Unit 6: Statistical Inference and Statistical Significance (t-tests)

Unit 6 Post Hole:

State the null hypothesis of a test for statistical significance; reject (or not) the null hypothesis; and, draw an inference (or not) from a sample to a population.

Unit 6 Technical Memo and School Board Memo:

From your regression analyses in Memo 5, draw conclusions from your sample to the population when appropriate.

Unit 6 (and Units 7 and 8) Reading:

http://onlinestatbook.com/

Chapter 5, Probability	Chapter 6, Normal Distributions
Chapter 7, Sampling Distributions	Chapter 8, Estimation
Chapter 9, Logic Of Hypothesis Testing	Chapter 10, Testing Means
Chapter 11, Power	Chapter 12, Prediction

Unit 6: 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. (4 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 <u>descriptive statistics</u>. 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 <u>descriptive statistics</u>.**
  - **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 6: Technical Memo and School Board Memo

Work Products (Part II of II):

I. Technical Memo (continued)

- D. Regression Analysis. Answer your research question using <u>inferential statistics</u>. (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 6: 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 6: 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 6: Roadmap (SPSS Output)

Model Summary

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

a. Predictors: (Constant), FREELUNCH

Unit 5

ANOVA[b]

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

a. Predictors: (Constant), FREELUNCH

b. Dependent Variable: READING

Unit 9

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

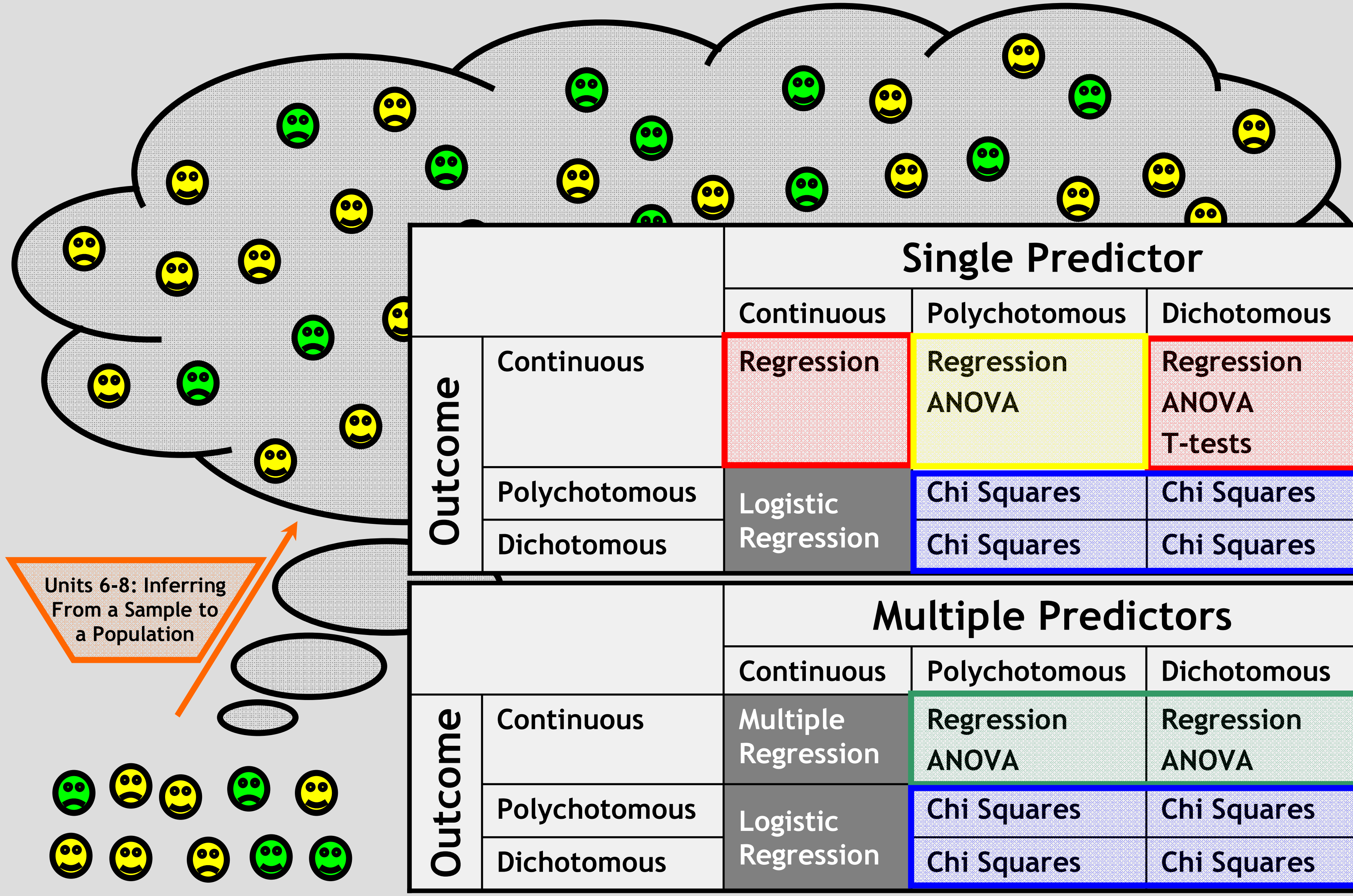
Coefficients[a]

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 6: Road Map (Schematic)



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

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

Epistemological Minute

The Type/Token Distinction—

<u>Dog Dog Cat</u>

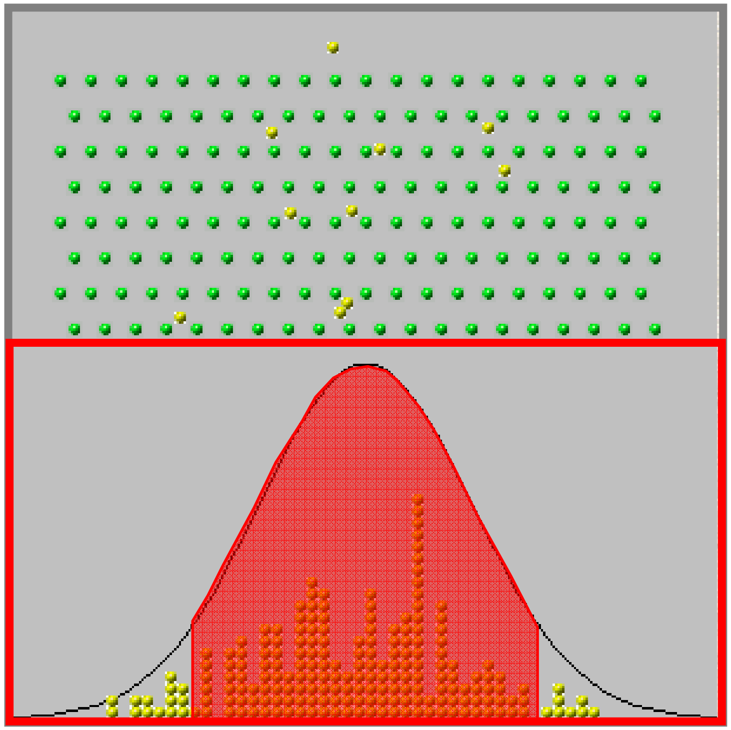
How many words are on the above line?

If you answered "three," you were thinking in terms of word <u>tokens</u>.

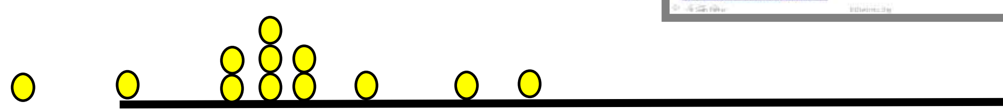
If you answered "two," you were thinking in terms of word <u>types</u>.

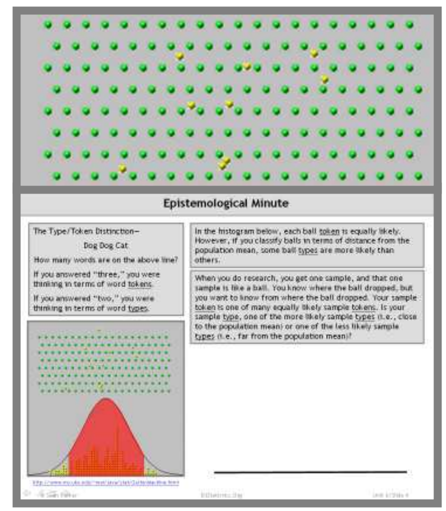
In the histogram below, each ball <u>token</u> is equally likely. Imagine they were numbered as in a lottery. However, if you classify balls in terms of distance from the population mean, some ball <u>types</u> are more likely than others.

When you do research, you get one sample, and that one sample is like a ball. You know where the ball dropped, but you want to know *from where* the ball dropped. Your sample <u>token</u> is one of many equally likely sample <u>tokens</u>. Is your sample <u>type</u>, one of the more likely sample <u>types</u> (i.e., close to the population mean) or one of the less likely sample <u>types</u> (i.e., far from the population mean)?



http://www.ms.uky.edu/~mai/java/stat/GaltonMachine.html





The Dialectic (and Diarrhea) of Statistical Inference

The Critics

In Unit 6, 7 and 8, we are going to step into abstraction. Each element of abstraction is part of an elaborate dance. The purpose of the dance is to craft a statistical argument that justifies an inference from a sample to a population. In the next few slides, I will attempt to personify the argument/dance/dialectic. I will frame it as a press conference between you and your critics.

You may wonder why we bother with the rigmarole. We bother because nobody cares about our sample, not even us. The population is the thing, albeit an abstract thing.

You might think of the population as an unwieldy number of subjects, too many for our data gathering limitations. Although this may help you pedagogically, it is not quite right. Whenever we want to draw a conclusion about causes, development or types, our population is not only large, it is infinite. (I told you that population is an abstract thing!) Causal reasoning is abstract; it involves an invisible "Necessary Connexion" in addition to succession and correlation. Developmental reasoning is abstract; it involves an invisible trajectory in addition to change over time. Classificatory reasoning (e.g., about differences in types) is abstract; it involves categories that we construct in order to understand the world.

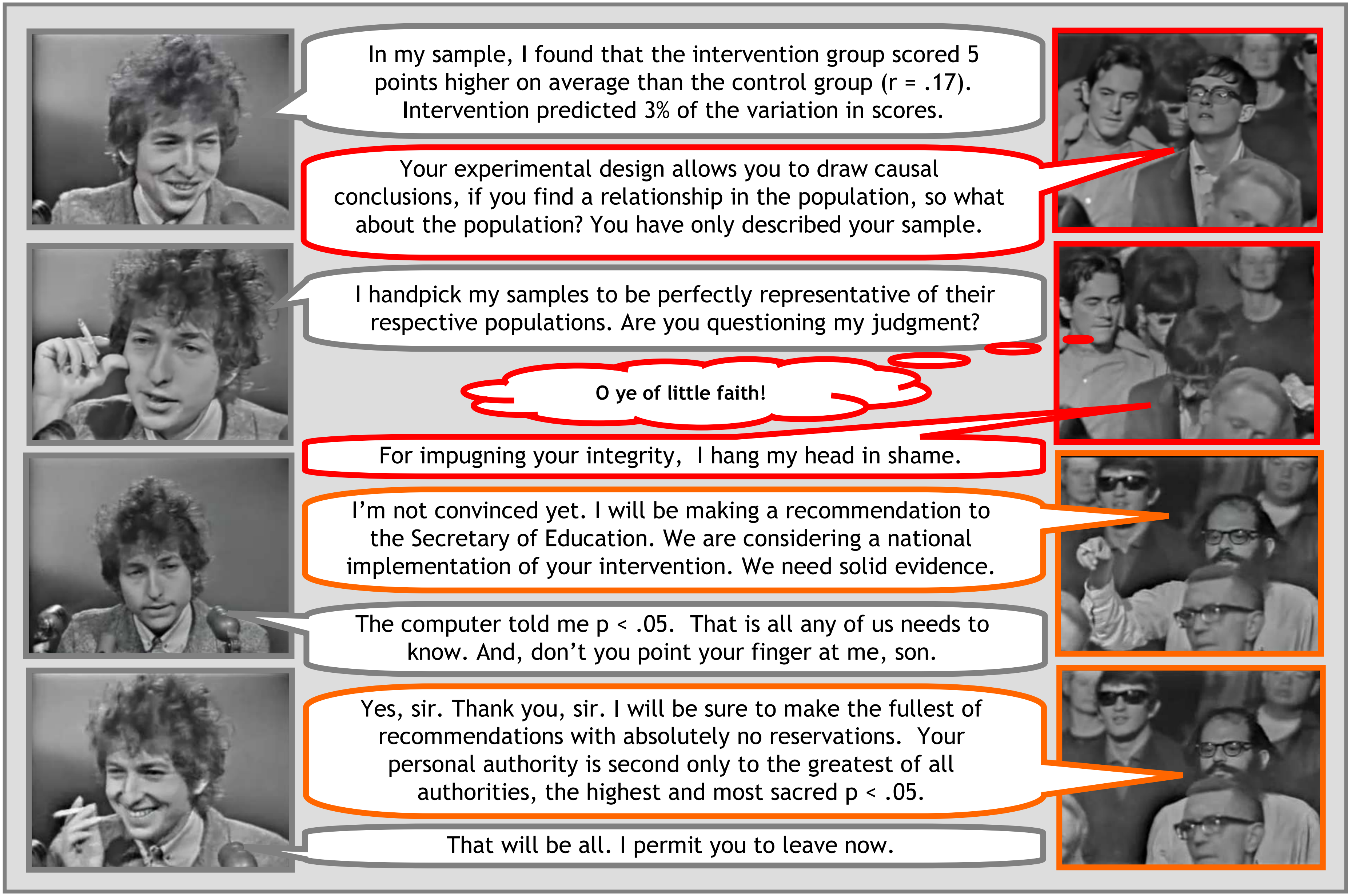
To illustrate the point, suppose you are thinking about the students in your class of twenty. You might be wondering if there is an SES achievement gap. Following are two questions, only the second of which gets at your question:

Do ***the*** low SES students perform worse than ***the*** high SES students in your class?
Do low SES students perform worse than high SES students in your class?

