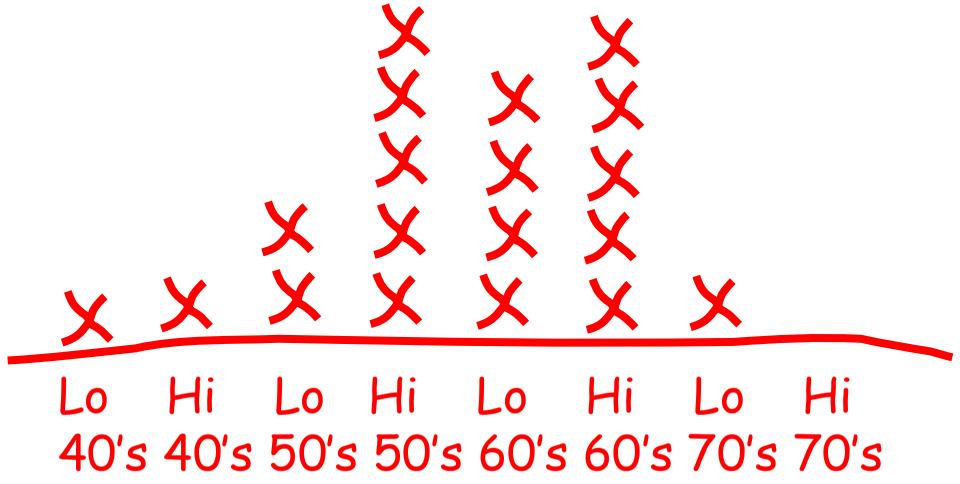


	SchID	StudID	MathAch
1	6053	39	60
2	6053	42	69
3	6053	43	44
4	6053	44	61
5	6053	45	65
6	6053	53	67
7	6053	55	57
8	6053	56	64
9	6053	57	67
10	6053	60	65
11	6053	63	70
12	6053	64	59
13	6053	67	54
14	6053	68	53
15	6053	69	58
16	6053	71	59
17	6053	76	60
18	6053	77	48
19	6053	82	59

With pencil and paper, we can quickly get a handle on small data sets by using stem-and-leaf plots.

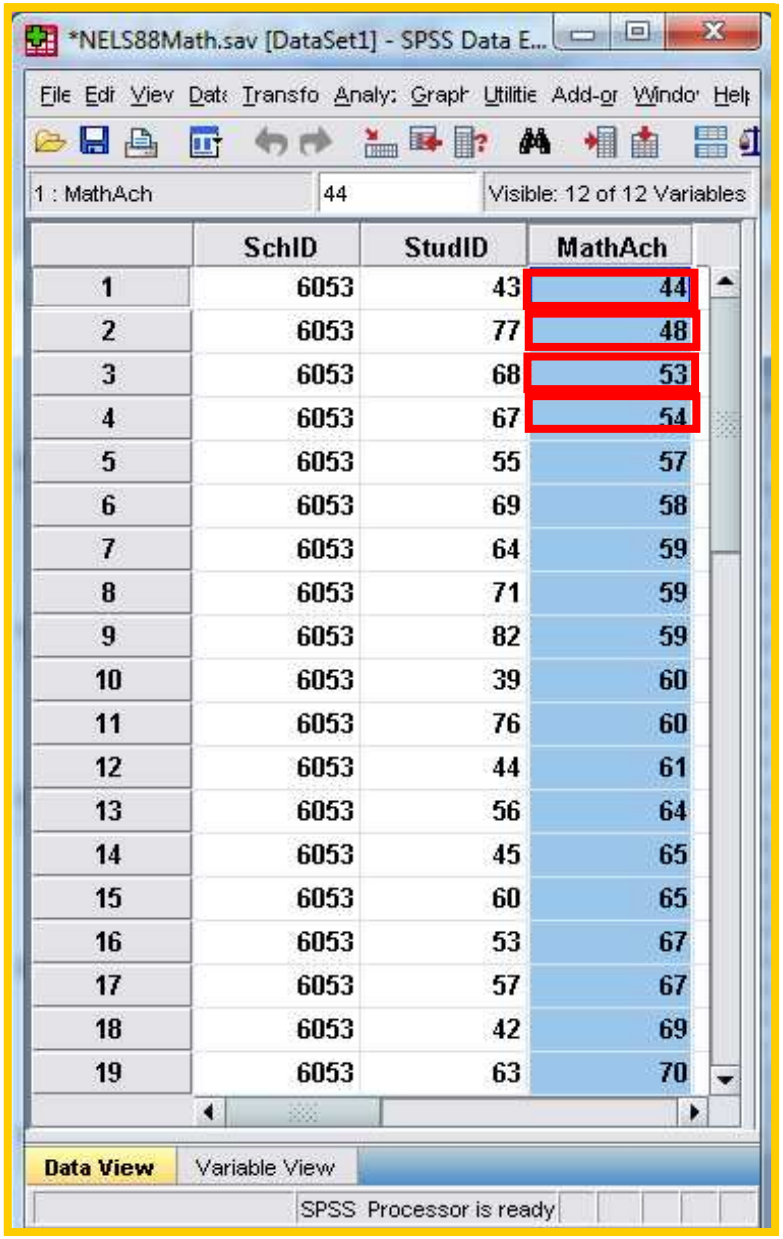
What does math achievement "look" like?
Is a score of 60 high or low?

We can get more fine-grained information if we give ourselves more "bins." Metaphorically, we are sorting observations into bins.



Now, we can see that a score 60 is neither high nor low, but middling.

Intro to Stem and Leaf Plots (Leading to Histograms!) Part III of IV



	SchID	StudID	MathAch
1	6053	43	44
2	6053	77	48
3	6053	68	53
4	6053	67	54
5	6053	55	57
6	6053	69	58
7	6053	64	59
8	6053	71	59
9	6053	82	59
10	6053	39	60
11	6053	76	60
12	6053	44	61
13	6053	56	64
14	6053	45	65
15	6053	60	65
16	6053	53	67
17	6053	57	67
18	6053	42	69
19	6053	63	70

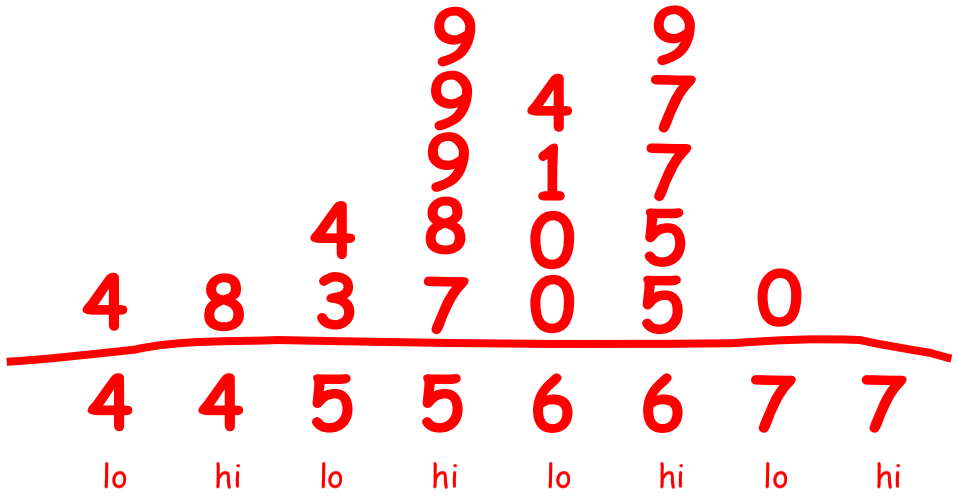
Another way to get more fine-grained information is to include the actual values in our plot.

First, sort the data from low to high. This in itself is helpful for "seeing" math achievement! See that a score of 60 is in the middle.

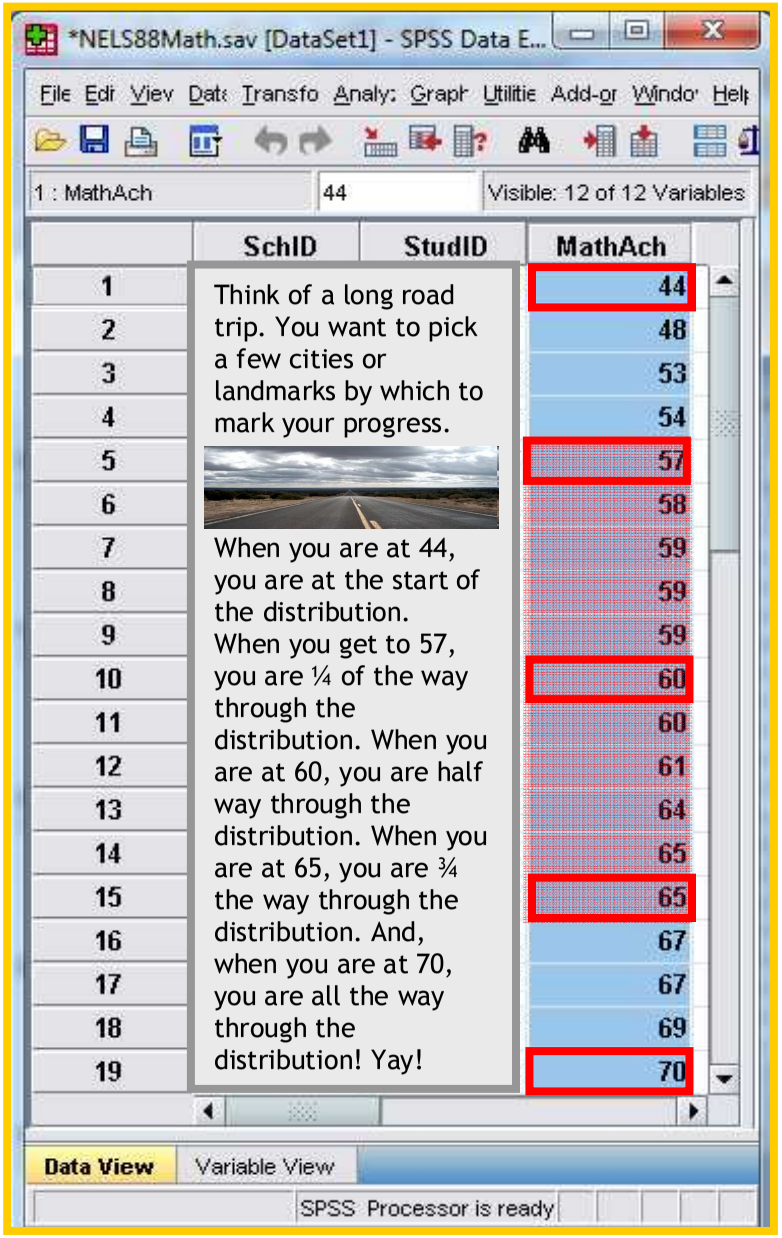
For 44, We put a 4 in the 4(lo) column to make 44.

For 48, We put an 8 in the 4(hi) column to make 48.

In general, the column tells us the tens digit and the stacked number tells us the ones digit.

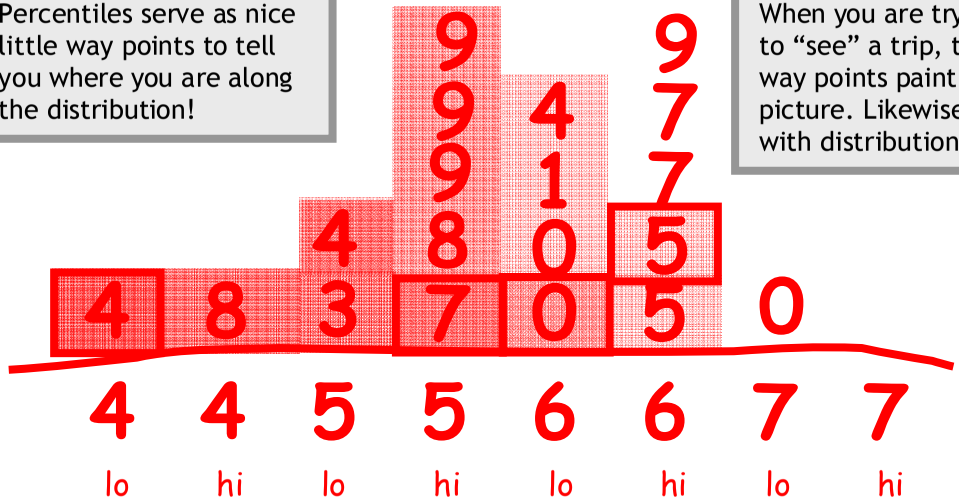


Intro to Stem and Leaf Plots (Leading to Histograms!) Part IV of IV



What value is greater than exactly/roughly 0% of the values?
0th Percentile/ Minimum: 44
What value is greater than exactly/roughly 25% of the values?
25th Percentile/ 1st Quartile/ Tukey's Lower Hinge: 57
What value is greater than exactly/roughly 50% of the values?
50th Percentile/ Median: 60
What value is greater than exactly/roughly 75% of the values?
75th Percentile/ 3rd Quartile/ Tukey's Upper Hinge: 65
What value is greater than exactly/roughly 99% of the values?
99th Percentile/ Maximum: 70
Now, we can use the above numbers to give us more info.
75th-25th Percentile/ Interquartile Range/ Midspread: 8
We subtract the 25th percentile (here, 57) from the 75th percentile (here, 65) to get the midspread, which tells us that the bulk of the values fall within 8 points. The midspread is nice to know in and of itself, but we'll also use it to detect outliers.

Percentiles serve as nice little way points to tell you where you are along the distribution!



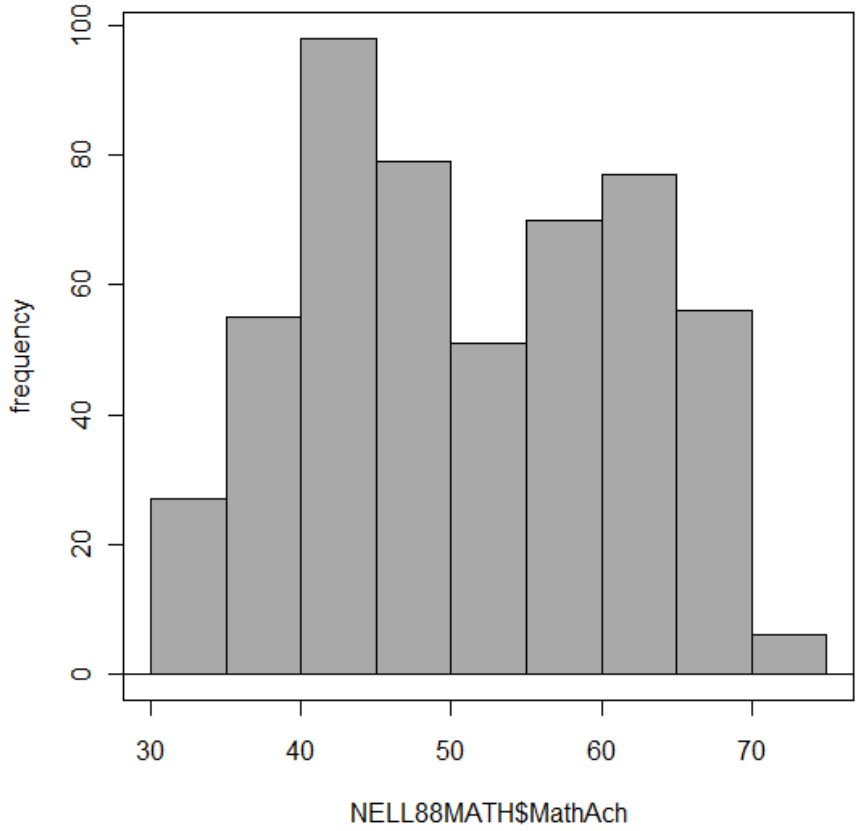
When you are trying to "see" a trip, the way points paint a picture. Likewise with distributions.

Our First Histogram!