Whereas the first question is about the particular students in your class who happen to have one label or the other, "low" or "high," the second question is about types of students. You can definitively answer the first question based on your twenty students, but, for the second the questions, you may not have a large enough sample size.

You

Diarrhea of Statistical Inference (Part I of I)



Dialectic of Statistical Inference (Part I of II)

In my random sample, I found that the intervention group scored 5 points higher on average than the control group ($r = .17$, $p < .05$). Intervention predicted 3% of the score variation.

I suspect that there is no intervention effect, that the relationship you observe in your sample is merely an artifact of sampling error and not reflective of the population.

Well, the $p < .05$ tells us that you might be right (it's not $p = 0$), but if you were right, and thus there were no relationship in the population, we would only observe a relationship so strong (or stronger) less than 5% of the time. That is pretty unlikely.

I see. My suspicion of 0.00000 relationship is not very plausible.

Okay, so we do not want to conclude that the relationship is exactly zero in the population, but that does not mean the relationship is exactly .17 as you observe in your sample.

That's right. We don't want to conclude that the Pearson correlation in the population is exactly .17. Likewise, we do not want to conclude an intervention effect of exactly 5 points. However, they are unbiased estimates of the population values.

How precise are those estimates?

Using the same standard errors that we used to calculate $p < .05$ and reject the null hypothesis, we can construct 95% confidence intervals. Thus, our estimate for the population correlation is $.17 \pm .14$ and, for the intervention effect, 5 ± 4.

Dialectic of Statistical Inference (Part II of II)

Your intervention is boosting children's scores on average. A 1-point boost is your lower-bound estimate for that average, and a 9-point boost is your upper-bound estimate for that average. Should your intervention be an educational funding priority?

That is a difficult question. I can tell you all about ***statistical*** significance, but your question is about ***practical*** significance. To determine whether the intervention is worth implementing, we need to conduct a benefits-costs analysis: Do the benefits of the intervention outweigh its costs? Among the costs, we must consider opportunity costs: How does our intervention stack up against similarly targeted interventions?

My wife is an economist. I'll have her people call your people.

Back to statistical significance—The whole process of inference from a sample to a population is heavily laden with assumptions. If those assumptions do not hold, then it's all lies.

Yes. If my assumptions weren't tenable, my standard errors and, consequently, p-values and confidence intervals would be biased, and I would not be reporting them. Give me a break.

Hey, we're all friends. Were there *any* worrisome assumptions?

Independence, normality, linearity and outliers were okay, but there was a little heteroskedasticity; there was a little less variation in the intervention group than the control group. Heteroskedasticity won't bias our magnitude and strength estimates, but it will bias our precision estimate (i.e., standard error). Next semester, Sean will show me how to fix it.

Notes on the Dialectic (and Diarrhea) of Statistical Inference

All of our inferential machinery relies on the randomness of our sample. The population to which we infer is the population from which we randomly sampled.

The p-value of .05 is worshipped by many people. Do not take advantage of the worshippers, and do not become a worshipper yourself. It is a false idol.

<u>Unit 6</u> is about answering the critic who suspects that there is no relationship in the population. The critic's concern is sampling error. By accident of randomness, we always expect to observe some relationship.

<u>Unit 7</u> is about quantifying the uncertainty for our estimate. We use the standard error to build a confidence interval, which provides a plausible range of values encompassing (we hope) the exact population value.

Statistical significance and practical significance are two crucially distinct concepts. Determining the practical significance of a finding requires deep substantive knowledge and, perhaps, a little economics.

<u>Unit 8</u> is about the assumptions that underpin the trustworthiness of our standard error. Standard errors are the true workhorses of both statistical significance and confidence intervals.

Unit 6: Research Questions

Theory 1: Since depression leads to introversion, and reading is an introverted activity, depressed children will be stronger readers than non-depressed children.

Research Question 1: In children of immigrants, reading achievement is positively correlated with depression levels.

Theory 2: Since depression conflicts with cognitive functioning, and reading is a cognitively demanding activity, depressed children will be weaker readers than non-depressed children.

Research Question 2: In children of immigrants, reading achievement is negatively correlated with depression levels.

Data Set: ChildrenOfImmigrants.sav

Variables:
  Outcome—Reading Achievement Score (*READING*)
  Predictor—Depression Level (*DEPRESS*)

Model: $READING = \beta_0 + \beta_1 DEPRESS + \varepsilon$

ChildrenOfImmigrants.sav Codebook

Portes, Alejandro and Ruben G. Rumbaut
Children of Immigrants Longitudinal Study
(1992, 1995)

"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).

Subset of data: Random sample of 880 participants obtained through the website.

Selected references:

Portes, Alejandro, & Ruben G. Rumbaut (2001). *Legacies: The Story of the Immigrant Second Generation*. Berkeley CA: University of California Press.

More information is available at: http://cmd.princeton.edu/ (Center for Migration and Development, Princeton University)

ChildrenOfImmigrants.sav Codebook

Variable Name	Variable Description	Characteristics
ID	Identification #	Integers
Reading	Stanford Reading Achievement Score	Range: 527-830 Mean: 669
FreeLunch	% students in school who are eligible for free lunch program	Range: 0-92.30 Mean: 45.27
Male	Sex dummy variable	1=Male 0=Female
Depress	Depression scale (Higher score means more depressed)	Range: -1.68 - 5.57 Mean: 0.00
SES	Composite family SES score	Range: -1.66 - 2.09 Mean: -0.04

The Children of Immigrants Data Set

ChildrenOfImmigrants.sav [DataSet1] - SPSS Data Editor

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

1 : ID 4414

	ID	Reading	FreeLunch	Male	Depress	SES
1	4414	558	70.8	0	0.95820	-1.37
2	282	570	27.5	0	0.50933	-0.38
3	3848	580	92.3	1	0.94126	-1.02
4	342	581	27.5	1	-0.72716	0.25
5	3805	584	38.2	1	1.96360	0.07
6	4301	584	82.0	1	0.33707	-0.81
7	3548	586	82.0	0	-0.47796	-1.10
8	2593	589	38.2	0	2.45614	-1.65
9	3545	589	82.0	0	3.14966	-0.35

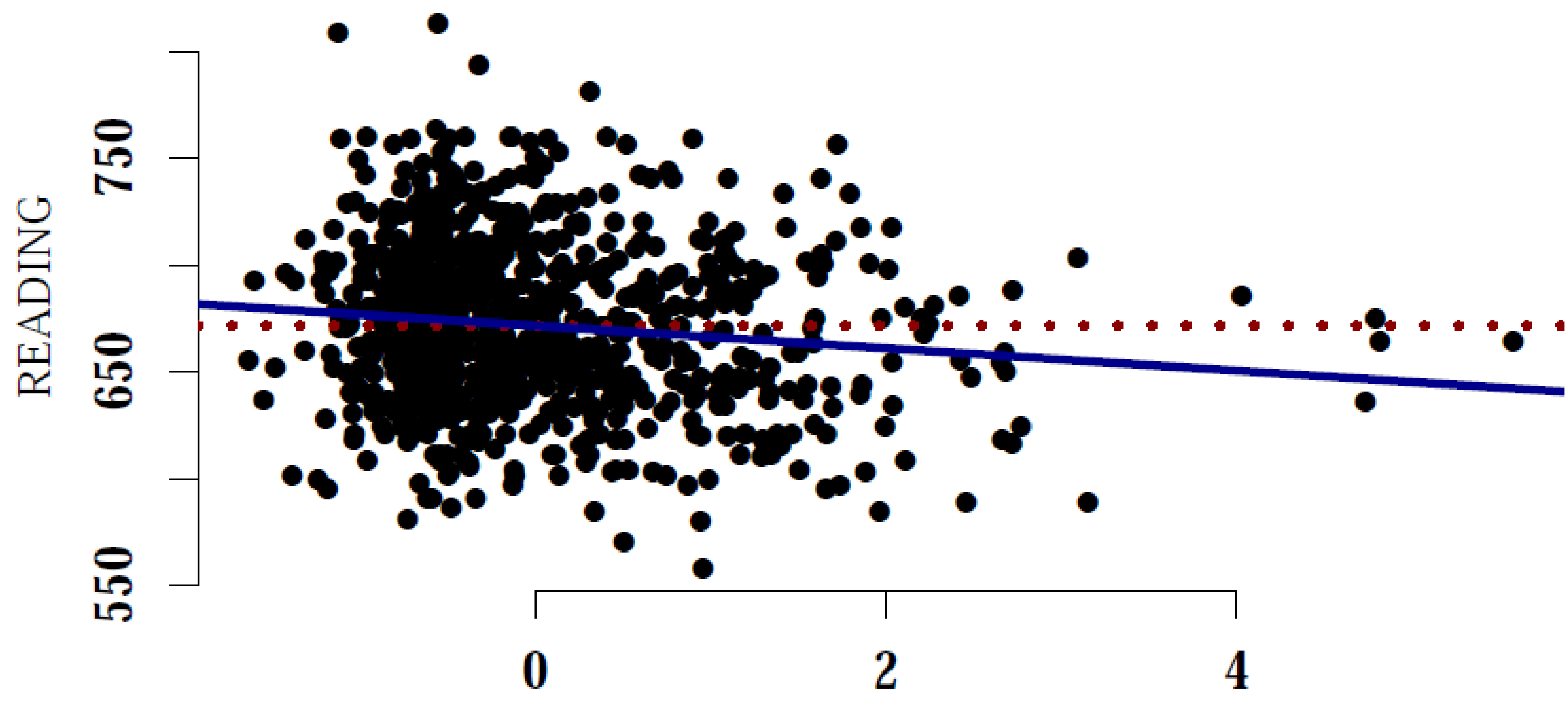
ChildrenOfImmigrants.sav [DataSet1] - SPSS Data Editor

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

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	ID	Numeric	4	0		None	None	8	Right	Scale
2	Reading	Numeric	3	0	Stanford Readi...	None	None	8	Right	Scale
3	FreeLunch	Numeric	4	1	% of Students i...	None	None	8	Right	Scale
4	Male	Numeric	1	0	Male = 1, Fem...	{0, Female}...	None	8	Right	Nominal
5	Depress	Numeric	8	5	Depression Sc...	None	None	8	Right	Scale
6	SES	Numeric	5	2	Composite Fa...	None	None	8	Right	Scale
7										

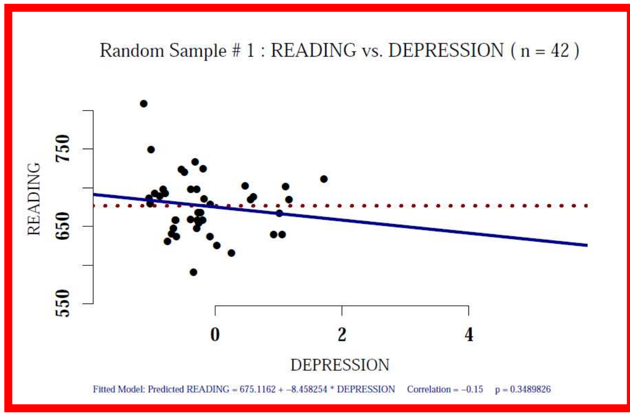
Let's *Pretend* That This Large Sample is *The* Population

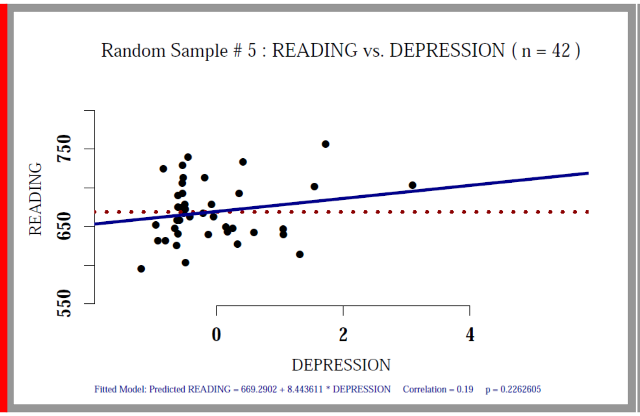
Full Sample: READING vs. DEPRESSION (n = 880)