When you have 19 observations a stem-and-leaf plot is nice to do by hand, but when you have, say, 519 observations, it's best to let the computer do the work! Although most statistical software packages can do stem-and-leaf plots, histograms are the industry standard for painting the picture of a univariate distribution (i.e., distribution of one variable).

```
3 | 2233333
3 | 444444555555555
3 | 66666666667777777777777
3 | 888888899999999999
4 | 00000000000001111111111111111111111111
4 | 22222222222222222222233333333333333333333333
4 | 44444444444444444444555555555
4 | 66666666666677777777777777777
4 | 88888888888888888889999999999999999
5 | 000000000000001111111
5 | 22222233333333333333333
5 | 444444444444444455555
5 | 66666666666666666777777777
5 | 88888888888888888899999
6 | 0000000000000000000111111111
6 | 22222222222222222223333333333333333333
6 | 444444444444445555555555555555
6 | 667777777777777777777
6 | 888888888899999999999999
7 | 000000000111111
```

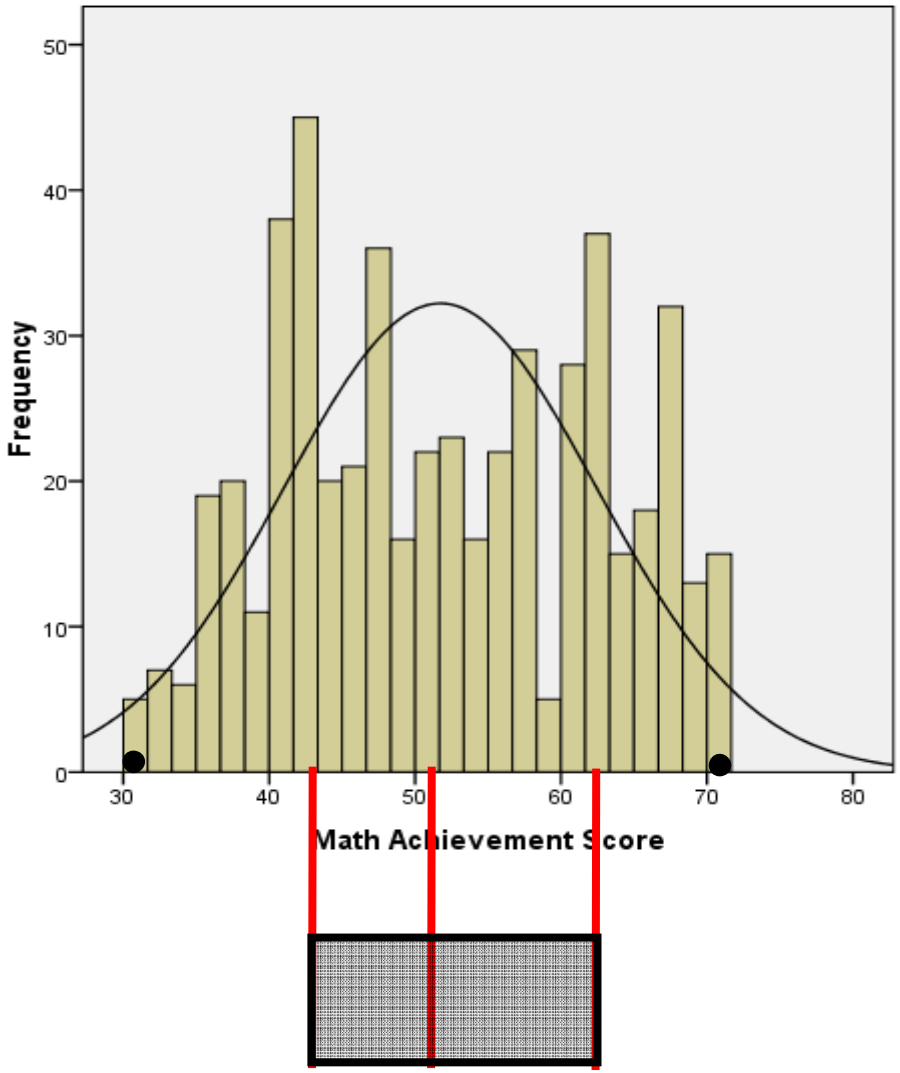
Stem-and-Leaf Plot in R



Histogram in R

Exploring Math Achievement: Location and Spread

Figure 2.1. Histogram and univariate statistics for math achievement scores (n = 519).



Let's call the median (i.e., the 50th percentile) "the **location**" of the distribution.

In our sample, the median math score is 51.

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

Let's call the interquartile range (i.e., the 75th-25th percentile, or midspread) "the **spread**" of the distribution.

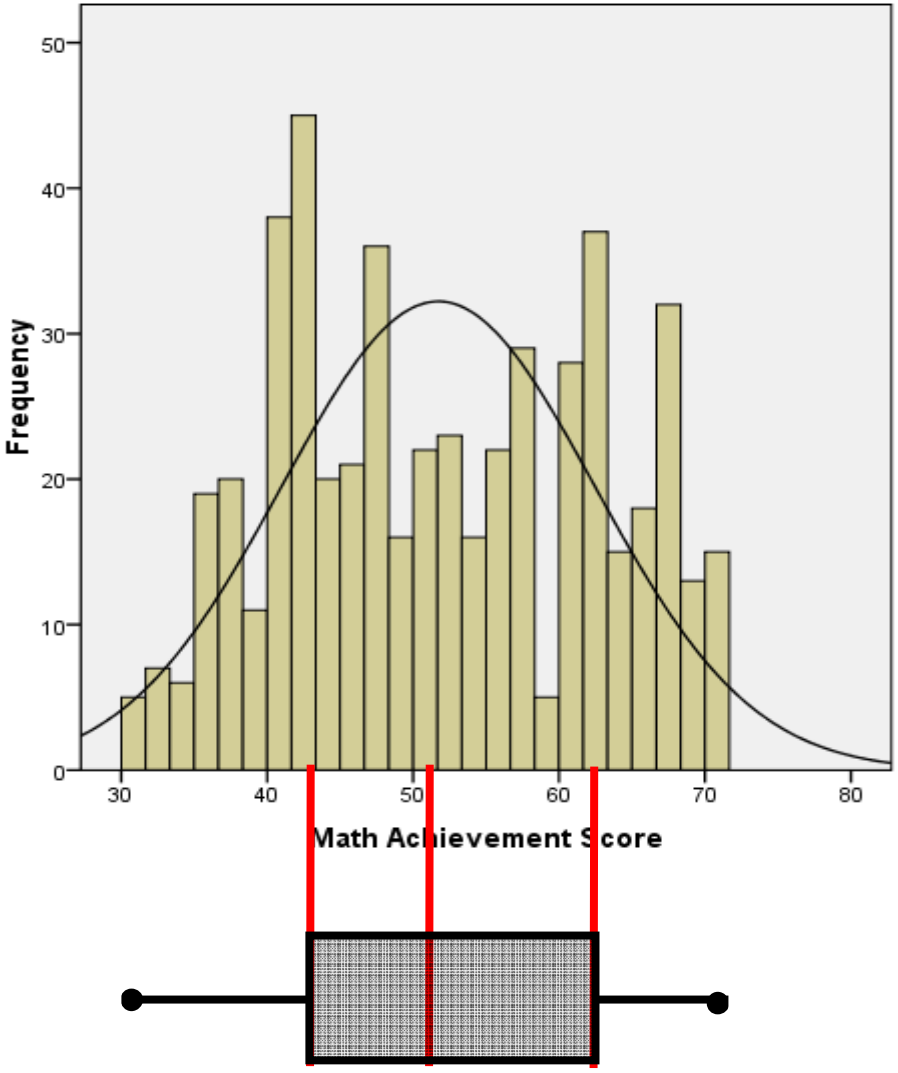
Most of the math scores are between 43 and 62, so the midspread is 19.

While we're considering **spread**, let's look at the min and max values.

Math scores range from 30 to 71 points.

A Method for Detecting Univariate Outliers (The RLB and RUB)

Figure 2.1. Histogram and univariate statistics for math achievement scores (n = 519).



The RUB is 90.5. No upper outliers! Do you understand the process?

What's happening here is something typical of data analysis. We want to detect outliers (i.e., observations that are "far away"), but how far is far? The answer will always be relative to the distribution, so we dig into the distribution to create an internal measuring stick, in this case, the midspread. We use the midspread to measure distance. It's like measuring your height in terms of feet, the length of your own feet!

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

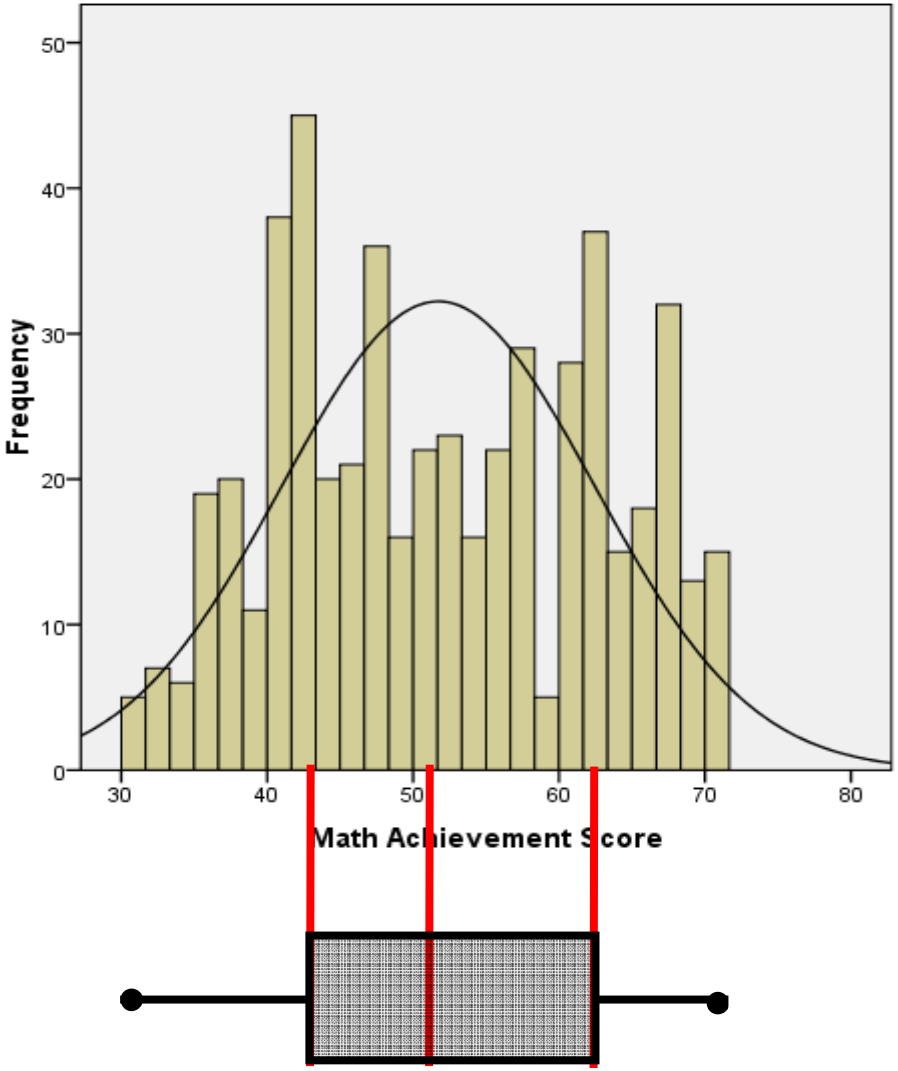
While we're still considering **spread**, let's look at reasonable upper and lower boundaries for outliers. We'll use 1.5 times the midspread range. We'll subtract 1.5 times the midspread from the 25th percentile, and we'll add 1.5 times the midspread to the 75th percentile. We will deem to be outliers any values that fall outside the reasonable lower boundary (RLB) and the reasonable upper boundary (RUB).

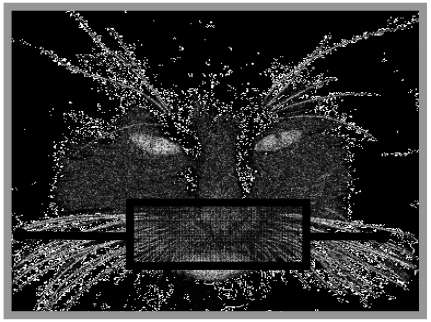
The midspread is 19. We subtract the 1.5 times the midspread (1.5*19 = 28.5) from the 25th percentile (43) to get 14.5 as an RLB.

See the whisker? This is called a box-and-whisker plot! Since the RLB is lower than the minimum, we trim the whisker, and conclude that there are no lower outliers.

Exploring Math Achievement: Shape

Figure 2.1. Histogram and univariate statistics for math achievement scores (n = 519).





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

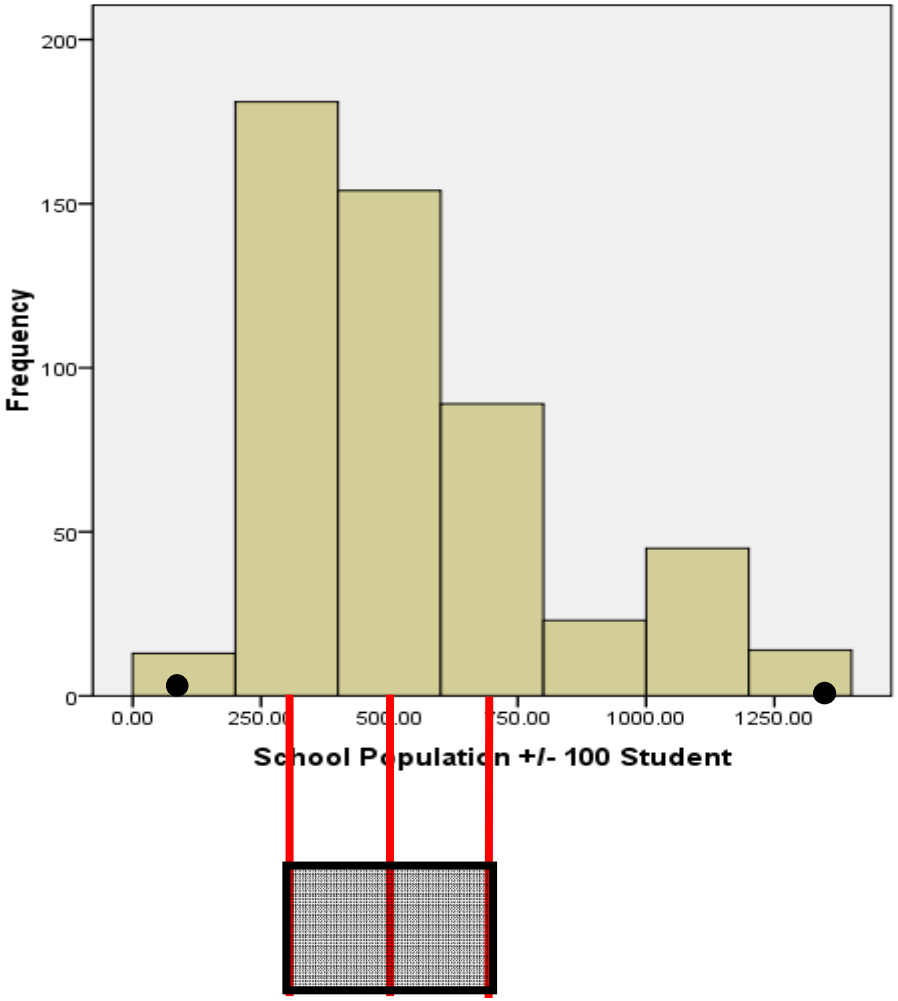
Now, let's consider **shape**. We will use a best-fitting normal curve as a baseline for purposes of comparison; that mysterious mountain in the middle is a normal curve overlay. The normal curve is just a (theoretical) basis of comparison. We can ask, does our distribution look like a normal distribution? If not, how does our distribution differ?

A normal distribution is a distribution with one peak; it is unimodal. Our distribution has two peaks; it is bimodal!

The distribution of math scores is bimodal, with one peak around 42 and another slightly smaller peak around 62.

Exploring School Size: Location and Spread

Figure 2.2. Histogram and univariate statistics for students' school sizes (n = 519).



Let's call the median (i.e., the 50th percentile) "the **location**" of the distribution.

In our sample, the median student goes to a school with a population of about 500.

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

Let's call the interquartile range (i.e., the 75th-25th percentile, or midspread) "the **spread**" of the distribution.

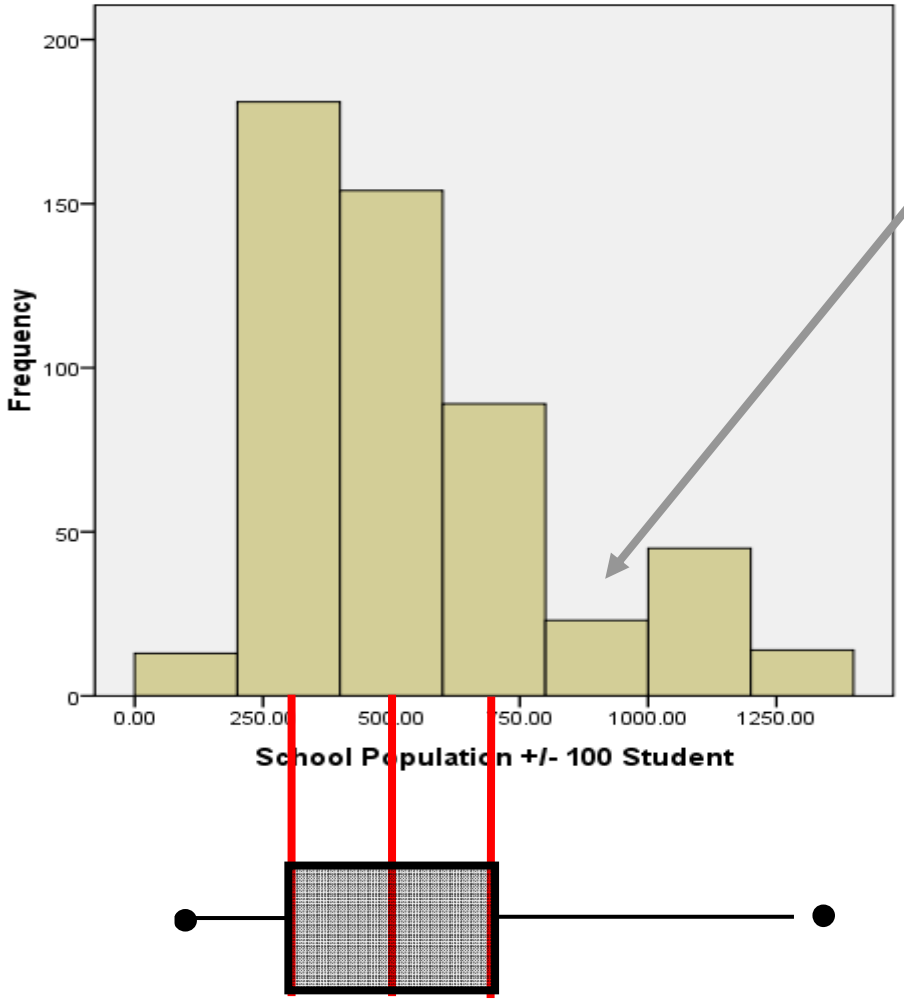
Most of the school populations are between 400 and 700, so the midspread is 300.

For **spread**, let's also look at the min and max.

Some students go to small schools of about 100 students, and other students go to large schools of about 1300 students .

Exploring Math School Size: Outliers (Spread Continued)

Figure 2.2. Histogram and univariate statistics for students' school sizes (n = 519).



Not every nook and cranny is meaningful. Not every zig and zag is noteworthy. In fact, few are. Develop soft eyes, eyes that can look beyond the idiosyncrasy to the underlying pattern.

Notice the dip. Who cares? I don't. If you were to give the histogram a shake, the dip would disappear as the towers toppled a little to the left and right. Often the difference between being in one bin or another (to the left or right) is measurement error (e.g., a sneeze at the right or wrong time while filling out the survey or taking the test). Forget the dip.

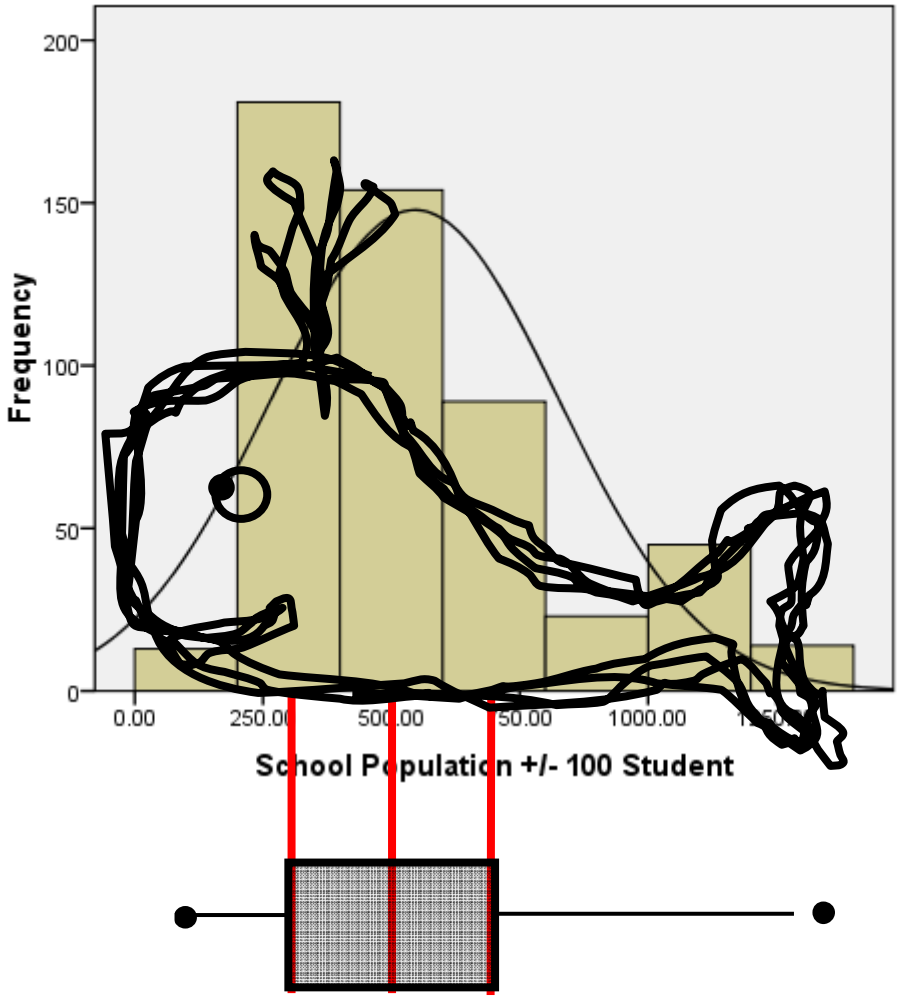
Minimum		100.0000
Maximum		1300.0000
Percentiles	25	300.0000
	50	500.0000
	75	700.0000

For **spread**, let's also look at reasonable upper and lower boundaries for outliers. We'll use 1.5 times the midspread.

No observations fall below our RLB of -150, of course, but students who go to the largest schools in our sample (approximately 1300) fall above our RUB of 1150, so we conclude that there are outliers in the upper range of our distribution.

Exploring School Size: Shape

Figure 2.2. Histogram and univariate statistics for students' school sizes (n = 519).



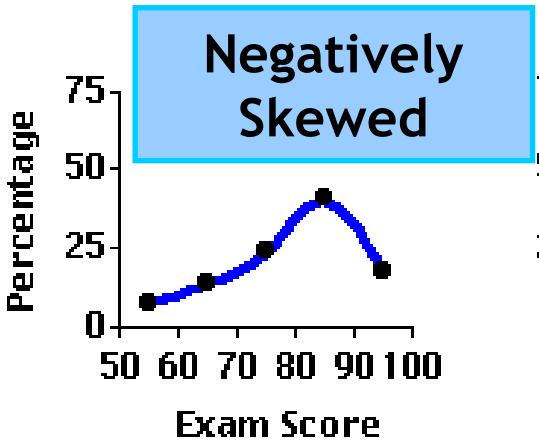
Now, let's consider **shape**. We will use a best-fitting normal curve as a baseline for purposes of comparison. When we consider shape, we not only consider the number of peaks (unimodality vs. bimodality), we also consider the peakiness of the peaks and the length of the tails.

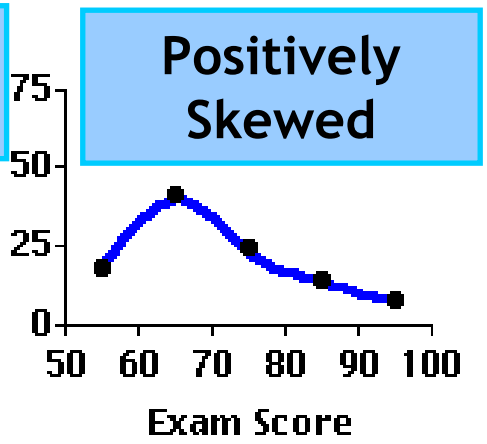
The fancy word for "peakiness" is kurtosis, but please use the non-fancy terms. A leptokurtic distribution is peaky. A platykurtic distribution is flat. A mesokurtic distribution is neither peaky nor flat, and the paradigmatic mesokurtic distribution is the normal. (The normal distribution is the baby bear of distributions, which is why we overlay it for purposed of comparison. It's juuuuuuuust right.)

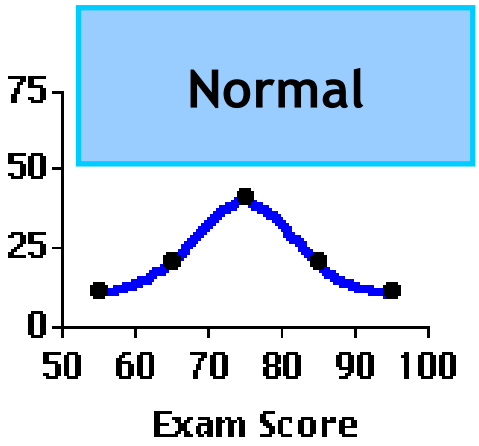
If kurtosis is about the middle of the distribution, then the tails are about the edges. Are the edges symmetrical? Or, does one edge trail off, forming a long tail? Note, contrary to some usages, skew is in the tail of the whale. If there is long upper tail, the skew is positive. If there is a long lower tail, the skew is negative. Again, the normal distribution is just right, it has no skew; it is, by definition, perfectly symmetric.

The distribution of student's school population sizes is positively skewed. (Note that it's also somewhat peaky (i.e., leptokurtic), but heavy skewness generally dominates the shape, making peakiness ignorable.)

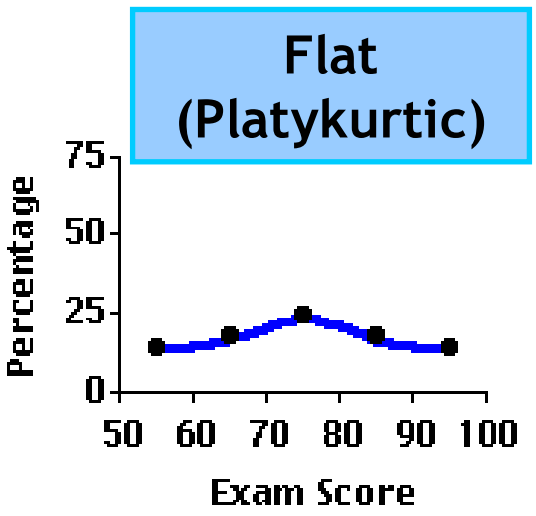
Shapes of Distributions

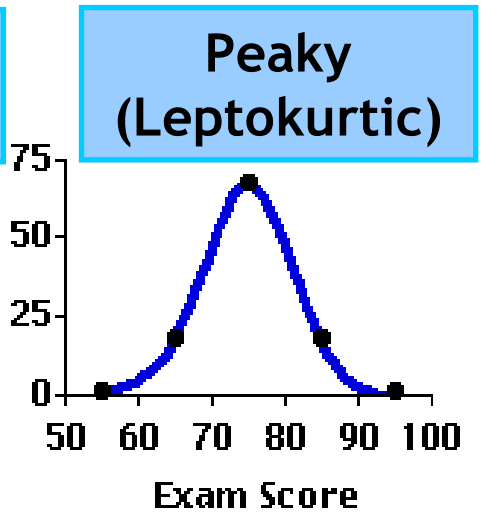


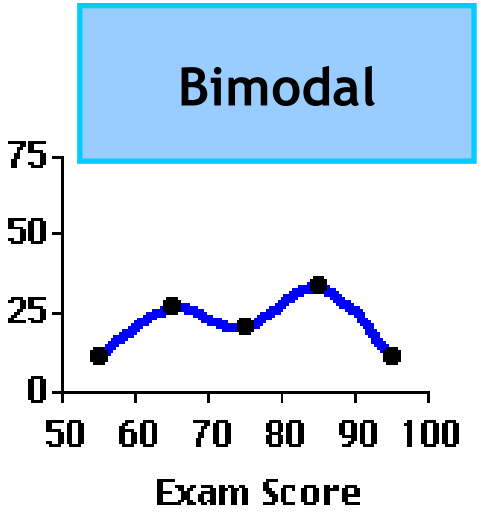




Normal distributions (by definition) are symmetric (i.e., zero skewed) and neither flat nor peaky (i.e., zero kurtotic, or mesokurtic).







http://faculty.vassar.edu/lowry/ch2pt1.html

Univariate Exploratory Data Analysis

SPread

Location

And

SHape

When conducting exploratory data analysis for single variables, generate a histogram or stem-and-leaf plot, and look for SPLASH: SPread, Location And SHape.

You probably want to assess shape first. For starters, draw a best fitting normal curve with the computer, or use your imagination if you don't have electronic resources. Use the normal curve as your basis of comparison. Look for modality: is the distribution one-peaked (unimodal), two-peaked (bimodal) or multi-peaked (multimodal)? Look for skew: is there a long tail to the left (negative skew), or to the right (positive skew), or neither (symmetry)? Look for kurtosis: Is the distribution flat (platykurtic) or peaky (leptokurtic) or "normalish" (mesokurtic)?

Then, assess the location by determining the median (or 50th percentile). The median marks the point where 50% of the observations are greater and 50% are less.

Finally, assess the spread. What is the interquartile range (i.e., midspread)? What are the minimum and maximum values? Are there any outliers? Note your standard for detecting outliers.

You can now do the Unit 2 Post Hole: Use exploratory data analytic techniques to describe the distribution of a variable.

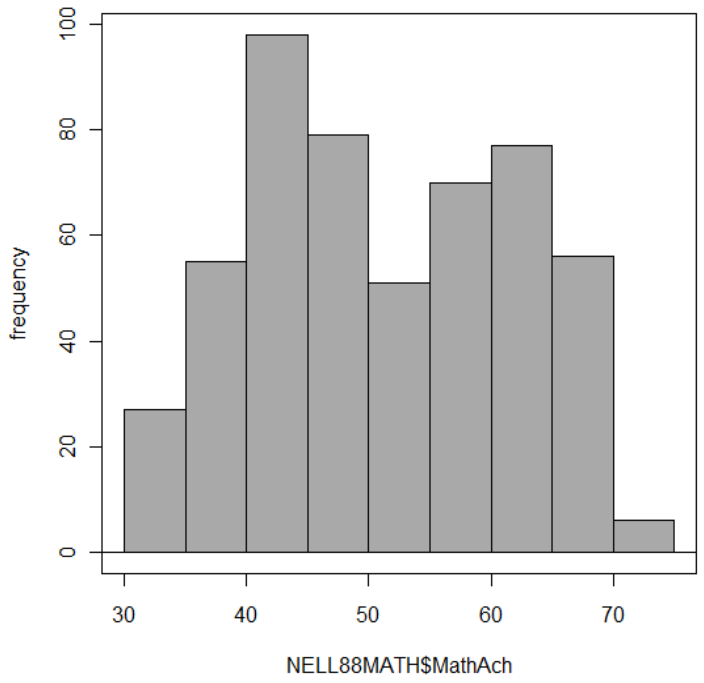
Practice problems are at the end of the presentation.

Dig the Post Hole

Unit 2 Post Hole:

Use exploratory data analytic techniques to describe the distribution of a variable.

Evidentiary materials: a histogram and percentiles.



mean	sd	0%	25%	50%	75%	100%	n
51.72254	10.70922	30	43	51	61.5	71	519

Spread: Use the midspread, min, max, RLB and RUB. Mention outliers or the lack thereof.