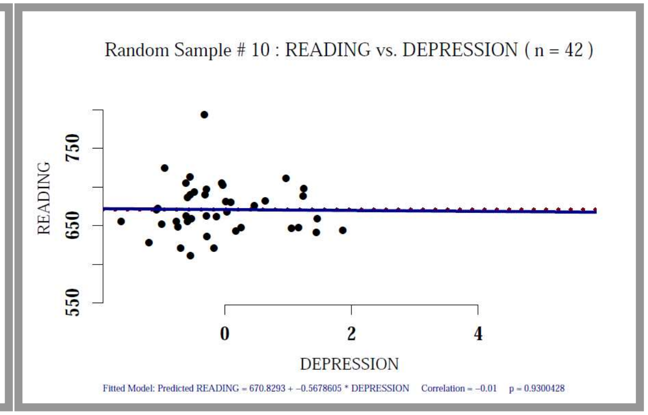


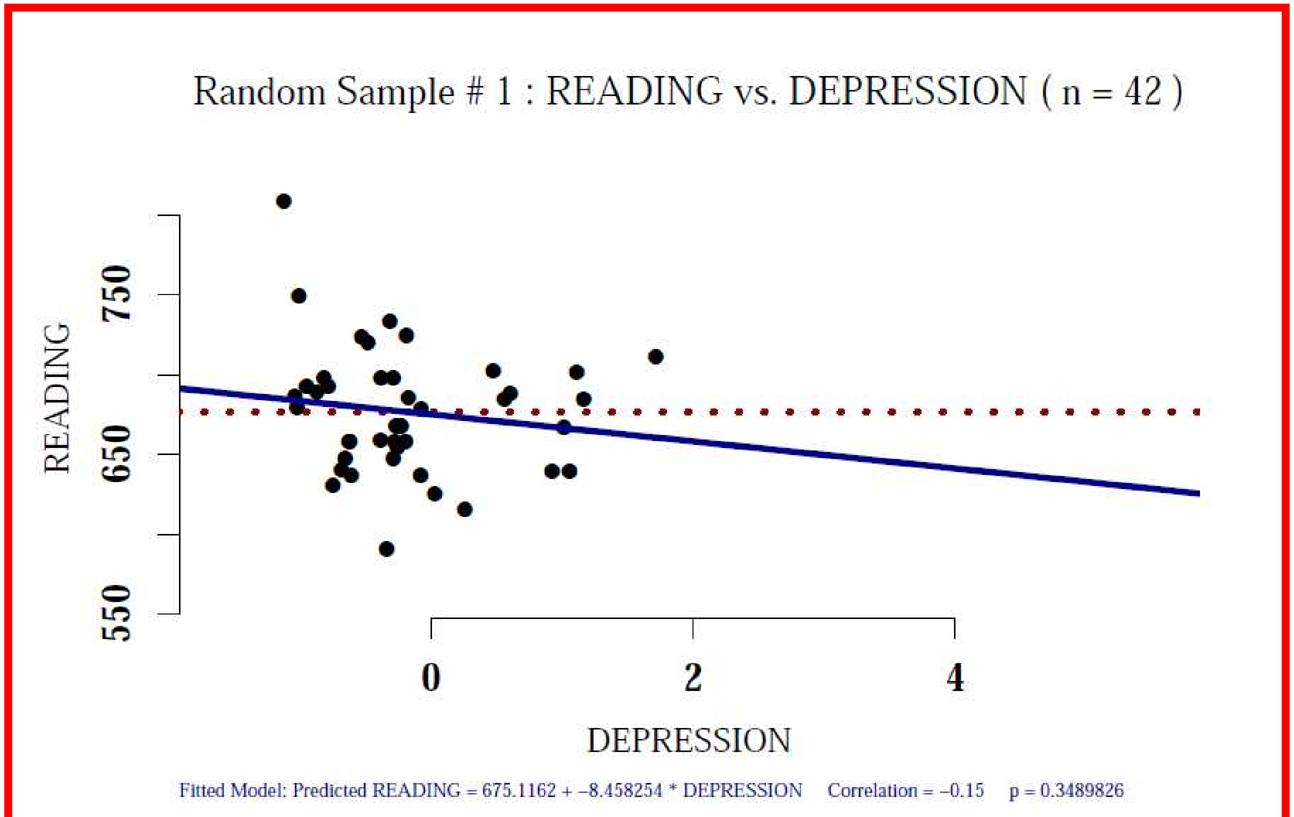
Fitted Model: Predicted READING = 671.6072 + –5.260247 * DEPRESSION   Correlation = –0.12   p = 0.0002475392

What Happens When We Take A Random Sample (n = 42)?

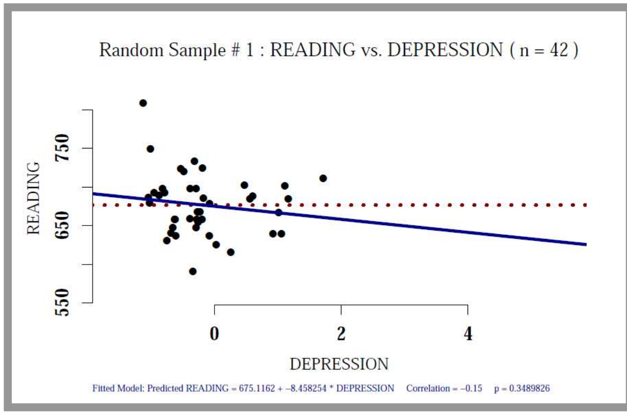


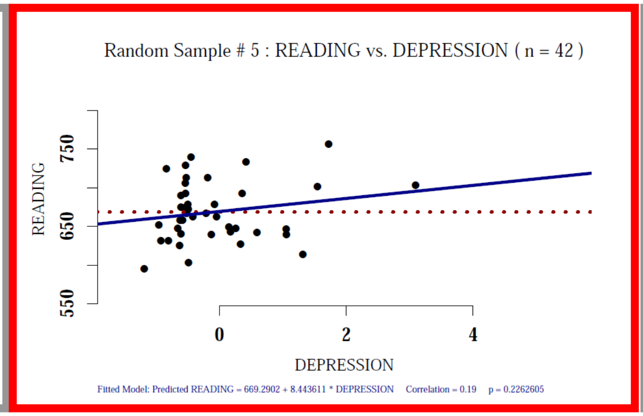


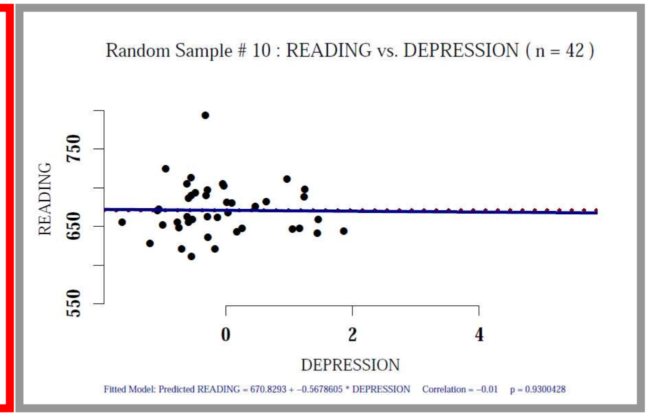


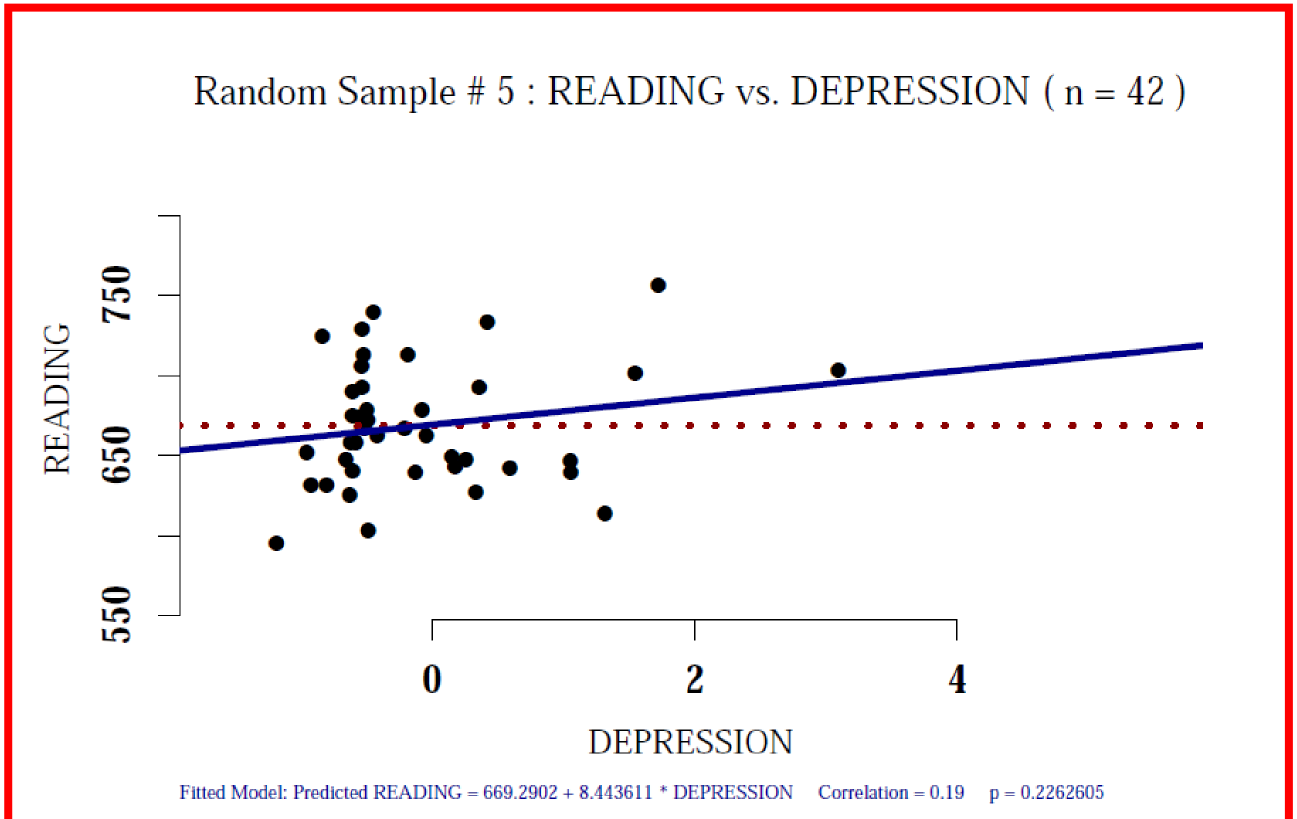


What Happens When We Take A Random Sample (n = 42)?