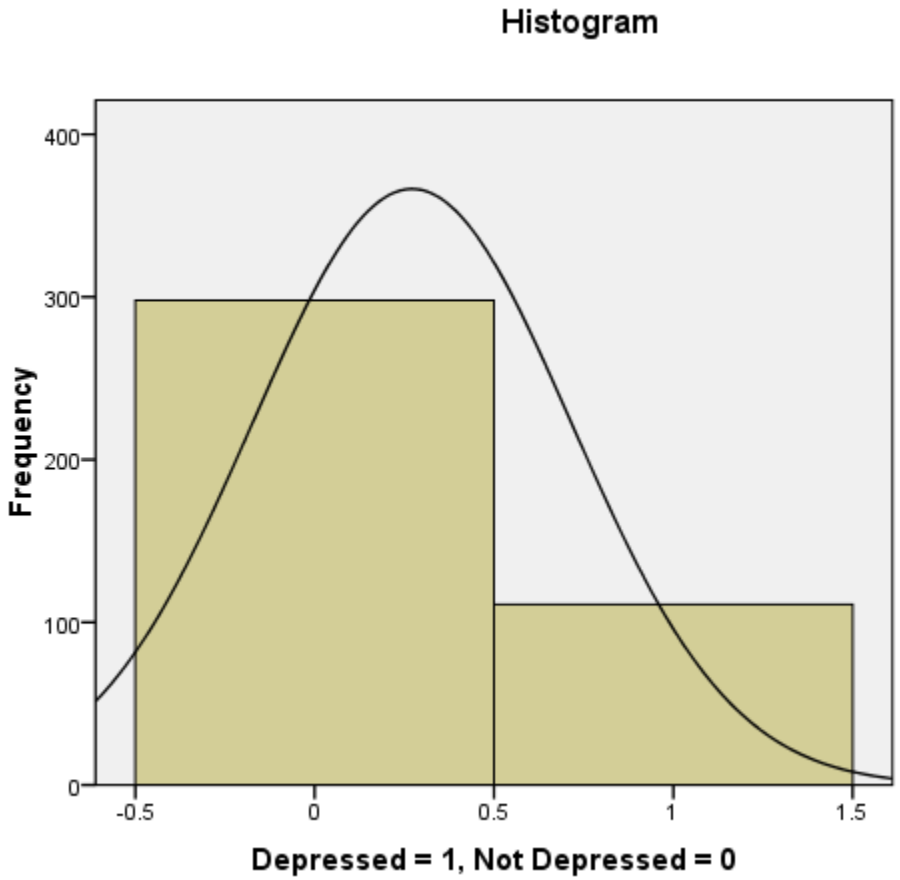
Location: Use the median (aka 50th percentile).

Shape: If the distribution is bimodal (or multimodal), this fact dominates, and skew and kurtosis are probably not worth mentioning. Else if the distribution is skewed, this fact dominates, and kurtosis is probably not worth mentioning. When the distribution is unimodal and symmetric, be sure to check kurtosis (by comparing to a normal curve).

Here is my shot (the parenthetical comments are optional but nice):

Spread: Scores range from 30 to 71. The midspread is 18.5. The RLB is 15 (43-1.5*18.5) suggesting no lower outliers. The RUB is 90 (61.5+1.5*18.5) suggesting no upper outliers.
Location: The median is 51.
Shape: The distribution is bimodal.

Dichotomies (and Polychotomies) are Easy!



Dichotomous variables can never be normally distributed. The median will always equal the min or the max. There can't be outliers. Don't even try to give dichotomies the full SPLaSH treatment. Instead, just report the frequency (or perhaps percentage) for each category.

Statistics

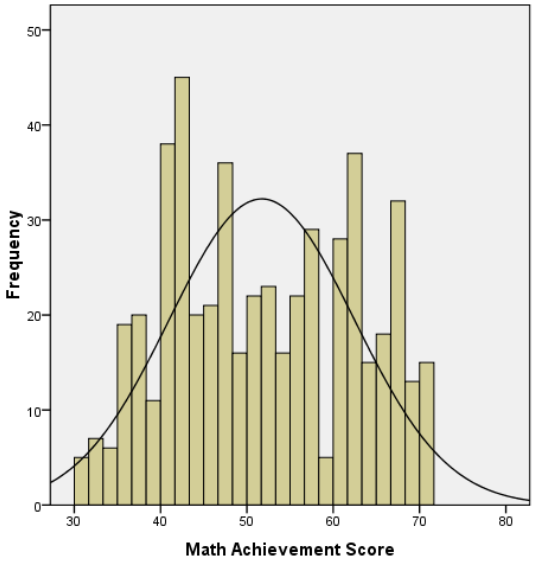
Depressed = 1, Not Depressed = 0

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

In our sample, 27% of the subjects were depressed, and 73% were not.

From where did I get the 27%? Look at the mean: .27. The mean is the proportion of 1s in a distribution with only 0s and 1s. More on the meaning of "mean" in Unit 3.

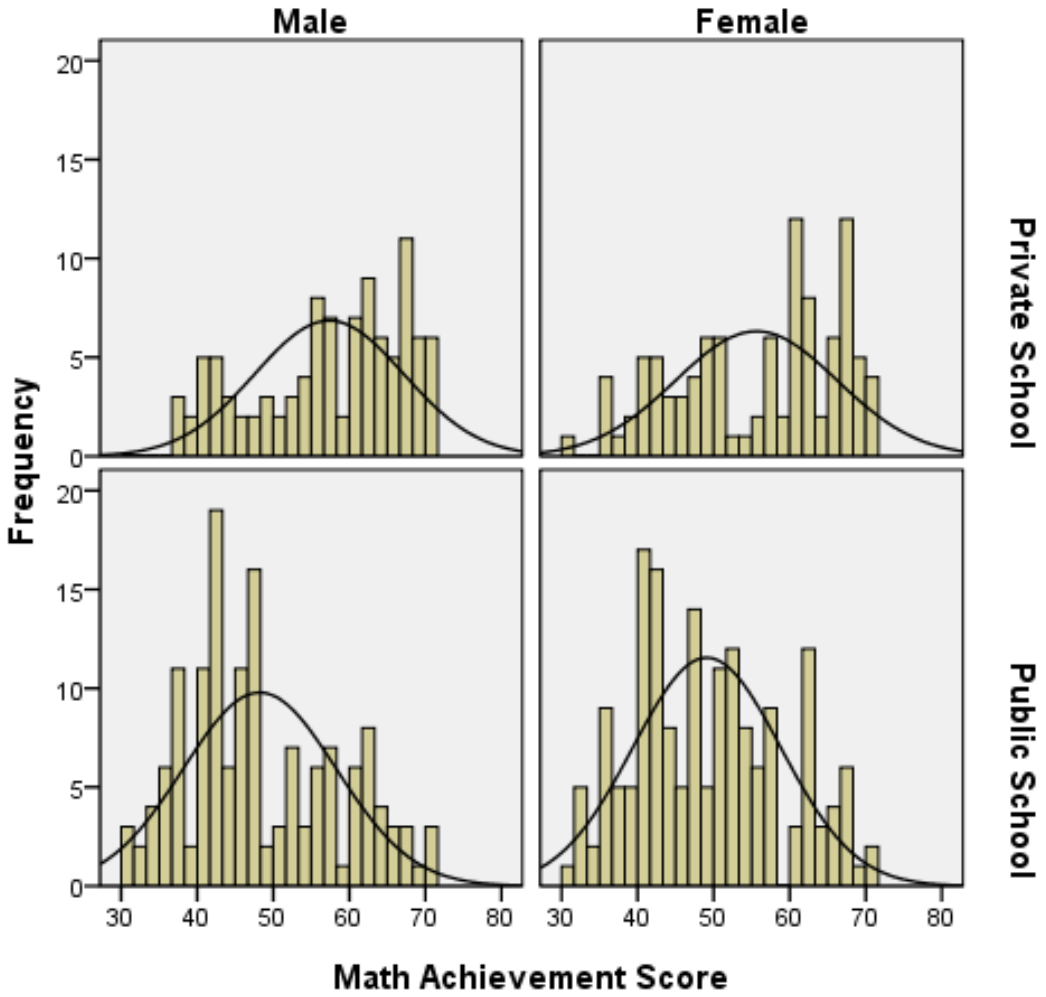
Exploring Math Achievement: Digging Deeper



Know that there are histograms within histograms. Wonder about them!

There are also histograms within scatterplots…

Figure 2.3. Histograms of math achievement scores partitioned by sex of the student and the public/private status of the student's school.



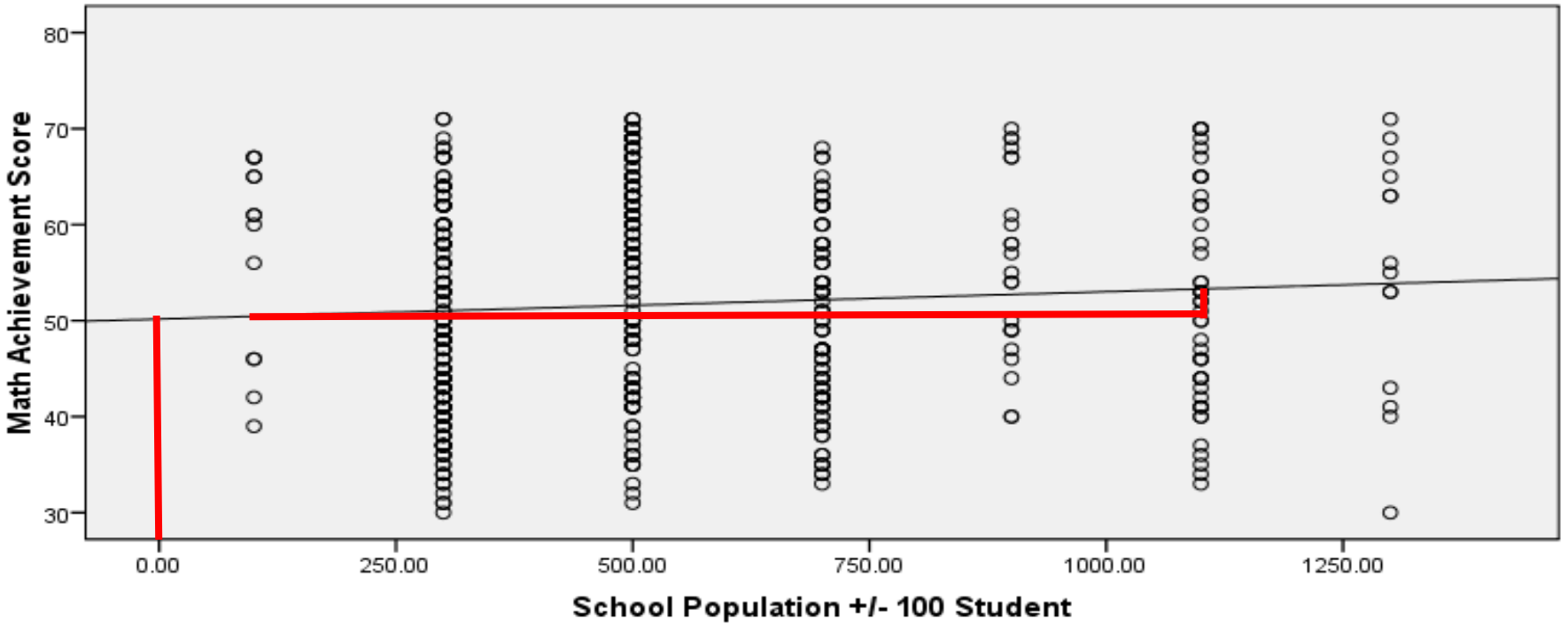
Exploring Math Achievement and School Size

Coefficients[a]

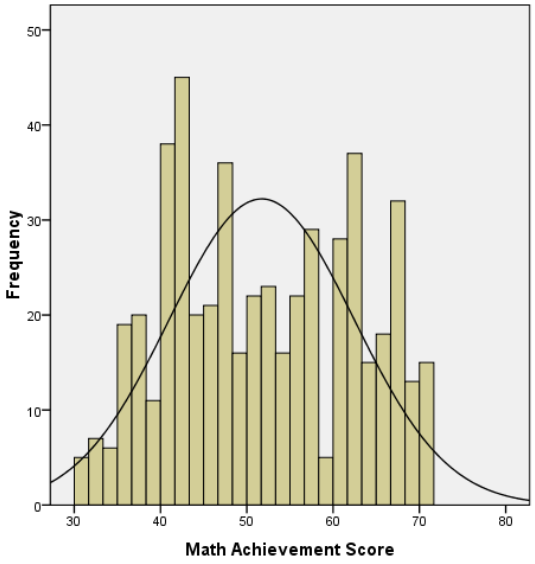
Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	50.167	1.029		48.767	.000
	School Population +/- 100 Student	.003	.002	.075	1.700	.090

a. Dependent Variable: Math Achievement Score

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



Exploring Math Achievement and School Size



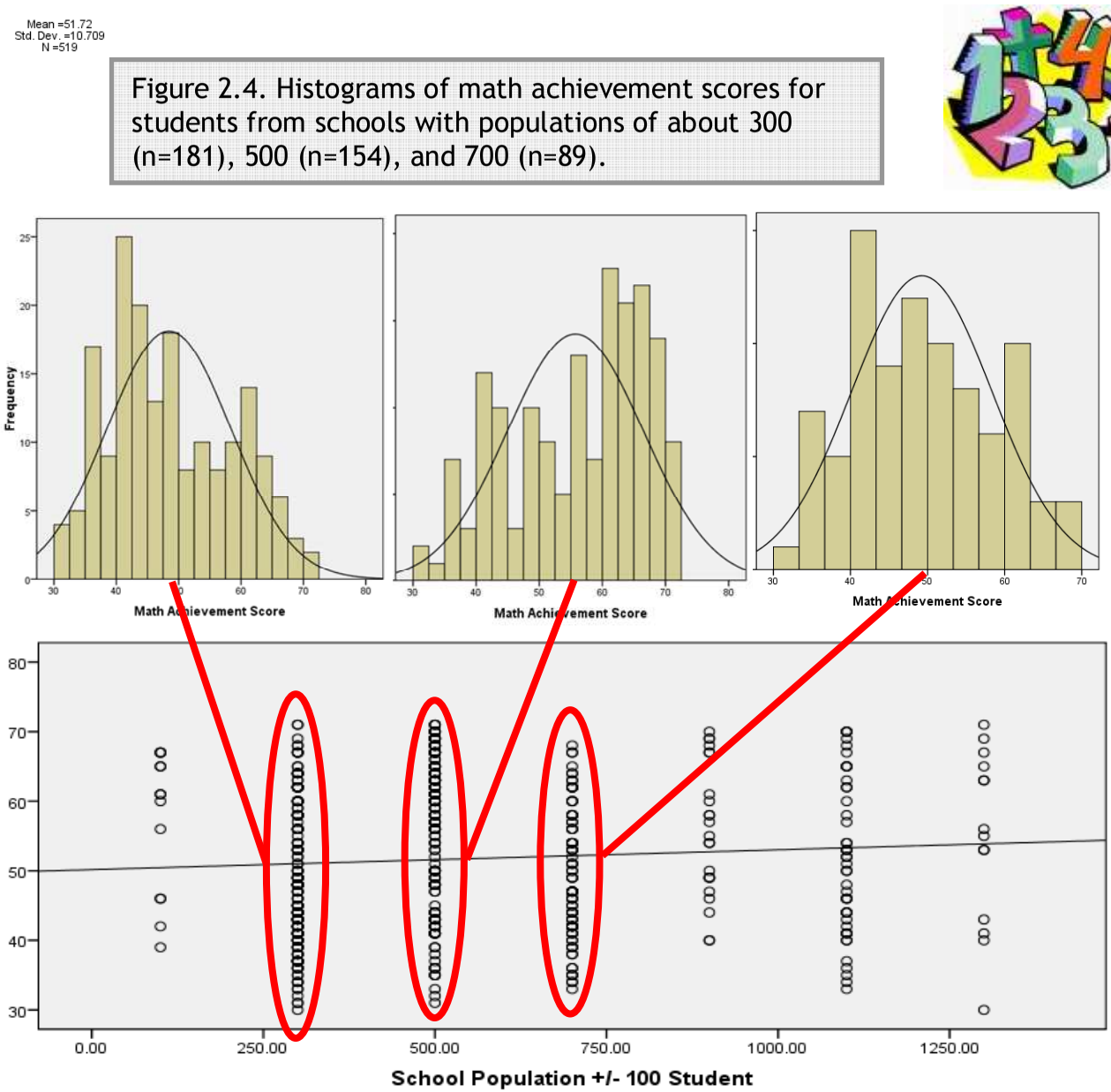
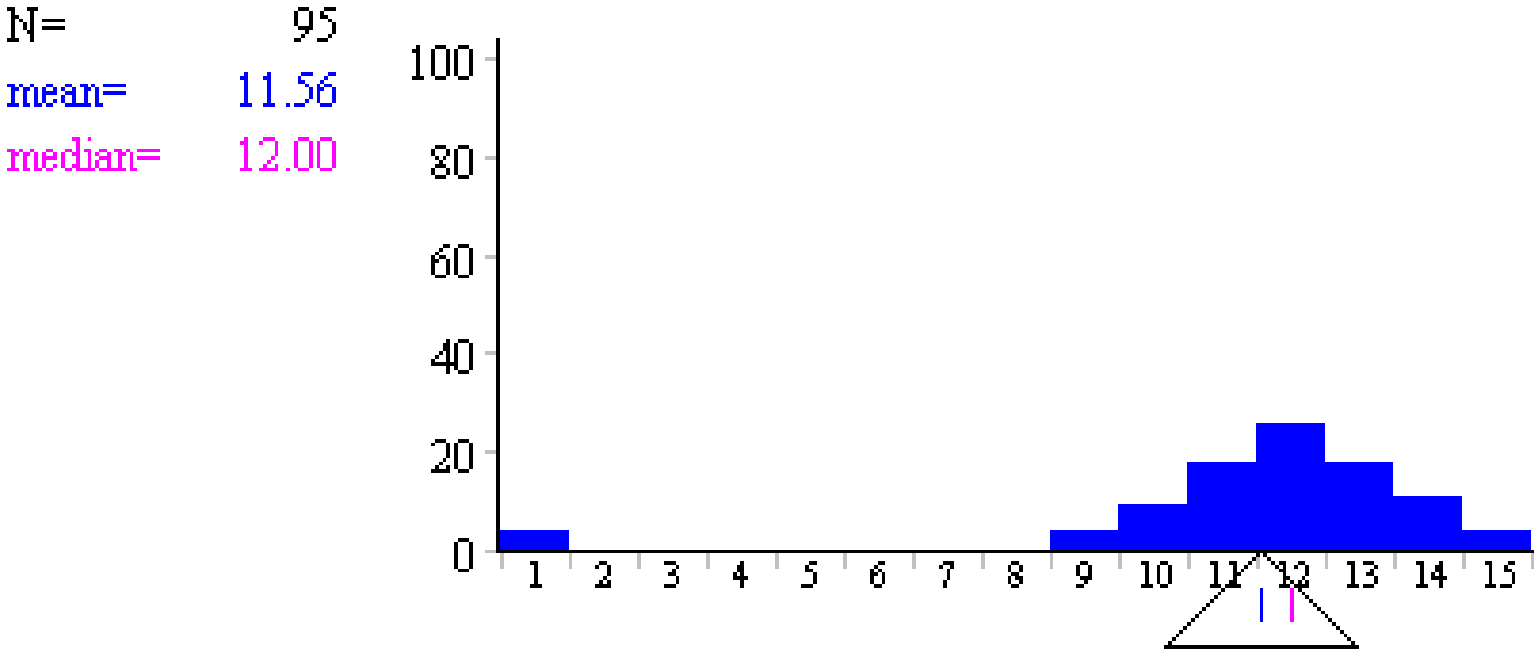


Figure 2.4. Histograms of math achievement scores for students from schools with populations of about 300 (n=181), 500 (n=154), and 700 (n=89).

There are several reasons to distrust this fitted regression line. We will explore them in detail this semester. However, the reason I would like to emphasize here is the one reason to which we will give the least attention: non-independence. Note that there are only a few schools in this sample (represented by a handful of students each). Only one school has about 100 students, and only one school has about 1300 students. If, for example, the 100-student school were a bad school, and the 1300-student school were a good school, the schools would determine our results (not the students' school population sizes per se).

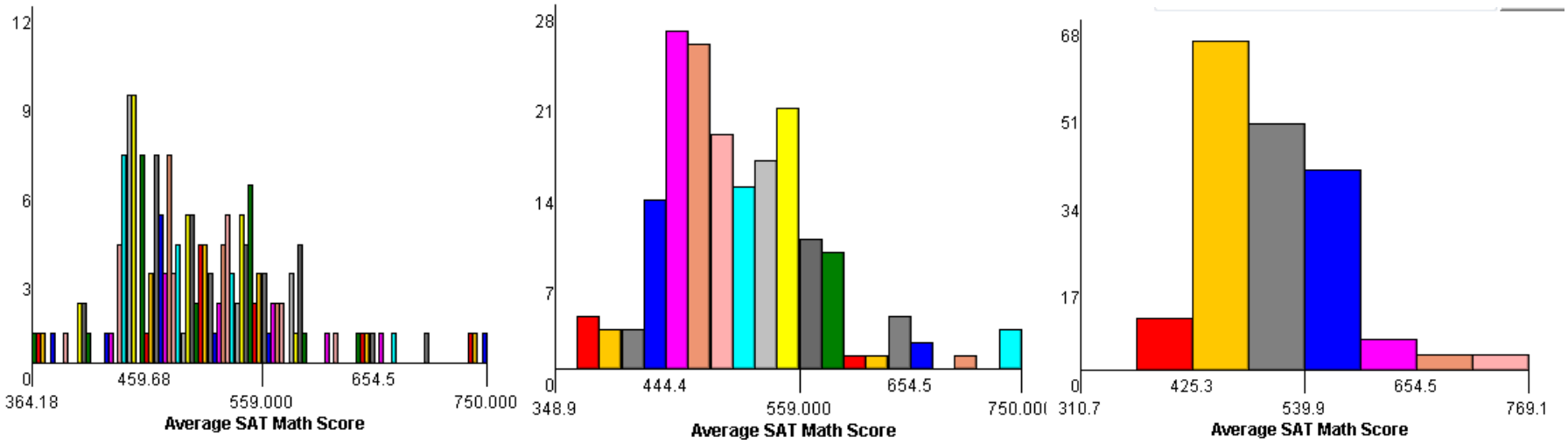
Outlier Resistant vs. Outlier Sensitive Statistics



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

When the median and the mean differ, more than half the sample will be above or below average (the mean). Can you explain that to the school board? (It helps to talk about the average income in the room and what happens when Bill Gates walks through the door—"The Bill Gates Effect.")

Exploring the Effect of Bin Size



http://www.shodor.org/interactivate/activities/Histogram/

Do not be fooled by arbitrary choices such as the size of bins or the length of axes.

Other Explorations:

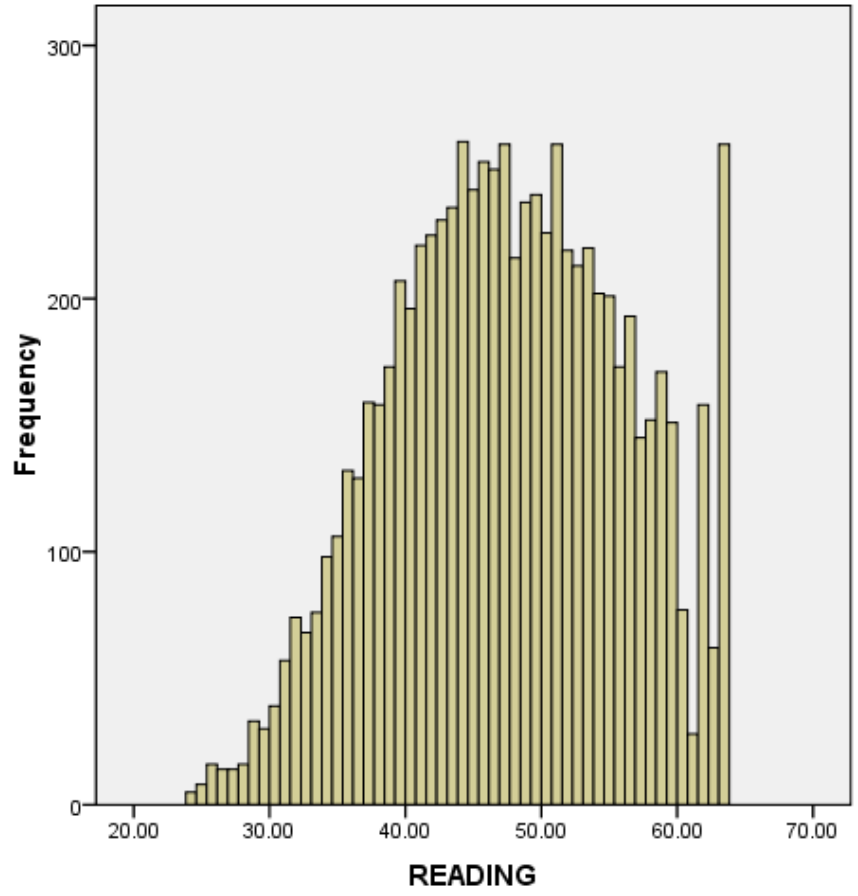
http://www.shodor.org/interactivate/activities/NormalDistribution/

http://www.shodor.org/interactivate/activities/StemAndLeafPlotter/

http://www.shodor.org/interactivate/activities/Regression/

Answering our Roadmap Question

Unit 2: In our sample, what does reading achievement look like (from an outlier resistant 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 ranges from 24 to 63. The scores are fairly normally distributed except for a ceiling effect caused by a top test score of 63 attained by 261 students. The median student's score is 47. The midspread is 13 points (from a score of 41 to a score of 54). There are no extreme outliers.

Unit 2 Appendix: Key Concepts

- Do not confuse “percentage differences” with “percentage point differences.”
- Rule of Thumb #1: Rules Of thumb only work when they work. Use your own judgment.
- The normal distribution is not particularly loved by the gods. Rather, the normal distribution is a result of a common kind of randomness resulting from the accumulation of many chance events. It will play a pivotal role in the machinery of statistical hypothesis testing.
- For exploratory purposes, look with soft eyes. We are trying to see beyond the sample into the population. It’s a little mystical, I know.
- Outlier resistant statistics such as the median and midspread can help us look with soft eyes. They minimize the influence of outliers.
- When the median and the mean differ, more than half the sample will be above or below average (the mean). Can you explain that to the school board? (It helps to talk about the average income in the room and what happens when Bill Gates walks through the door—”The Bill Gates Effect.”)
- Do not be fooled by arbitrary choices such as the size of bins or the length of axes.

Unit 2 Appendix: Key Interpretations

The Shape of a Distribution:

"The distribution of math achievement scores is bimodal."

"The distribution of SAT scores is approximately normal."

"The distribution of annual household income is positively skewed."

The Spread of a Distribution:

"The distribution ranges from 60 to 102. The midspread of the distribution is 27 points. Using 1.5 times the midspread to determine the reasonable upper and lower bounds for outliers, we notice that there are five positive outliers (87, 87, 91, 93, and 102) but no negative outliers."

"In our sample of 7,800 students, the distribution of reading scores ranges from 24 to 63. The scores are fairly normally distributed except for a ceiling effect caused by a top test score of 63 attained by 261 students. The median student's score is 47. The midspread is 13 points (from a score of 41 to a score of 54). There are no extreme outliers."

Dichotomies:

"In our sample, 27% of the subjects were depressed, and 73% were not."

Unit 2 Appendix: Key Terminology

- **Spread, Location and Shape (SPLASH):** Also known as the midspread/min/max/RLB/RUB, median and kurtosis/skew/modality.
- **Midspread:** The range of the middle 50% of observations when you sort all observations from low to high.
- **Median**: A measure of central tendency of a distribution (aka an average, or the location) represented by the middle observation when you sort all observations from low to high.
- **Kurtosis:** The peakiness or flatness of a distribution.
- **Skew:** The tailedness of a distribution, where skew is in the whale's tail.
- **Modality:** The number of peaks a distribution has, where one-peaked distributions are "unimodal" and two-peaked distributions are "bimodal."
- **Reasonable Upper/Lower Bound for Outlier Detection (RUB/RLB):** The RUB is the 75th percentile plus one and a half midspreads. The RLB is the 25th percentile minus one and a half midspreads. These values can be helpful for detecting outliers.

Unit 2 Appendix: Key Terminology

Not Necessarily To Be Memorized

Percentage: If I scored 80% the first time I took a test, and then I scored 100% on the retake, my score went up 25% (because 80% + (80%*25%) = 100%).

Percentage Point: If I scored 80% the first time I took a test, and then I scored 100% on the retake, my score went up 20 percentage points.

Percentile: An n^{th} percentile is the value of your variable at which n percent of your values fall below.* Say the 25^{th} percentile of *MathAch* is 300, then 25% of the math achievement scores fall below 300.

Percentile Rank: In the example, a math achievement score of 300 has a percentile rank of 25. A percentile rank is a number from 0 to 99. A percentile is a number on the scale of the variable.

Median: The 50^{th} percentile. The middle observation when you line up the observations from lo to hi.

Upper Quartile: A marker of the top 25%, the 75^{th} percentile.

Lower Quartile: A marker of the bottom 25%, the 25^{th} percentile.

Tukey's Hinges: The 25^{th} and 75^{th} percentiles (basically).

Interquartile Range: The "distance" from the 25^{th} to the 75^{th} percentile (aka "Midspread").

Midspread: The "distance" from the 25^{th} to the 75^{th} percentile (aka "Interquartile Range").

Reasonable Upper/Lower Bound For Outlier Detection (RUB/RLB): The RUB is the 75^{th} percentile plus 1.5 midspreads. The RLB is the 25^{th} percentile minus 1.5 midspreads. This is a "rule of thumb."

Unit 2 Appendix: Key Terminology (draft)

Categorizing Variables

Measurement Perspective	Modeling Perspective: Continuous • spectrum-like	Modeling Perspective: Polychotomous • ≥3 categories	Modeling Perspective: Dichotomous • 2 categories
Ratio • interval • zero means none	*HOMEWORKHRS* *AGE*		
Interval • all units are equal • e.g., (1-0) = (10 - 9)	*READING* *SES*		
Ordinal • rankings • ordered categories		*COURSELEVEL* *MCAS* (*ADV*, *PROF*, *Ni*, *F*)	
Nominal • names • unordered categories		*RACE* *RELIGION*	*MALE* *FREELUNCH*

Unit 2 Appendix: Key Terminology (Draft)

Modeling:

Measurement:

String	Dichotomous	Tricky*	Continuous	
Nominal	Nominal	Ordinal	Interval	Ratio
RACER	*MEDALIST*	*RANK*	*TIMEB* (Minutes Behind Leader)	*TIMEA* (Total Minutes)
Meaghan	1	1st	0	41
Josepha	1	2nd	2	43
Kristin	1	3rd	3	44
Suzanne	0	4th	6	47
Katani	0	5th	7	48
Rachael	0	6th	8	49
Shelley	0	7th	9	50
Amy	0	8th	15	56
Ines	0	9th	16	57
Jennifer	0	10th	19	60

- The *RANK* variable has too many categories to be polychotomous, one for each observation. It's almost continuous but not quite. In Appendix A, I will (briefly) introduce tools that are analogous to regression but for ranks.
- Maybe *RANK* would be polychotomous if it were rankings in a qualifying heat, and the three categories were: qualified, alternate, dnq.
- Notice that for *TIMEB*, zero does not mean zero time.

Unit 2 Appendix: SPSS Syntax

```
**************************************************************************************.
*I'm going to produce univariate descriptive statistics and a histogram with a normal curve overlay
for the variable MATHACH.
*/NTILES=4 asks for quartiles.
*/STATISTICS=STDDEV MINIMUM MAXIMUM MEAN is fairly obvious.
*/HISTOGRAM NORMAL is also fairly obvious.
*Forget about the other lines for now.
**************************************************************************************.

FREQUENCIES VARIABLES=MATHACH
  /FORMAT=NOTABLE
  /NTILES=4
  /STATISTICS=STDDEV MINIMUM MAXIMUM MEAN
  /HISTOGRAM NORMAL
  /ORDER=ANALYSIS.
```

Unit 2 Appendix: R Syntax

```
#-------------------------------------------------------------------------------------------------------------------
# I'm going to produce univariate descriptive statistics and a histogram for the variable MATHACH.
#-------------------------------------------------------------------------------------------------------------------

Dataset <-
  read.spss("F:/CD140/Data Sets/NELS Math Achievement/NELS88Math.sav",
  use.value.labels=TRUE, max.value.labels=Inf, to.data.frame=TRUE)

Hist(Dataset$MathAch, scale="frequency", breaks="Sturges", col="darkgray")

library(abind, pos=4)

numSummary(Dataset[,"MathAch"], statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
```

Perceived Intimacy of Adolescent Girls (Intimacy.sav)