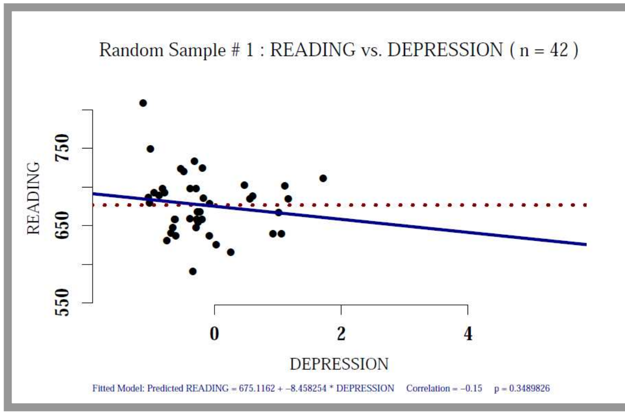


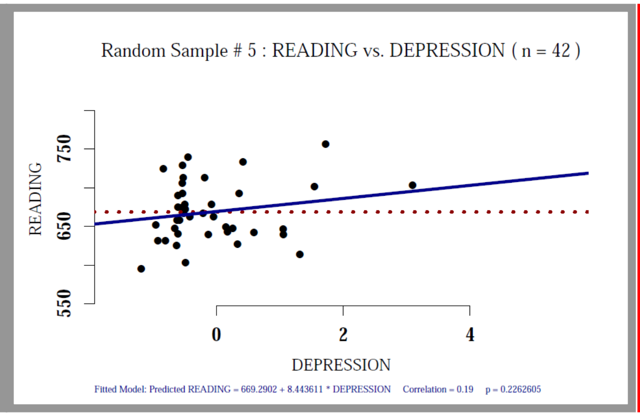


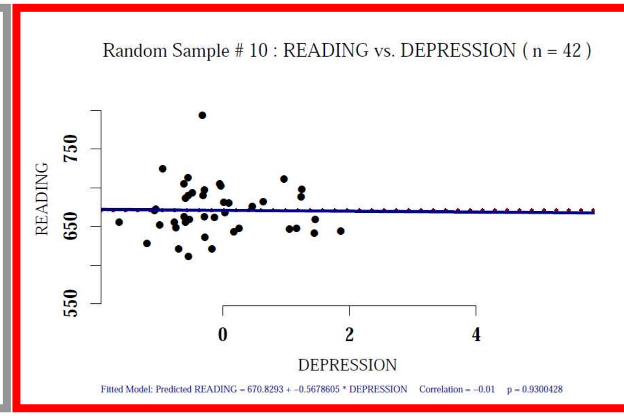


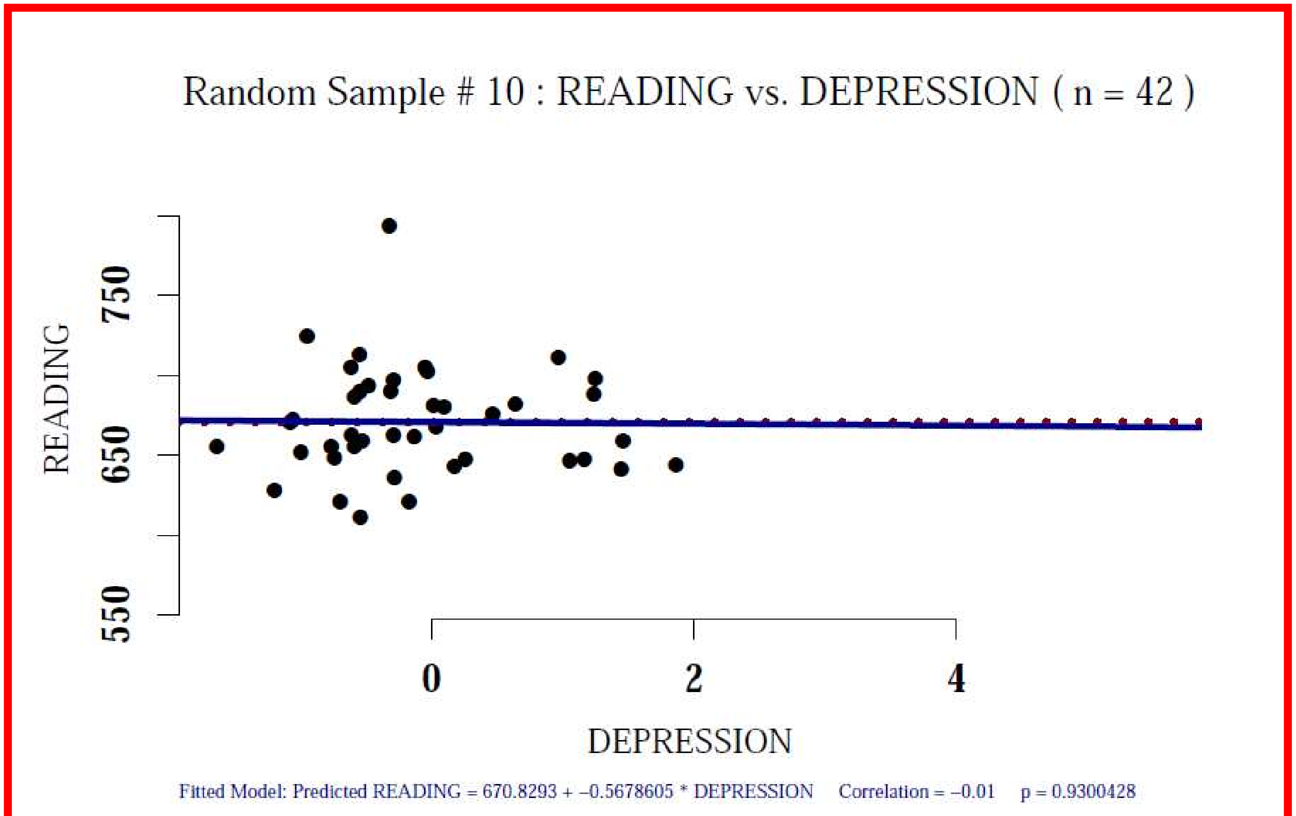


What Happens When We Take A Random Sample (n = 42)?

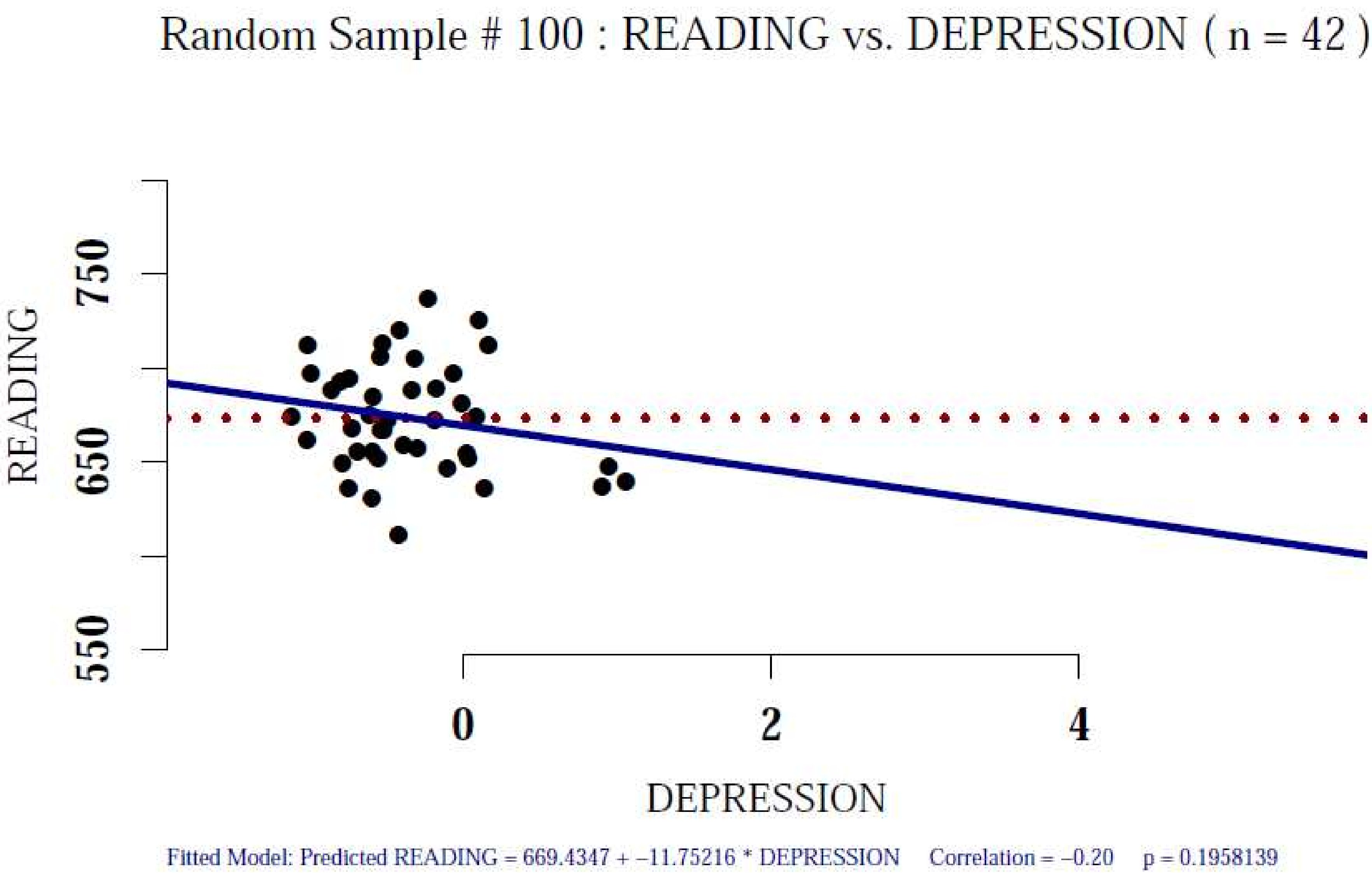




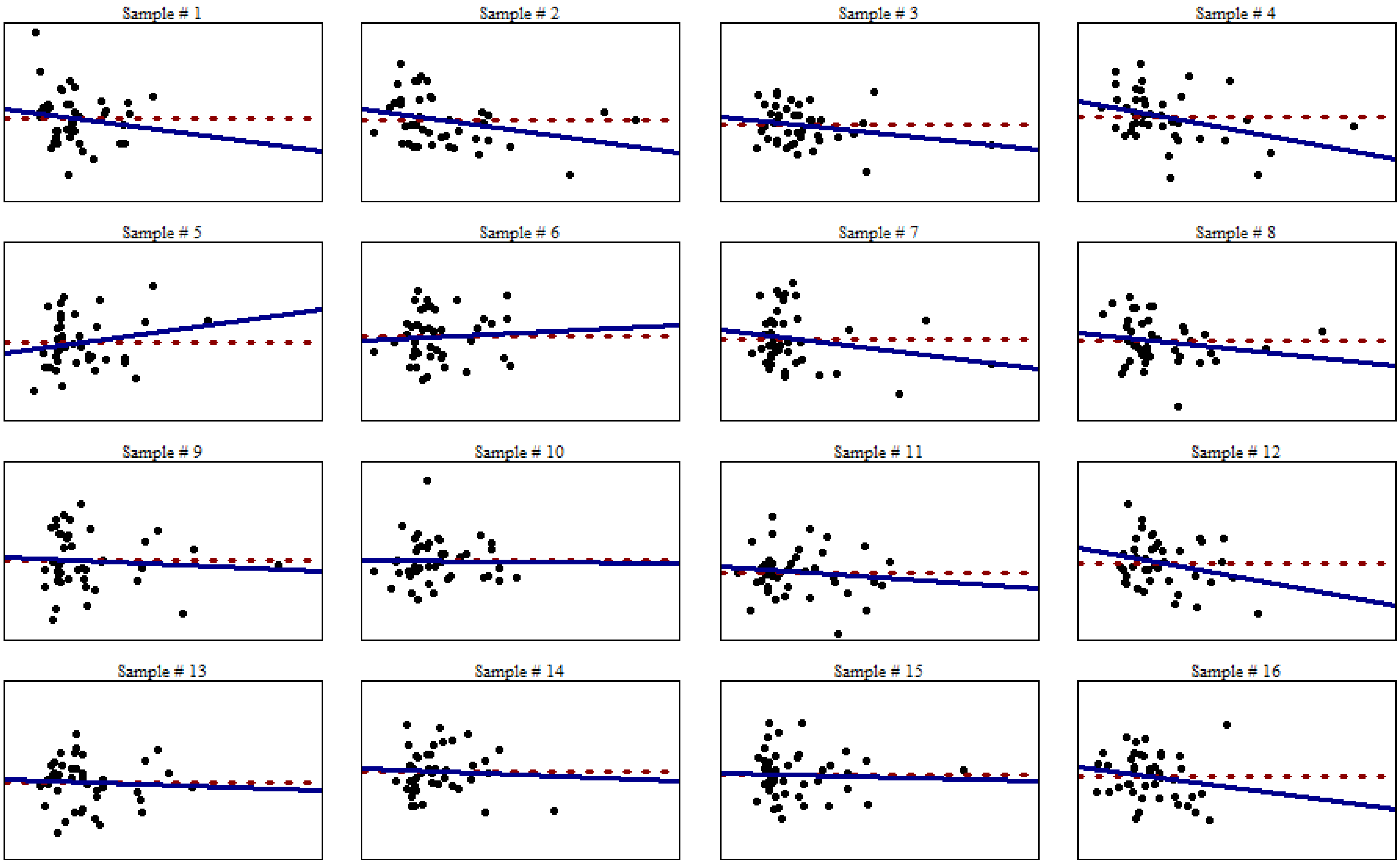




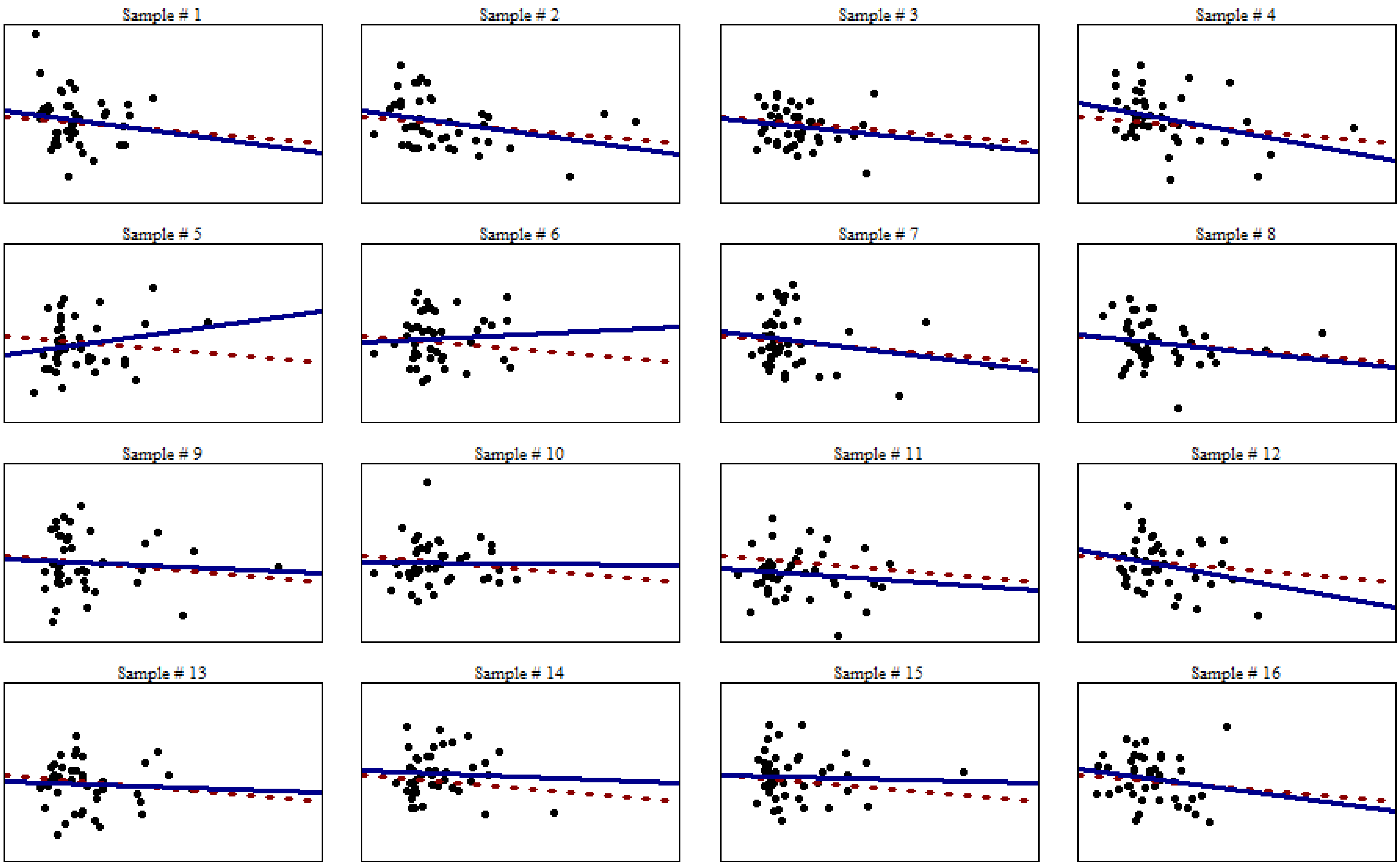
The Problem of Sampling Error: Every Sample Yields Different Results