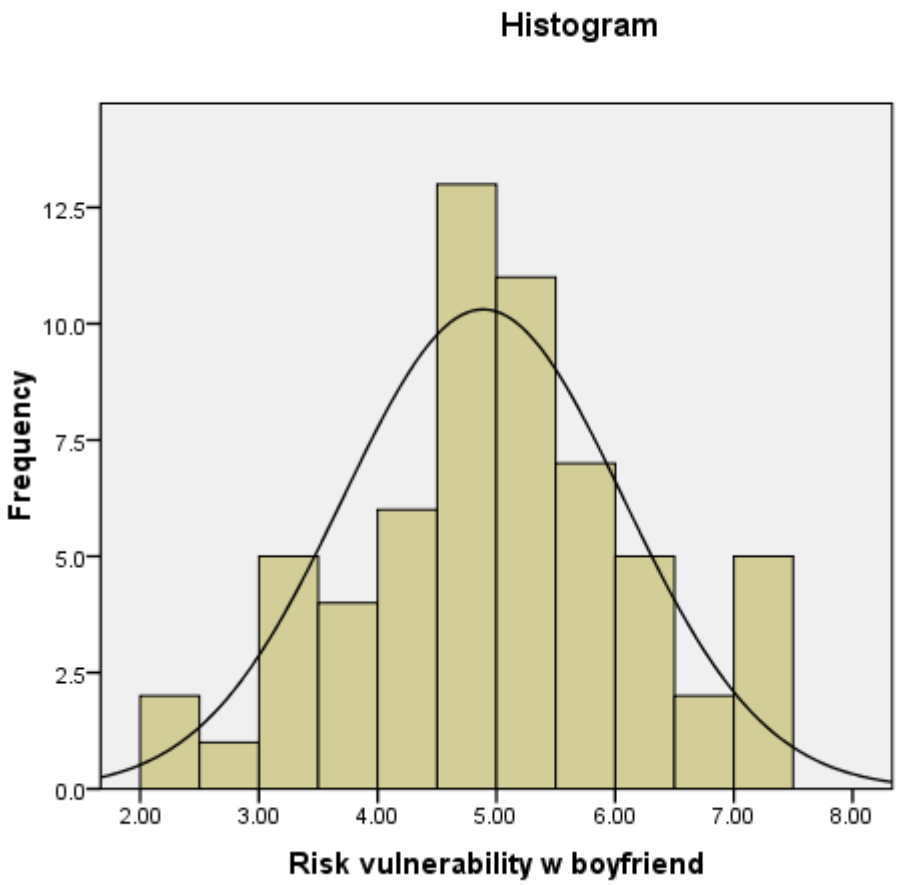
- **Overview: Dataset contains self-ratings of the intimacy that adolescent girls perceive themselves as having with: (a) their mother and (b) their boyfriend.**

- **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)

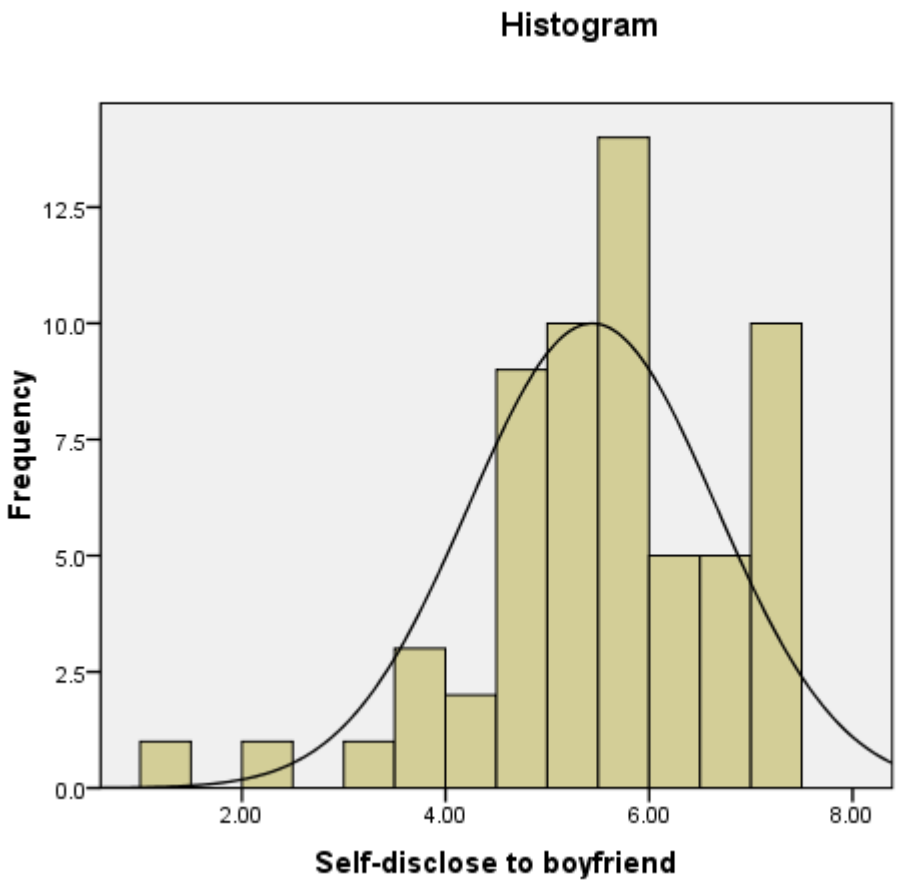


Statistics

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)

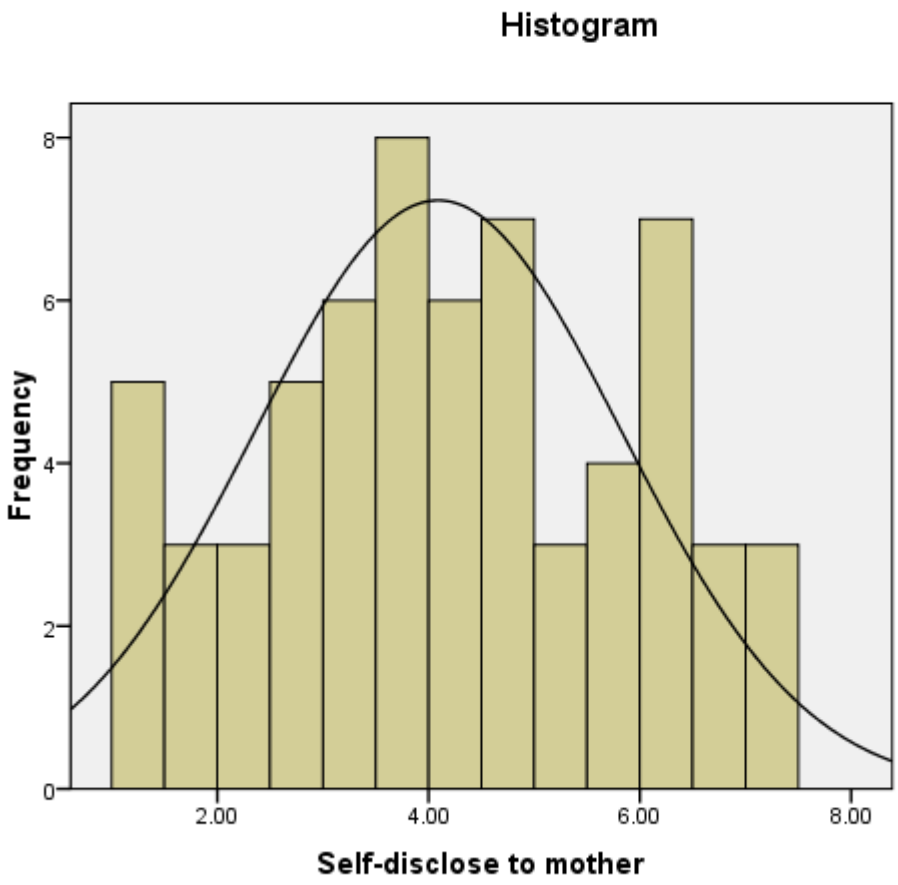


Statistics

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)

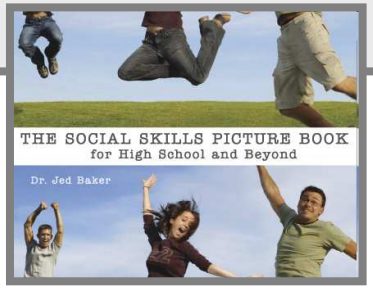


Statistics

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:**

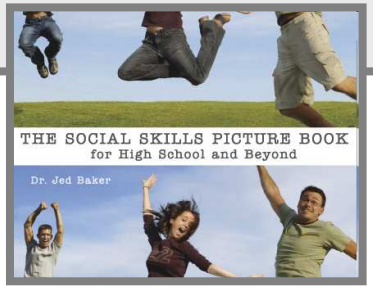
Variables about the student—

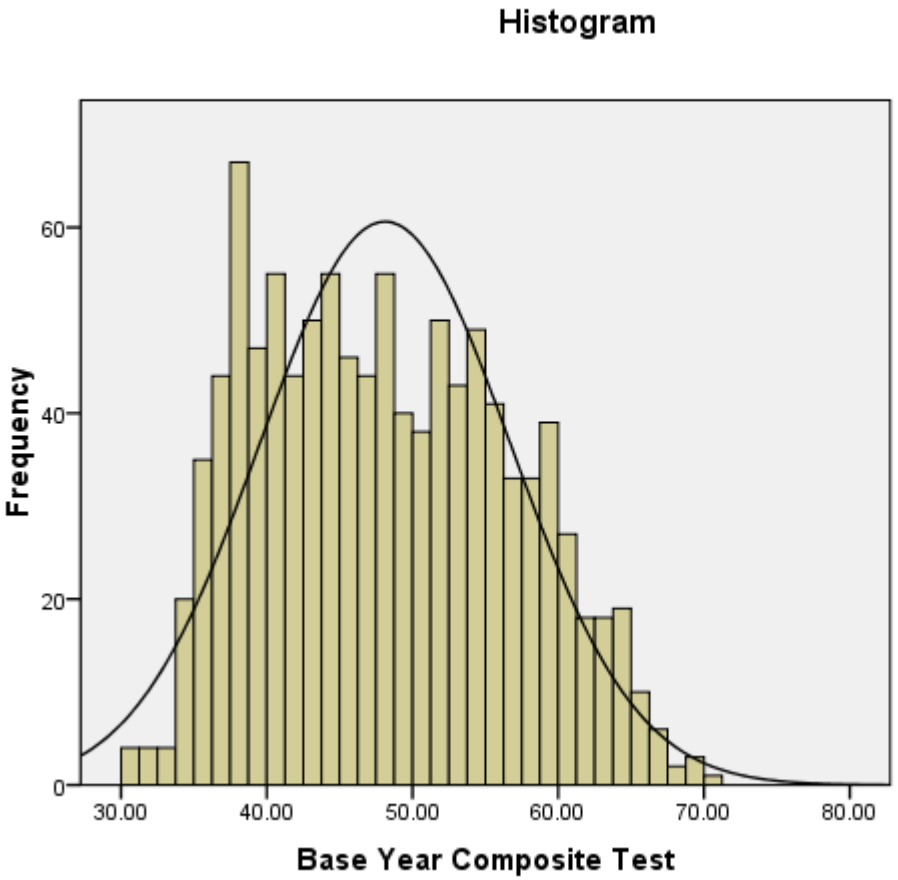
(Black) 1=Black, 0=Other
(Latin) 1=Latino/a, 0=Other
(Sex) 1=Female, 0=Male
(BYSES) Base year SES
(GPA80) HS GPA in 1980
(GPS82) HS GPA in 1982
(BYTest) Base year composite of reading and math tests
(BBConc) Base year self concept
(FEConc) First Follow-up self concept

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)



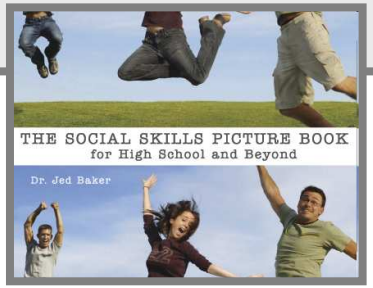


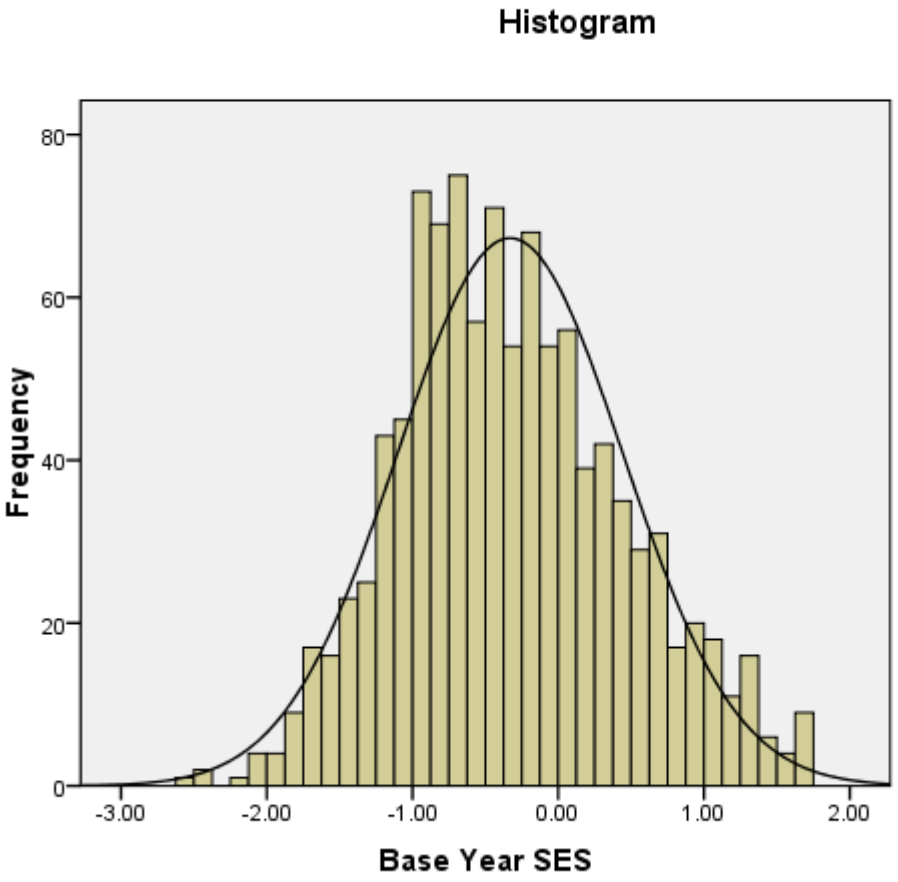
Statistics

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)