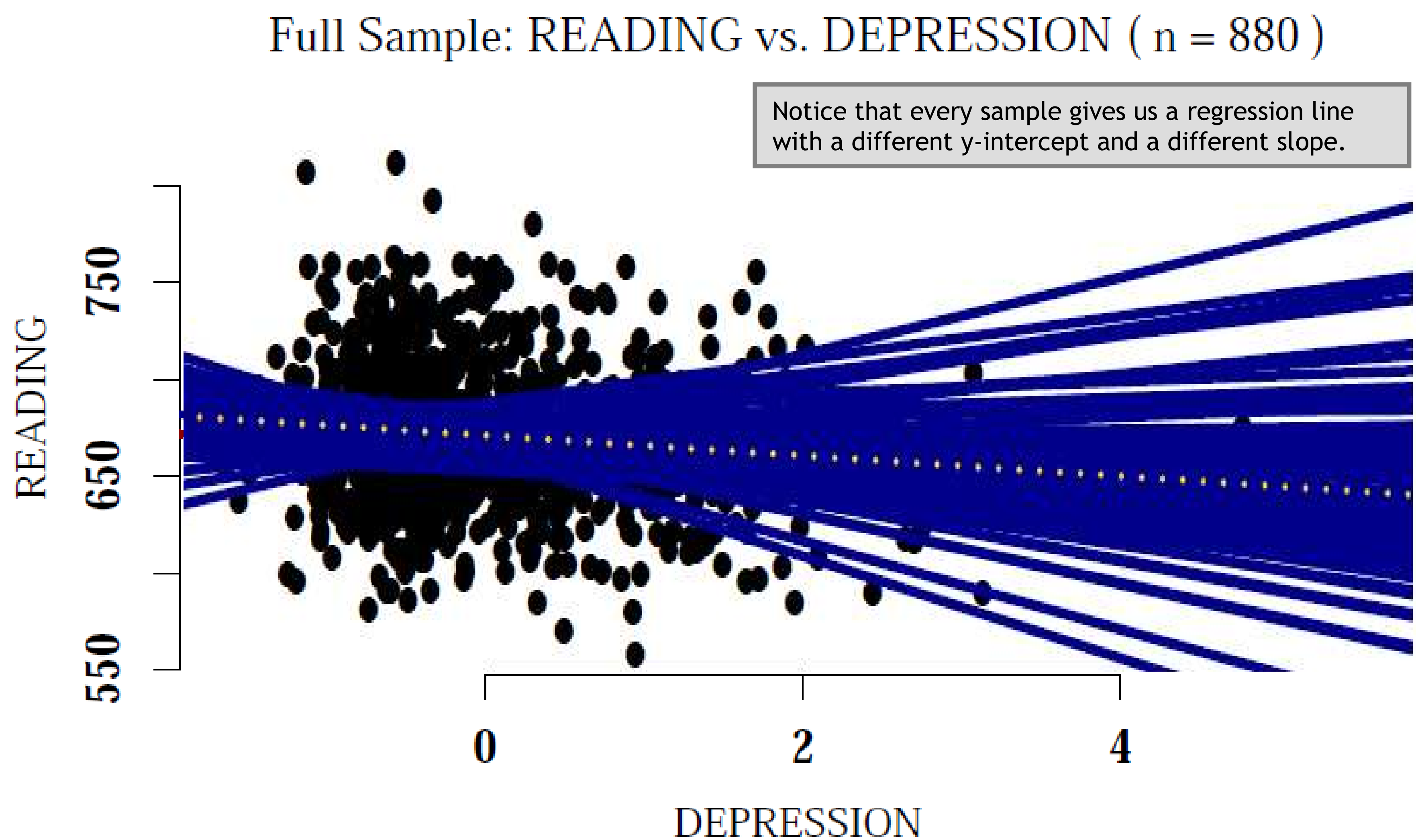
The Mean As A Reference Line (Red Dotted)



The “Population” Trend As A Reference Line (Red Dotted)



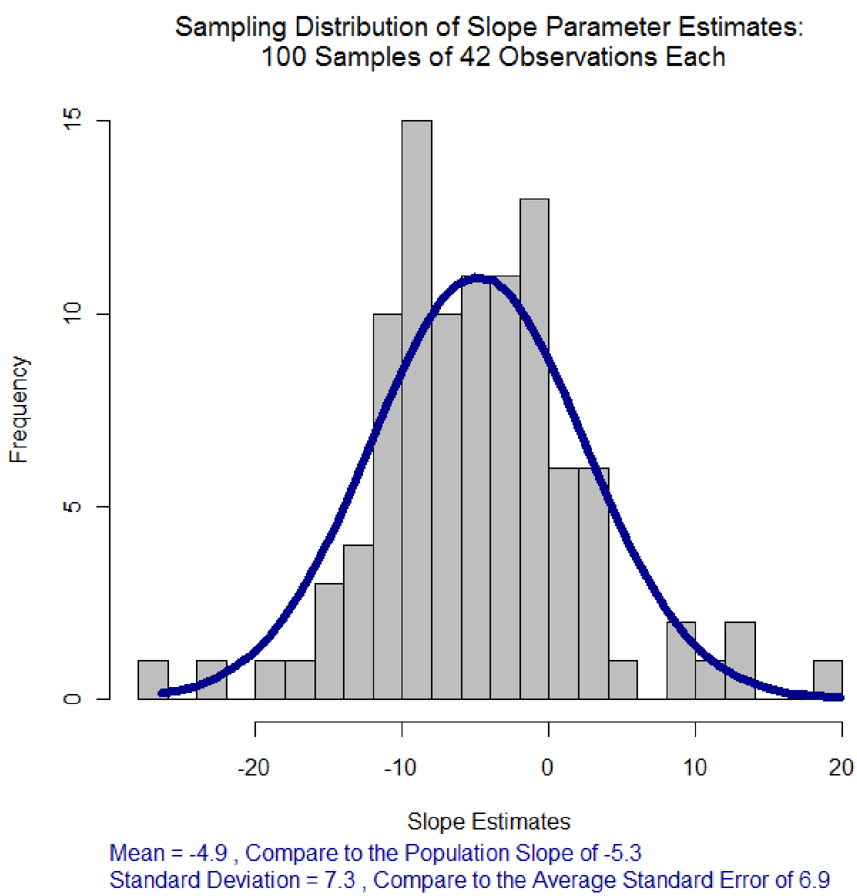
100 Regression Lines Superimposed On The “Population”

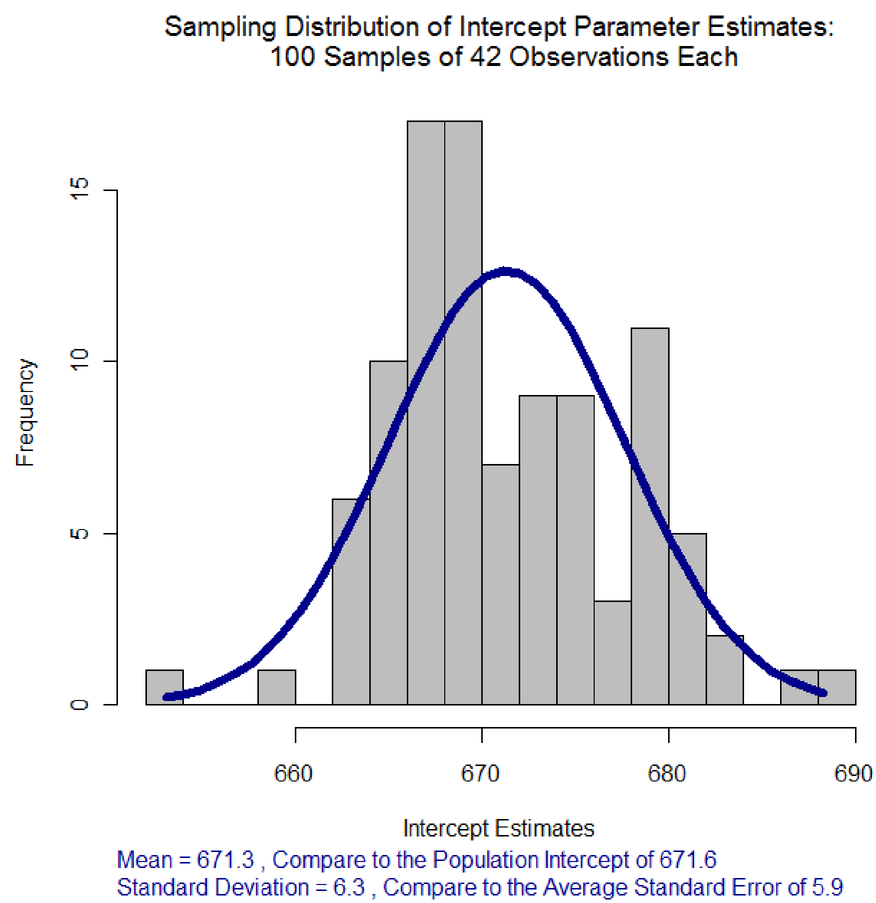


Notice that every sample gives us a regression line with a different y-intercept and a different slope.

Notice the butterfly pattern. This is not an accident. Imagine if we took the 100 slopes and made a histogram of them. Likewise, imagine if we made a histogram of the 100 y-intercepts.

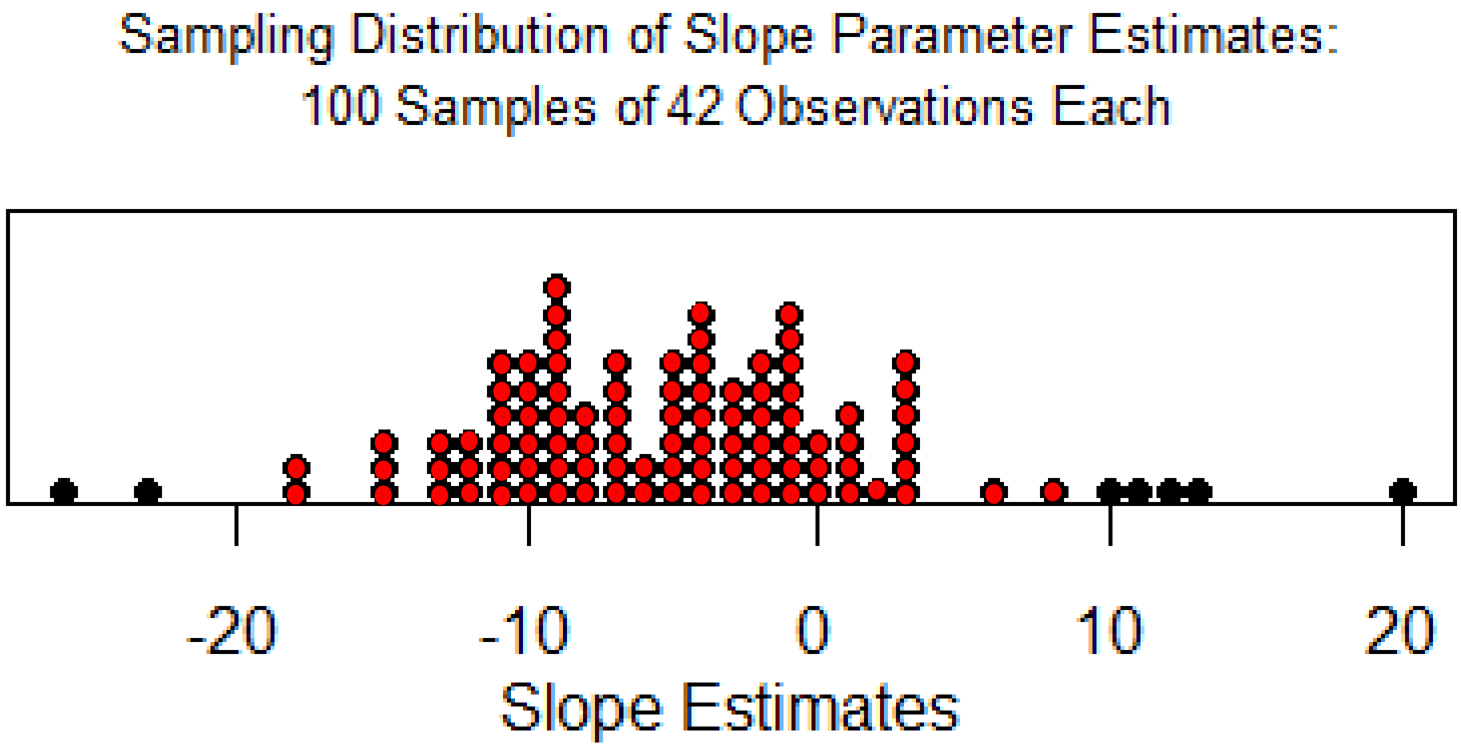
Two *Sampling Distributions*: Distributions of Sample Estimates





A <u>sampling distribution</u> is a distribution of statistics taken from many (equal sized, random) samples of the same population. Generally, in life, we get one sample. Nevertheless, it's a useful question to ask what would happen if we took many samples.