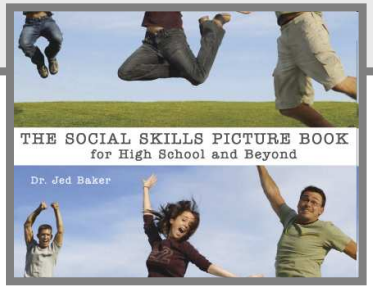


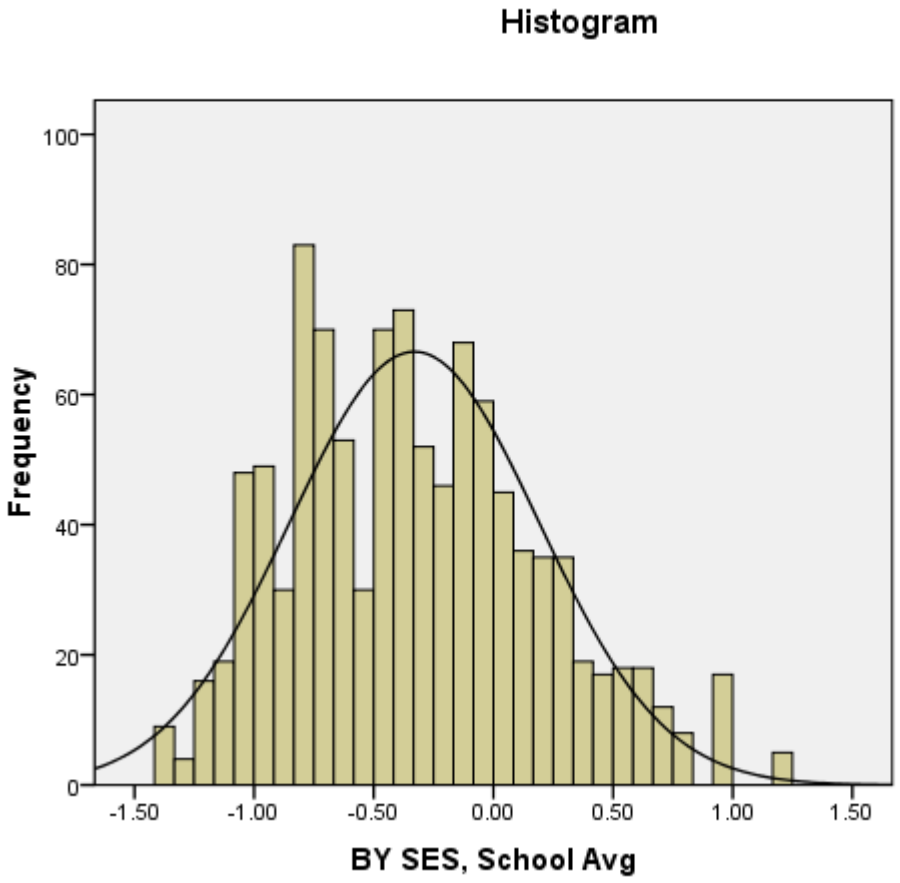
Statistics

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

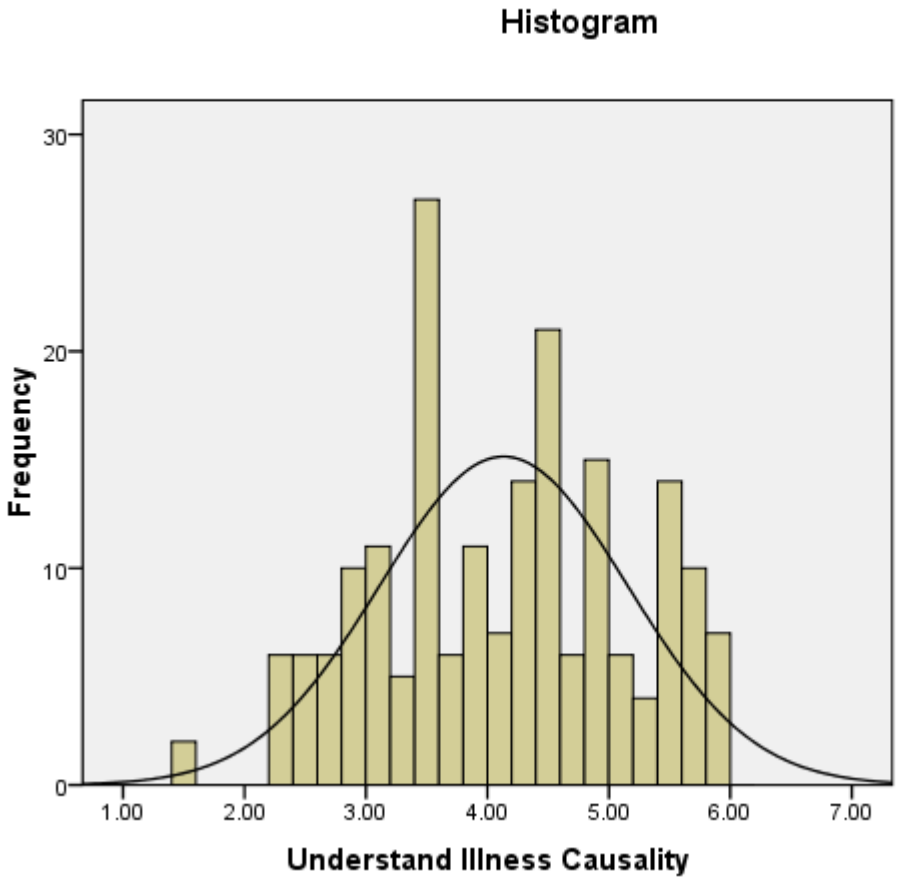
N	Valid	1044.0000
	Missing	.0000
Mean		-.3304
Std. Deviation		.5211
Minimum		-1.3625
Maximum		1.2240
Percentiles	25	-.7483
	50	-.3640
	75	.0103

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)

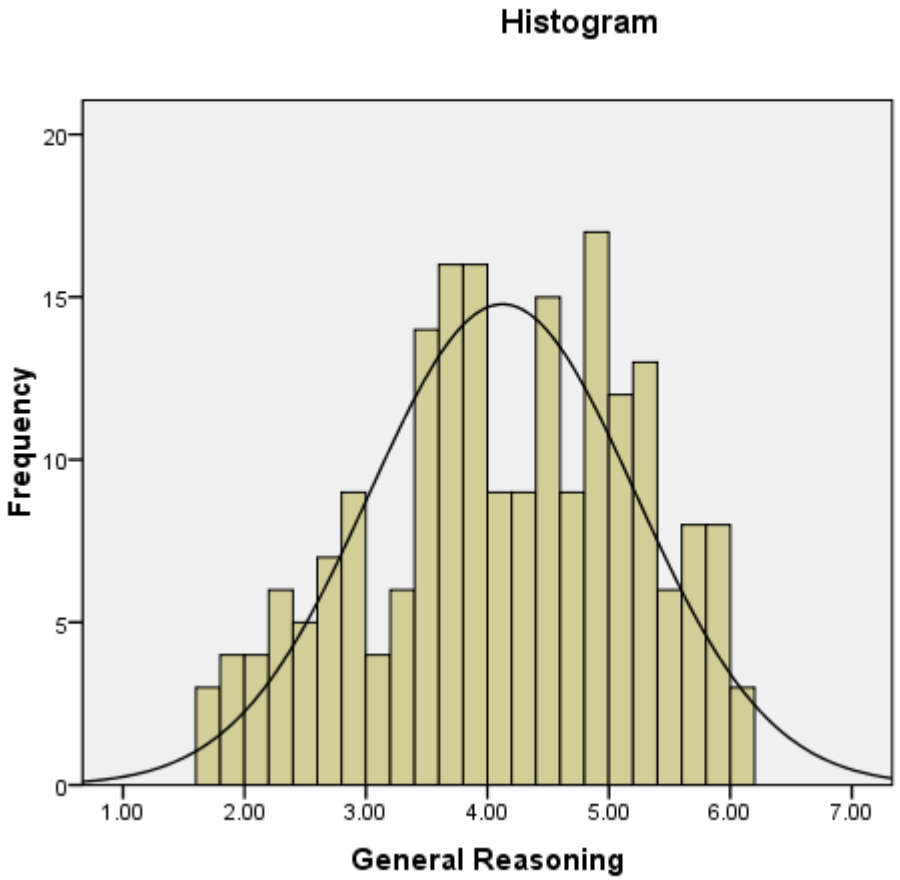


Statistics

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)

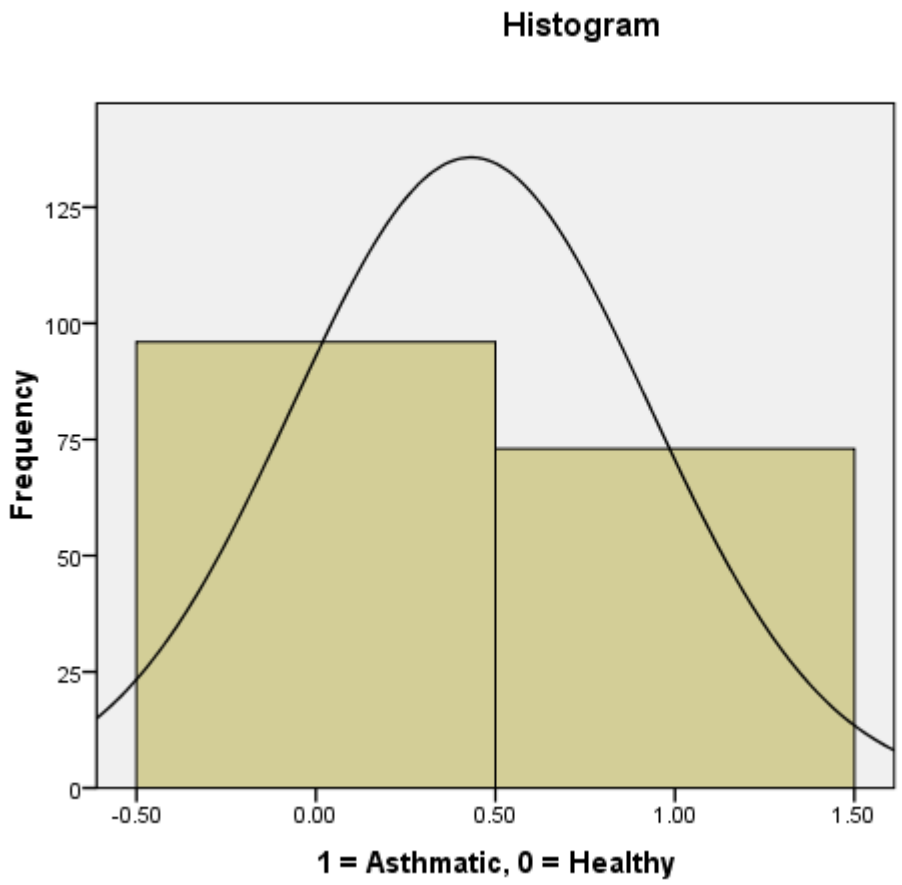


Statistics

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

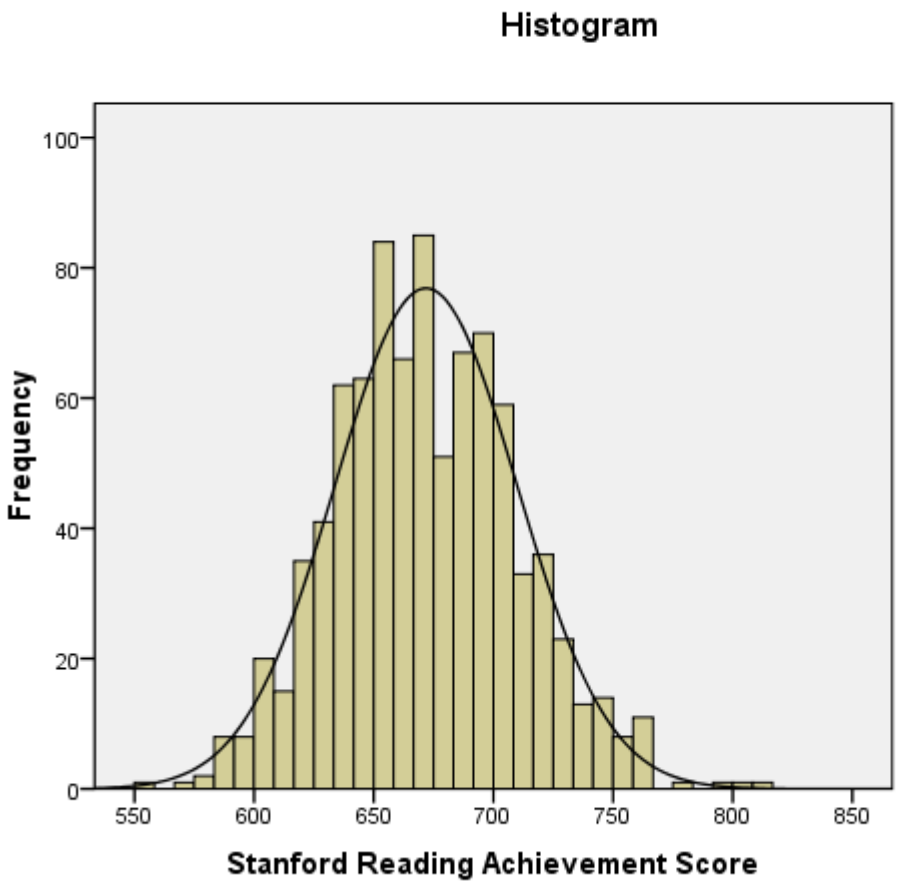
N	Valid	169.0000
	Missing	36.0000
Mean		.4320
Std. Deviation		.4968
Minimum		.0000
Maximum		1.0000
Percentiles	25	.0000
	50	.0000
	75	1.0000

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)

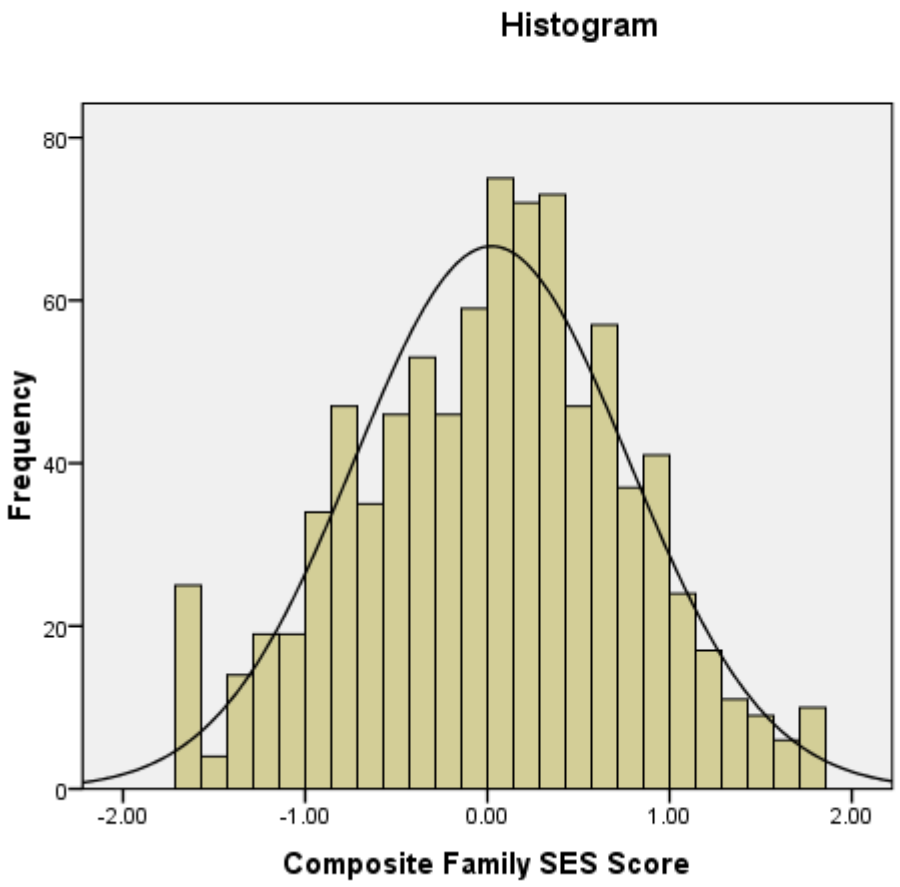


Statistics

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)

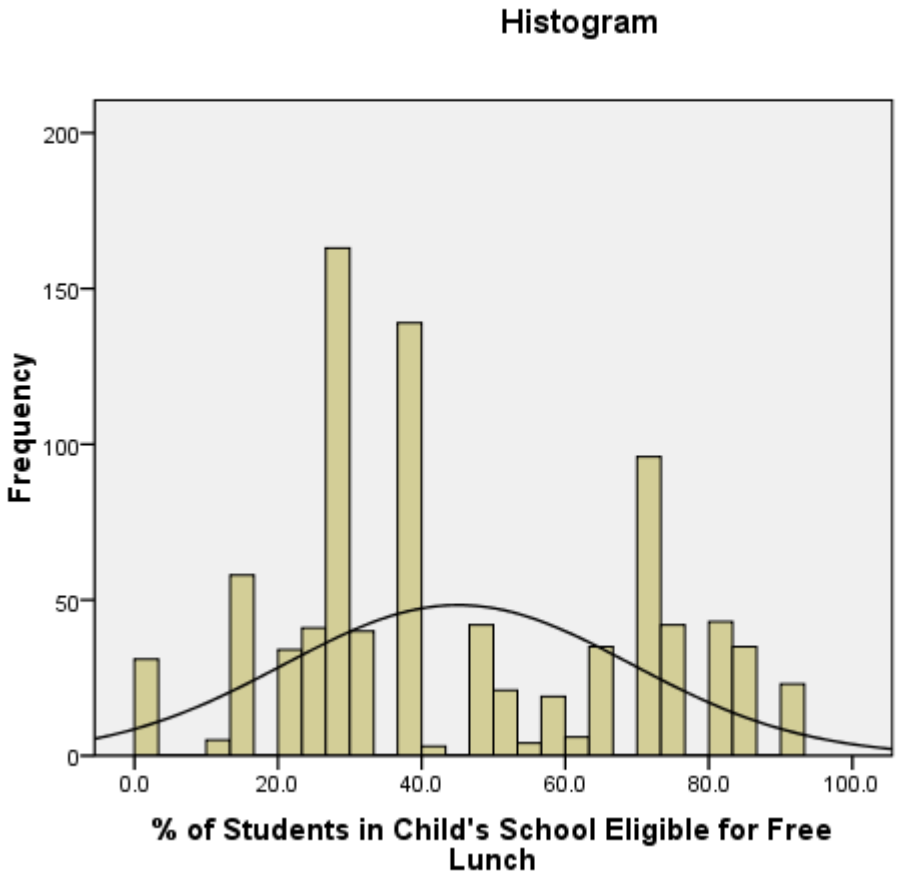


Statistics

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

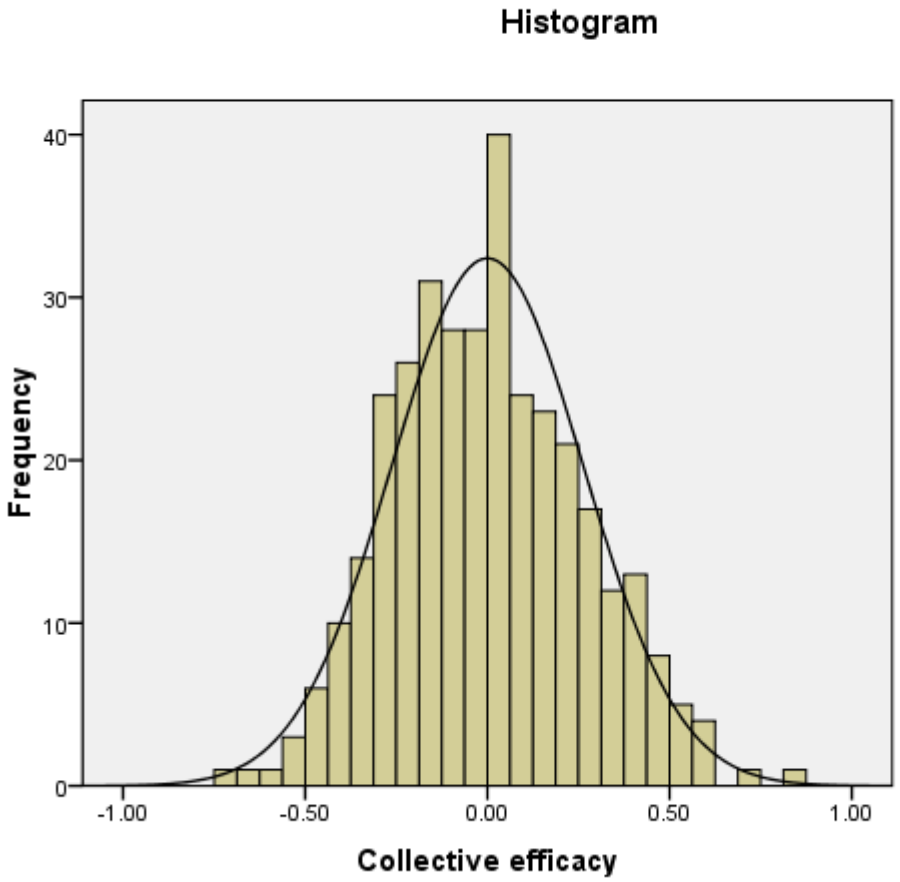
N	Valid	880.000
	Missing	.000
Mean		45.073
Std. Deviation		24.194
Minimum		.000
Maximum		92.300
Percentiles	25	27.500
	50	38.900
	75	72.200