Sampling Distribution In Dot Plot Form



Each dot is a sample. Each dot is months of research planning and data collection. You only get one dot. You never get to see the other dots. You only get to see your dot. If your dot says the slope is -5, then -5 is your best guess for the population slope. If your dot says that the slope is 20, then 20 is your best guess for the population slope!

Samples as Tokens: Each sample is equally likely. If we numbered the dots 1-100 and held a lottery where you had one of the 100 tickets, you would have the same chance of winning as the other 99 ticket holders. *Be afraid. Be very afraid.*

Samples as Types: Samples with slopes around -5 are more likely than samples with slopes around 20. Samples with slopes within ±2 standard deviations of the population slope are more likely. Suppose, in addition to numbering the dots for the lottery, we painted the dots red around -5, and we painted the corresponding tickets red as well. A red ball would be more likely to be drawn than a **black** ball. *(But, if you were holding a red ticket, you would not be more likely to win than a person holding a **black** ticket!)*

Sampling Distributions in Stem-And-Leaf Form

A sample distribution is the observed distribution of a sample. A sampling distribution is a hypothetical distribution of statistics from many samples.

```
The decimal point is 1 digit(s) to the right of the |

  -2 | 6
  -2 | 3
  -1 | 88555
  -1 | 3332221111111000000
  -0 | 9999999988887777766555555
  -0 | 44444444333332222221111111100
   0 | 0111123333333
   0 | 68
   1 | 0123
   1 |
   2 | 0
```

100 Slopes
mean = -4.9

"fighting" to be the population slope

```
The decimal point is 1 digit(s) to the right of the |

  65 | 3
  65 | 9
  66 | 2333344
  66 | 55555555566777777778888888888889999999
  67 | 0000011111222233333344444
  67 | 5555666778889999999
  68 | 1122233
  68 | 88
```

100 Intercepts
mean = 671

"fighting" to be the population intercept

SPLaSH: Spread, Location and Shape

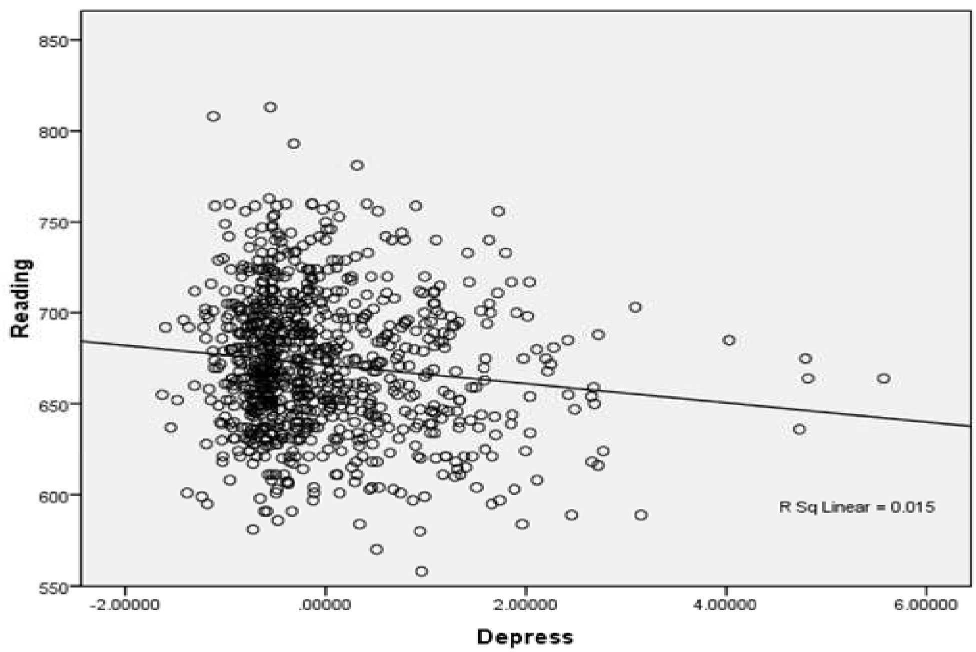
These distributions are "fighting" to be normal.

The means of these distributions are "fighting" to be the population values.

The standard deviations of these distributions are also predictably "fighting."

Examining the Full Sample (Pretending it is the Population)

Figure 6.2. A bivariate scatterplot of reading versus depression in children of immigrants (n = 880).



We are going to pretend that our large sample of 880 children of immigrants is the population. This is a pedagogical fiction; we NEVER get to observe the full population when our research question is designed to inform a theory about cause-and-effect, development or types.

Compare the mean of the sampling distribution of slopes to the "population" slope. Compare the mean of the sampling distribution of intercepts to the "population" intercept. The closeness is not an accident. Rather, the closeness is completely predictable from the Central Limit Theorem.

100 Slopes
mean = -4.9

"fighting" to be the population slope

100 Intercepts
mean = 671

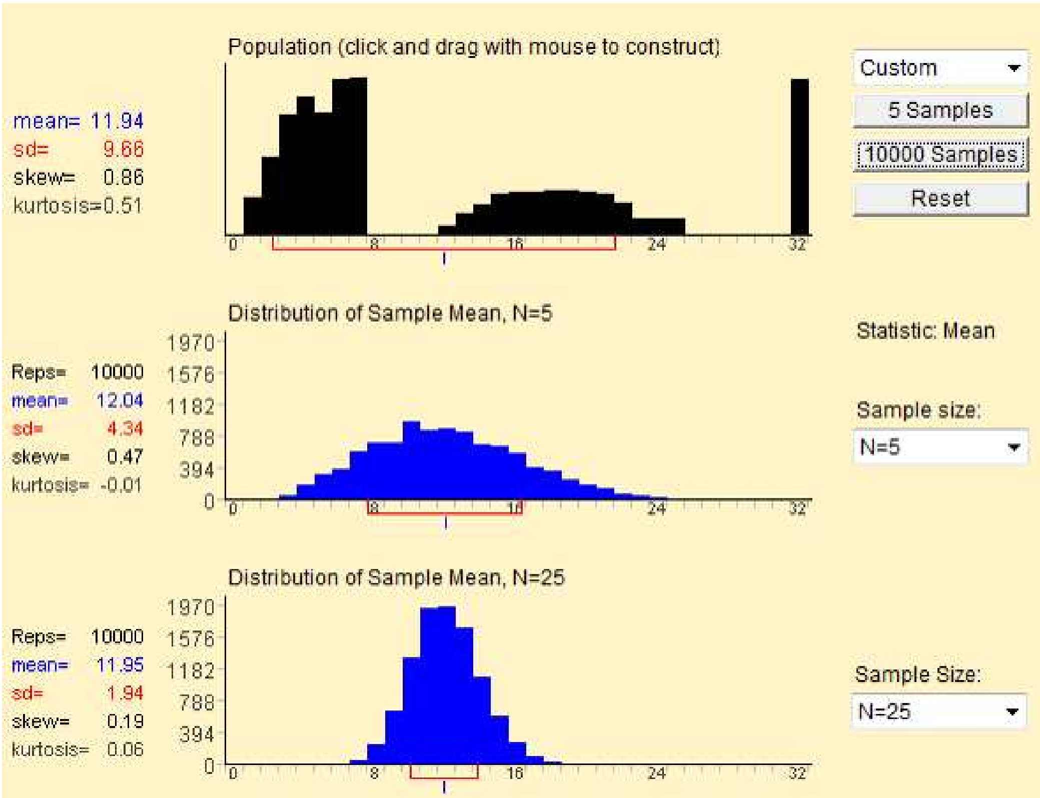
"fighting" to be the population intercept

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	671.607	1.275		526.746	.000	669.105	674.110
	Depress	-5.260	1.429	-.123	-3.680	.000	-8.066	-2.455

a. Dependent Variable: Reading

Introducing the Central Limit Theorem Through Simulations



Population Distribution

(I made it weird to illustrate a point.)

Sampling Distribution

The Means from 1000 Samples of 5 Subjects Each

Sampling Distribution

The Means from 1000 Samples of 25 Subjects Each

Notice the shapes, locations and spreads. The sampling distributions are normal. The means of the sampling distributions approach the mean of the population. The standard deviation of the sampling distributions decreases as sample size increases.

What is 9.66 divided by the square root of 5? What is 9.66 divided by the square root of 25?

http://onlinestatbook.com/simulations/sampling_dist1/sampling_dist1.html
http://onlinestatbook.com/simulations/sampling_dist_N/sampling_dist_N.html
http://onlinestatbook.com/simulations/CLT/clt.html

Central Limit Theorem

Sir Francis Galton (Natural Inheritance, 1889, from http://en.wikipedia.org/wiki/Central_limit_theorem) described the Central Limit Theorem as:

“I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error". The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.”

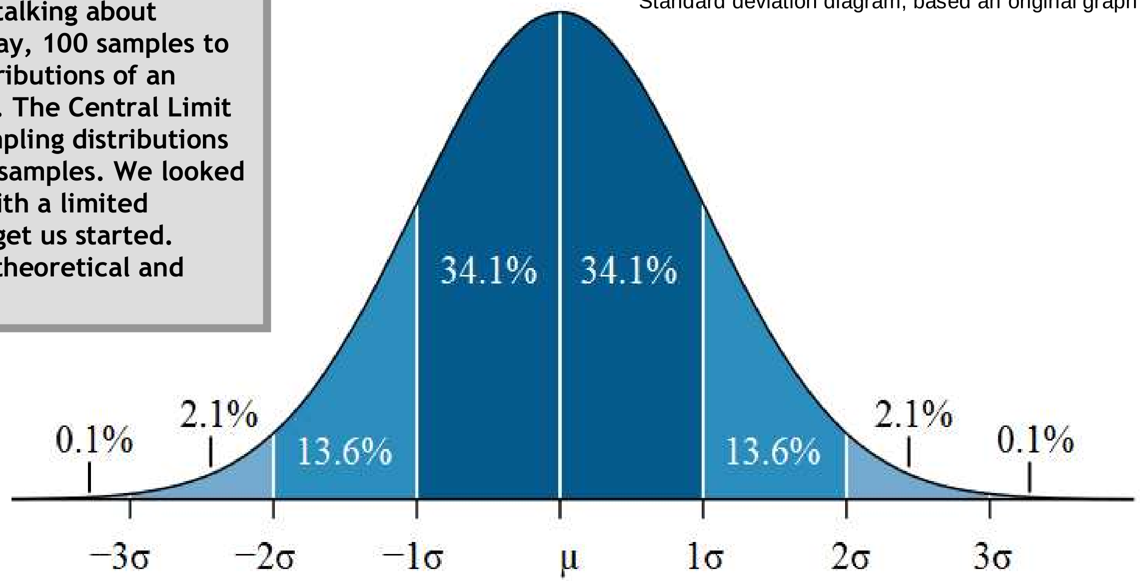
The Central Limit Theorem:
Given a population with a finite mean, μ, and a finite non-zero standard deviation, σ, a sampling distribution of the mean approaches a normal distribution with a mean of μ and a standard deviation of $\sigma/\sqrt{N}$ as N, the sample size for each sample in the sampling distribution increases.

The Upshot:
Whenever we estimate a parameter in our linear models (i.e., mean, y-intercept, or slope) from a sample, we can think of that parameter estimate as one point in a sampling distribution. We know that the sampling distribution is normal—SHAPE. We can guesstimate the standard deviation of the sampling distribution—SPREAD. Thus, only the million dollar question remains, “What is the mean of the sampling distribution”—LOCATION.

Shape and Spread

We are going to shift from talking about sampling distributions of, say, 100 samples to talking about sampling distributions of an infinite number of samples. The Central Limit Theorem tells us about sampling distributions with an infinite number of samples. We looked at sampling distributions with a limited number of samples just to get us started. Sampling distributions are theoretical and infinite.

Standard deviation diagram, based an original graph by Jeremy Kemp, in 2005-02-09



We know the shape of the sampling distributions for our means, slopes and y-intercepts. The shape would be perfectly normal were our sample size infinite. In fact, the shape is approximately normal for our finite sample sizes. Technically, our sampling distributions are not normal distributions, but t-distributions, but they are almost identical, especially when sample sizes are greater than 100. In practice we'll use the correct t-distribution (with our t-tests), but conceptually, it doesn't make a difference, so we'll focus on our old friend the normal distribution. We are very comfortable with the percentages underneath the normal distribution. Let's call the most extreme 5% of samples "weird." The weird samples constitute the 2.5% tips of the tails. Our probability of drawing a weird sample is 0.05.

Our definition of "weird" is our alpha level. Here, our alpha level is 0.05. Other common alpha levels are 0.1, 0.01, and 0.001. For this course, we'll stick with 0.05.

We can guesstimate the standard deviation of the sampling distribution using formulas based on observed standard deviations and sample size.

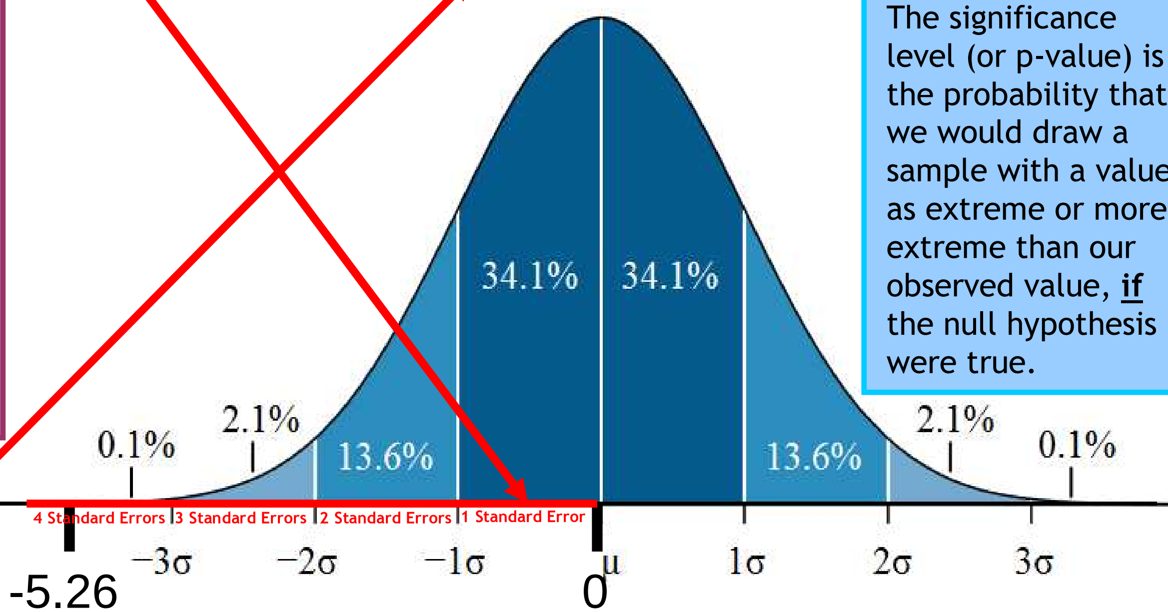
A standard deviation of a sampling distribution has a special name: standard error.

Location Location Location

The null hypothesis is a plausible worst case scenario for location. The null hypothesis (usually) states that there is no relationship in the population. We use the null hypothesis as a baseline, and we ask, "Would our sample be weird if the null hypothesis were true?" If our sample would be weird, we reject the null hypothesis and conclude that there is a relationship in the population.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	671.607	1.275		526.746	.000	669.105	674.110
	Depress	-5.260	1.429	-.123	-3.680	.000	-8.066	-2.455

If the null hypothesis were true, we would observe a slope of -5.26 in less than 0.001 of samples. Since our alpha level is 0.05 (and our p-value of 0.001 is less than our alpha level of 0.05), we reject the null hypothesis that there is no relationship between depression and reading in the population of immigrant children, and we conclude that there is a relationship in the population.



The significance level (or p-value) is the probability that we would draw a sample with a value as extreme or more extreme than our observed value, <u>if</u> the null hypothesis were true.

Our observed slope is -3.68 standard errors from zero.

Quiz: What is -5.26 divided by 1.429?

The negative correlation between depression levels and reading scores is statistically significant ($p < 0.05$).

Interpreting Significance Levels and Probability Values

- Set your alpha level ahead of time. An alpha level of 0.05 is the industry standard, but don't get me started on the topic—it's my dissertation. Your alpha level is your tolerance for false positives due to sampling error.
- State your null hypothesis. Usually your null hypothesis will be that there is no relationship in the population. Null hypotheses are always about the population.
- Note whether your significance level (i.e., sig. or probability value or p-value) is greater or less than your alpha level.
    - If your p-value is greater than your alpha level, then <u>do not reject</u> the null hypothesis; however, do not do not do not do not do not do not do not accept the null hypothesis, either. Rather, draw no conclusion.
    - If your p-value is less than your alpha level, then reject the null hypothesis and conclude that there is a relationship in the population. Do not try to pinpoint the magnitude or strength of the relationship (yet), but you may safely note the direction of the relationship.

In our sample of 880 children of immigrants, we observe a negative relationship between depression levels and reading scores. Children in our sample who differ by one unit of depression tend differ by 5.26 points on the reading achievement test. Our null hypothesis is that there is no relationship in the population of children of immigrants. However, if the null hypothesis were true, it would be very unlikely ($p < 0.05$) that we would draw such a sample. Therefore, we reject the null hypothesis and conclude that there is a negative relationship between depression and reading scores within the population of children of immigrants.

Danger, Will Robinson: If your sample is not random, none of this machinery works!

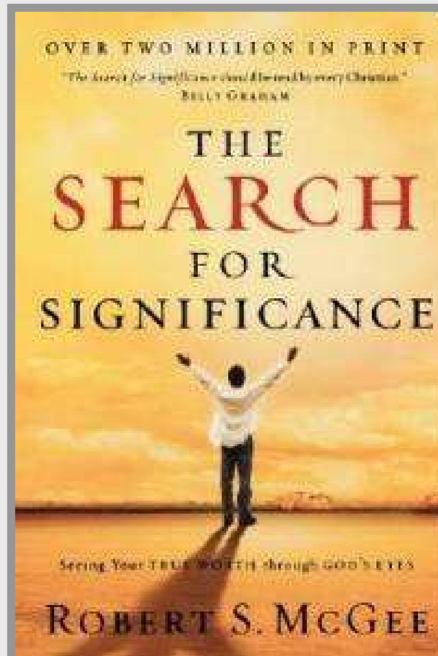
A sample statistic is <u>statistically significant</u> if, were the population statistic zero, it would be very unlikely to draw a random sample that yielded such a large (or larger) sample statistic. In any data analytic discourse, never use "significant" unless you mean "statistically significant" in which case use "statistically significant."

Reporting the Results of Simple Linear Regression

When reporting the results of simple linear regression:

1. Include your fitted model.
2. Use the R^2 statistic to convey the goodness of fit for the model (i.e., strength).
3. 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.
4. Describe the direction and magnitude of the relationship in your sample, preferably with illustrative examples. Make clear the substance of your findings.
5. Use confidence intervals to describe the magnitude of the relationship in the population
6. If simple linear regression is inappropriate for the data, then say so, briefly explain why, and forego any misleading analysis.

Keep in mind that, ***when*** the null hypothesis is true (which you'll never really now), you have a 5% chance of drawing a random sample that leads you to (mistakenly) reject the null (which is true). That's what it means to set your alpha level at .05. This is called *Type I Error*, when there is no relationship in the population, but you conclude that there is a relationship. In the parlance of dot plots of sampling distributions: when a sampling distribution is centered on zero, 5% of the time your sample will be a dot that is in the 2.5% lower extreme or the 2.5% upper extreme, so you will conclude that the sampling distribution is NOT centered on zero. It happens. In the parlance of "weirdness": We reject the null when our sample would be weird if the null were true, but 5% of the time our sample is actually weird! We are not developing infallible methods. We are developing reasonable practices that generally work over the long run. In 1,000 studies where the null is unwittingly true, we'll publish 50 (i.e., 5%), but at least we'll reject 950.



Never lose sight of the substantive meaning of the numbers.
You have what you need for the Unit 6 post hole. There is practice in back.

Dig the Post Hole

Unit 6 Post Hole:

State the null hypothesis of a test for statistical significance; reject (or not) the null hypothesis; and, draw an inference (or not) from a sample to a population.

Evidentiary materials: regression output (SPSS).

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

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			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	49.118	.115		428.169	.000
	FREELUNCH	-4.841	.198	-.267	-24.439	.000

a. Dependent Variable: READING

Tips:

- Use the three-part phrasing of the post hole to guide your three sentences.
- The null is about the population!
- Use an alpha level of .05.
- Your inference is about the population!
- **Never ever accept the null.** To accept the null is to conclude that the slope is 0.000000000… in the population. How could anybody know that?

Evidentiary materials: regression output (R).

```
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 '

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
```

Here is my answer:

The null hypothesis is that, in the population, there is no relationship between READING and FREELUNCH. Based on a p-value of less than .001, we reject the null hypothesis. We conclude that there is a relationship in the population.

Here is my answer *IF* the p-value were .34:

The null hypothesis is that, in the population, there is no relationship between READING and FREELUNCH. Based on a p-value of .34, which is greater than .05, we do not reject the null hypothesis. We cannot conclude that there is a relationship in the population.

Alien Reasoning, But We Have Some Down-Home Intuitions

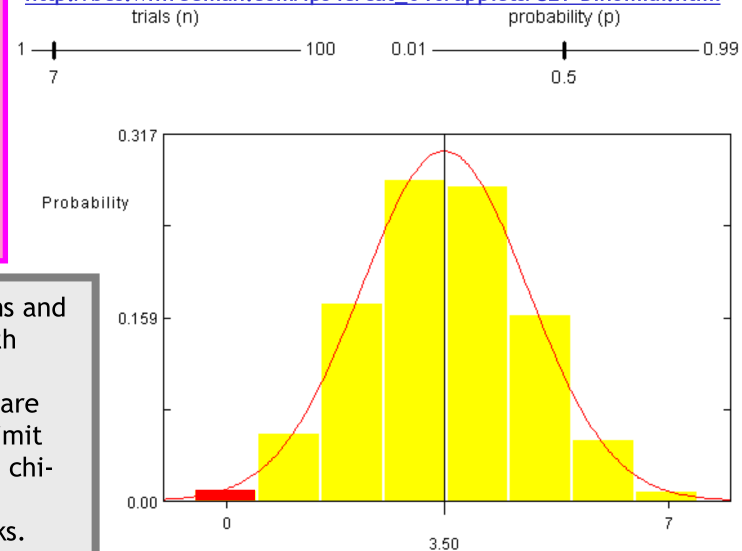
The p-value is not:

- How right your estimate is.
- How wrong your estimate is.
- How right the null is.
- How wrong the null is.

Your estimate and the null are both 100% wrong. Mathematically, each has a probabability = 0 of being the true population value.

The p-value is the probability that, if the null hypothesis were true, you would draw a sample as extreme (or more extreme) than the sample you happened to draw.

http://bcs.whfreeman.com/ips4e/cat_010/applets/CLT-Binomial.html

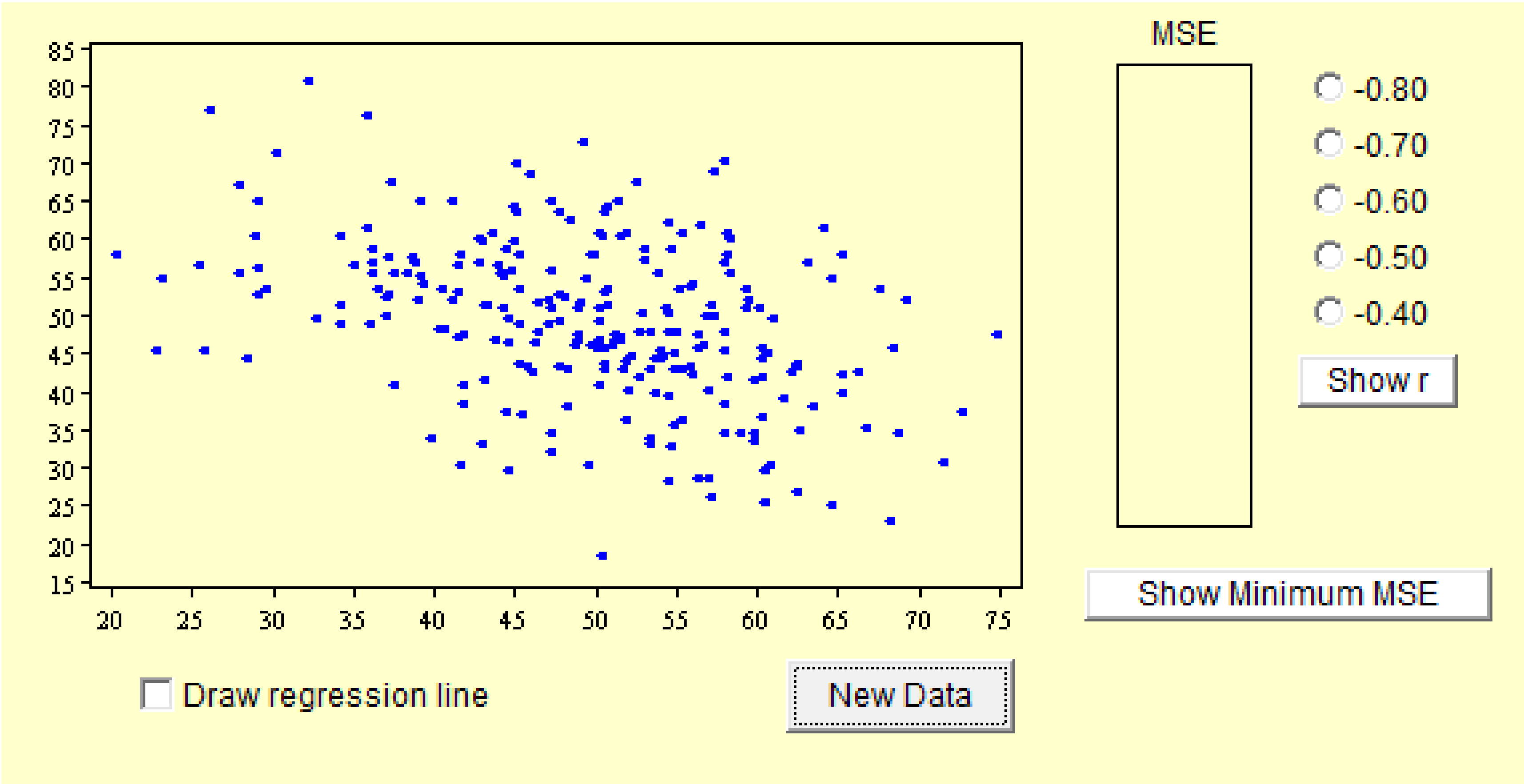


Note: sampling distributions of coin flips are different from sampling distributions of means and slopes, but they are almost the same, especially when trial/sample size is large. They both approach perfect normality as trial/sample size get large. Repeated coin flip trials form a binomial distribution, and repeated sample means and slopes form a t-distribution. There are all sorts of sampling distributions that statistician have figured out based on the Central Limit Theorem. We are learning the t-distribution, and later we will learn the F distribution and chi-square distribution. The hard part is the part we are doing right now: using sampling distributions. Once you can use one, you can use any. Here, we'll use the binomial for kicks.