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)

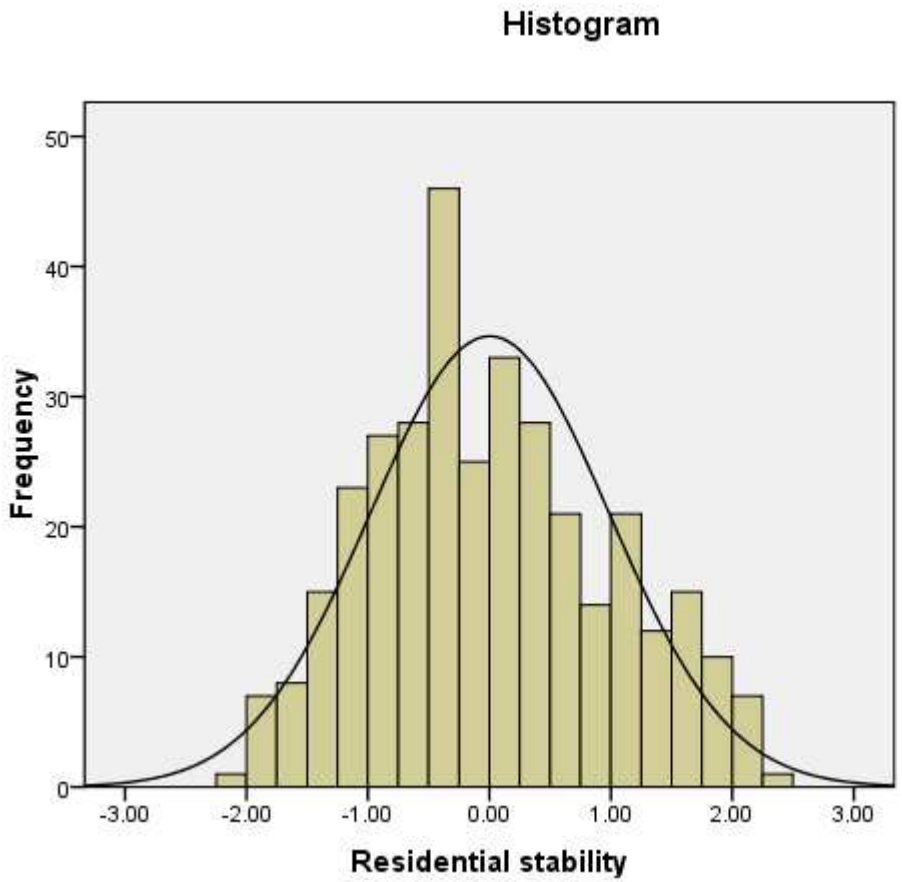


Statistics

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)

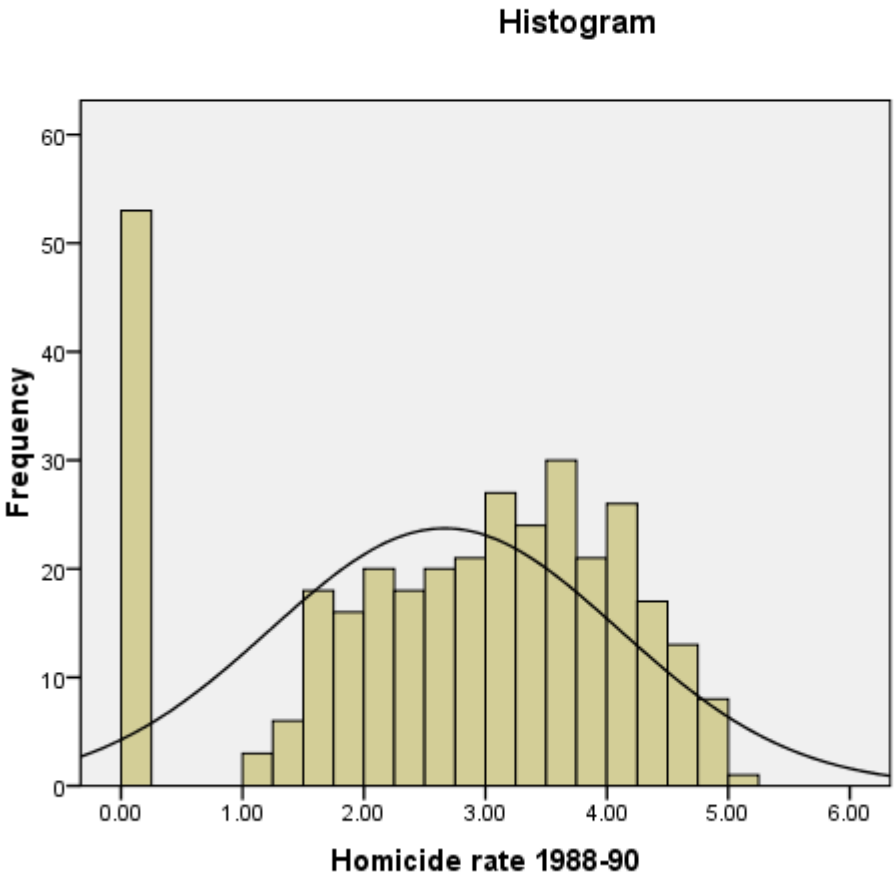


Statistics

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)

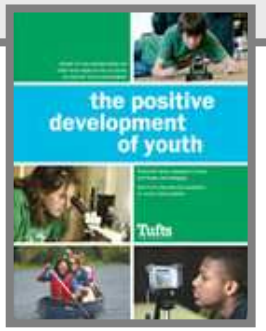


Statistics

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)



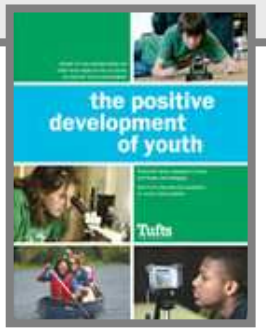
- 4-H Study of Positive Youth Development
- Source: Subset of data from IARYD, Tufts University

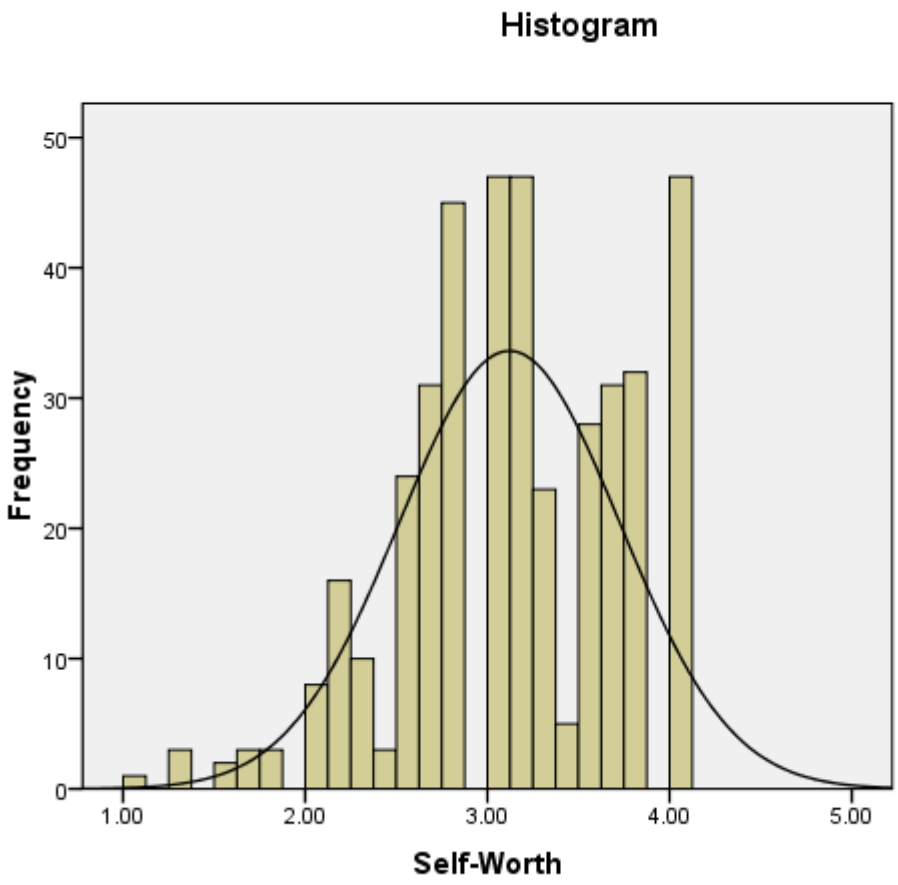
- Sample: These data consist of seventh graders who participated in Wave 3 of the 4-H Study of Positive Youth Development at Tufts University. This subfile is a substantially sampled-down version of the original file, as all the cases with any missing data on these selected variables were eliminated.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



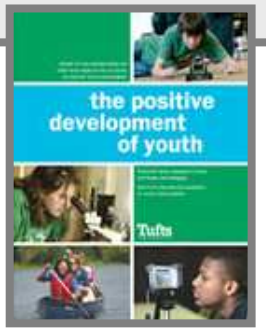


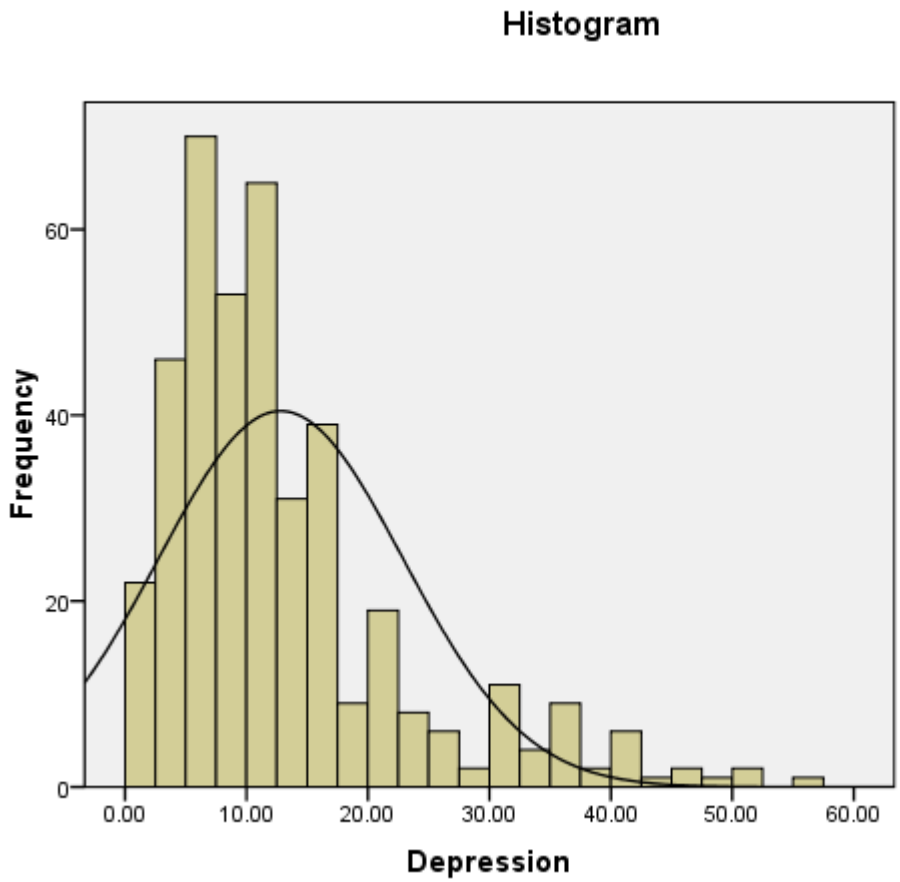
Statistics

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)

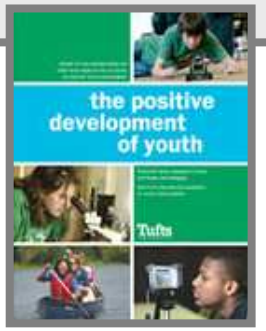


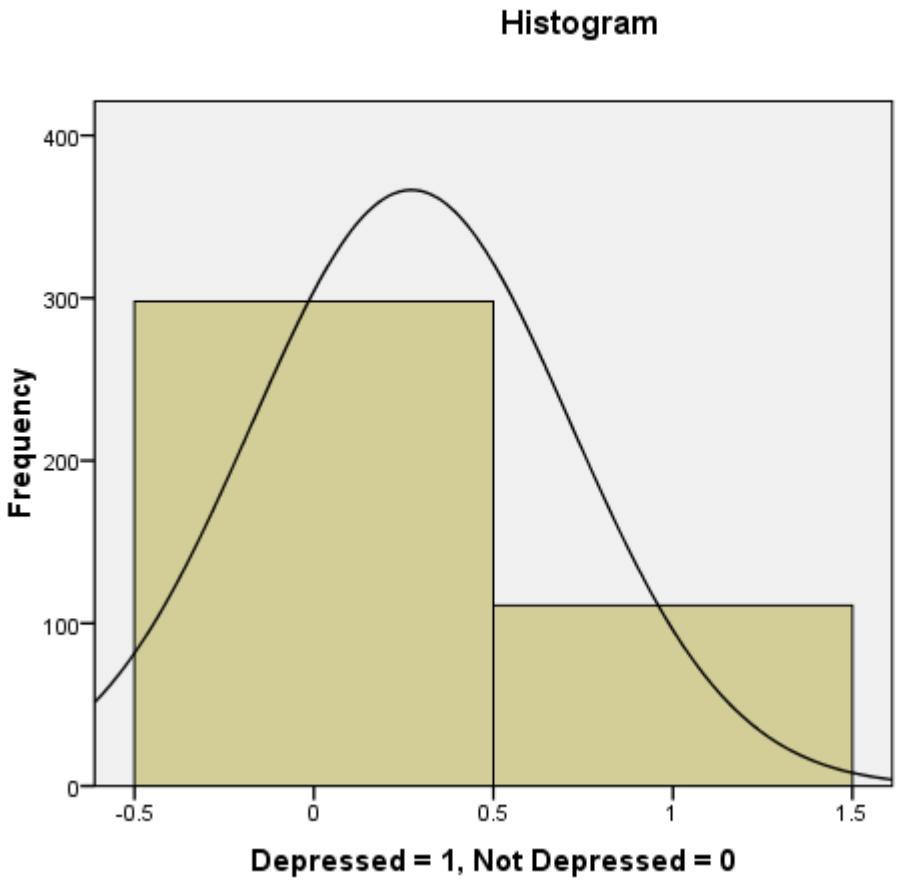


Statistics

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