When I play the coin-flip game with Uncle Jim, I start by assuming the game is fair. In other words, I start with the null hypothesis that each coin toss gives me a 50/50 chance of winning. The following probabilities assume the null hypothesis, but, at some point, I may want to reject the null hypothesis. That point is my alpha level. I want to be careful about calling Uncle Jim a "cheater" when I am just unlucky, so I set my alpha level low (perhaps at 0.05). Likewise, in research, I want to be careful about saying there's a relationship when I am just unlucky.

Probability of 1 Loss in A Row	Probability of 2 Losses in A Row	Probability of 3 Losses in A Row	Probability of 4 Losses in A Row	Probability of 5 Losses in A Row	Probability of 6 Losses in A Row	Probability of 7 Losses in A Row
p = .50	p = .25	p = .125	p = .0625	p = .03125	p = .015625	p = .0078125
LOSE	LOSE	LOSE	LOSE	LOSE	LOSE	LOSE

SURVIVE-r



http://onlinestatbook.com/stat_sim/reg_by_eye/index.html

Is Sean Smarter Than A Monkey?

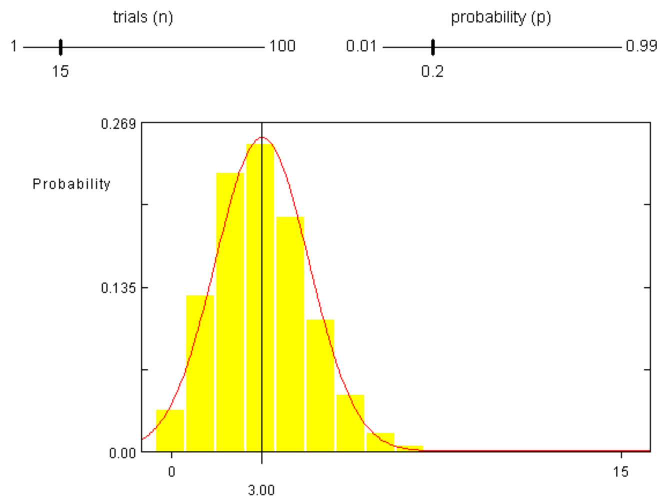
If Stats Monkey, the funky monkey, played Survive-r for 15 rounds, on average, he would get three right. In other words, if a billion monkeys played Survive-r for 15 rounds, the average monkey would get 3 right. Some monkeys, however, would guess correctly for all 15 rounds, and other monkeys would guess incorrectly for all 15 rounds. In light of this fact, how can we tell whether Sean is smarter than a monkey?

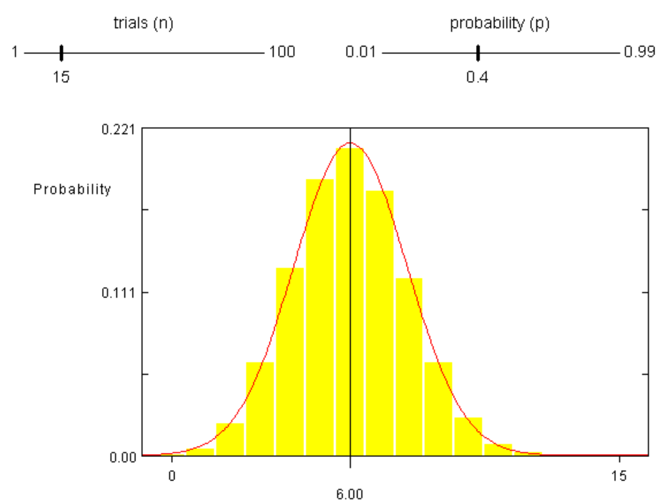
Stats Monkey always guesses B.

Why? He's a monkey! B is for "banana."

vs.

Stats Monkey		Sean Parker	
Round 1:	Right Wrong	Round 1:	Right Wrong
Round 2:	Right Wrong	Round 2:	Right Wrong
Round 3:	Right Wrong	Round 3:	Right Wrong
Round 4:	Right Wrong	Round 4:	Right Wrong
Round 5:	Right Wrong	Round 5:	Right Wrong
Round 6:	Right Wrong	Round 6:	Right Wrong
Round 7:	Right Wrong	Round 7:	Right Wrong
Round 8:	Right Wrong	Round 8:	Right Wrong
Round 9:	Right Wrong	Round 9:	Right Wrong
Round 10:	Right Wrong	Round 10:	Right Wrong
Round 11:	Right Wrong	Round 11:	Right Wrong
Round 12:	Right Wrong	Round 12:	Right Wrong
Round 13:	Right Wrong	Round 13:	Right Wrong
Round 14:	Right Wrong	Round 14:	Right Wrong
Round 15:	Right Wrong	Round 15:	Right Wrong
Total Right: ________		Total Right: ________	





Because Stats Monkey is answering the 15 questions at random (with a 1/5 or 0.20 chance for each), any one set of his answers is drawn from a binomial distribution of 15 trials with probability 0.20.

When Sean has a showdown with Stats Monkey, what makes him smarter than the monkey is not that he will necessarily get more right but that he is drawing from a binomial distribution located to the right of the monkey's binomial distribution. Let's say that, by guessing educatedly, for each question, Sean's probability of answering it correctly is 0.40.

Notice how difficult it would be to show that Sean is smarter than a monkey if we did not have the distributions. In life, we never have the sampling distributions!

Statistical Significance Testing Step-by-Step-by-Step

1. The Way Way Beginning
    1. What is your population of interest? Preschool students? Premature babies? Haitian immigrants? Public schools? Autistic children? Female athletes? Military veterans? School districts?
    2. How are you going to randomly sample from your population?
    3. What is your alpha level, your tolerance for false positives due to sampling error? Here you are anticipating Mr. Null. Both you and he know that any relationship you find in your sample is just an estimate of the population relationship. Different samples will yield different estimates. If there is no relationship in the population, you will still find a relationship (perhaps tiny) in your sample. How scared are you of saying there is a relationship in the population when there really is none? The lower you set your alpha level, the more scared you are. (Our $\alpha = .05$.)
2. Collect your data, a <u>random</u> sample, to answer your research question about the population.
3. Obtain a relevant estimate from your data. E.g., slope, correlation or R^2 statistic.
4. Mr. Null suspects your estimate is just an artifact of sampling error. Respond:
    1. State the null hypothesis. "In the population, there is no relationship between X and Y."
    2. Consider your estimate as one equally likely estimate among many possible estimates in a distribution of estimates. I.e., consider the sampling distribution.
        1. By the central limit theorem, you know the shape of the sampling distribution--normal.
        2. By the central limit theorem (and a little guesstimation) , you know the standard deviation of the sampling distribution—the standard error.
        3. You do not know the location, but Mr. Null suspects it is zero.
    3. Obtain the t-statistic by counting the number of standard errors between your estimate and Mr. Null's zero. You are (for the purposes of argument) setting the sampling distribution on zero (location =0, population mean = 0, $\mu = 0$) and asking if your sample estimate would be an outlier. (Recall that +/- 2 standard deviations from the mean is the beginning of outlier territory.)
    4. Using percentages underneath the normal distribution, calculate your p-value (i.e., how likely it would be to draw your sample estimate (or greater) from the sampling distribution if (1) the sampling distribution were really located at zero or equivalently (2) the sampling distribution mean were zero or equivalently (3) the population value that you estimated were zero.
    5. If your p-value is less than your alpha level ($\alpha = .05$.), reject the null hypothesis and conclude that there is (some) relationship in the population (probably not exactly your estimated relationship). If your p-value is equal to or greater than your alpha level, then do not reject the null (i.e., keep it open as a plausibility, even though it is probably wrong).

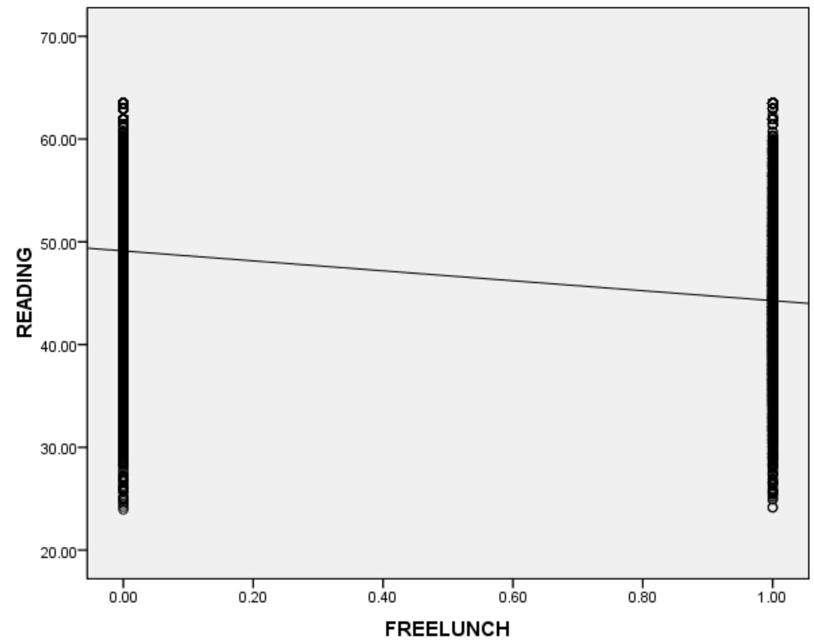
Answering our Roadmap Question

Unit 6: In the population, is there a relationship between reading achievement and free lunch?

$$Reading = \beta_0 + \beta_1 FreeLunch + \varepsilon$$

In the population of U.S. students (in 1988), students who were eligible for free/reduced lunch tended to score lower in reading. In our sample, the average difference in reading scores, 4.8, was statistically significantly different from 0 (p < 0.001).

Consider that, due to sampling error alone, we would never expect the two averages to be perfectly perfectly equal equal. We expect a difference no matter what, so we ask: Is the observed difference too large to be explained by chance alone? In this case, yes it is.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	49.118	.115		428.169	.000
	FREELUNCH	-4.841	.198	-.267	-24.439	.000

a. Dependent Variable: READING

Unit 6 Appendix: Key Concepts

- We NEVER get to observe the full population when our research question is designed to inform a theory about cause-and-effect, development or types.
- The Central Limit Theorem:

Given a population with a finite mean, μ, and a finite non-zero standard deviation, σ, the sampling distribution of the mean approaches a normal distribution with a mean of μ and a standard deviation of $\sigma/\sqrt{N}$ as N, the sample size, increases.

- Whenever we estimate a parameter in our linear models (i.e., mean, y-intercept, or slope) from a sample, we can think of that parameter estimate as one point in a sampling distribution. We know that the sampling distribution is normal—SHAPE. We can guesstimate the standard deviation of the sampling distribution—SPREAD. Thus, only the million dollar question remains, "What is the mean of the sampling distribution"—LOCATION.
- Statistical Null Hypothesis Testing in A Few Easy Steps

(1) Set your alpha level ahead of time. An alpha level of 0.05 is the industry standard, but don't get me started on the topic—it's my dissertation. Your alpha level is your tolerance for false positives due to sampling error.

(2) State your null hypothesis. Usually your null hypothesis will be that there is no relationship in the population. Null hypotheses are always about the population.

(3)Note whether your significance level (i.e., sig. or probability value or p-value) is greater or less than your alpha level.

> (3a) If your p-value is greater than your alpha level, then do not reject the null hypothesis; however, do not do not do not do not do not do not do not accept the null hypothesis, either. Rather, draw no conclusion.
>
> (3b) If your p-value is less than your alpha level, then reject the null hypothesis and conclude that there is a relationship in the population. Do not try to pinpoint the magnitude or strength of the relationship (yet), but you may safely note the direction of the relationship.

- If your sample is not random, none of this machinery works!
- The p-value is not: How right your estimate is. How wrong your estimate is. How right the null is. How wrong the null is. Rather, your estimate and the null are both 100% wrong; mathematically, each has a probabability = 0 of being the true population value. The p-value is the probability that, if the null hypothesis were true, you would draw a sample as extreme (or more extreme) than the sample you happened to draw.

Unit 6 Appendix: Key Interpretations

- In our sample of 880 children of immigrants, we observe a negative relationship between depression levels and reading scores. Children in our sample who differ by one unit of depression tend differ by 5.26 points on the reading achievement test. Our null hypothesis is that there is no relationship in the population of children of immigrants. However, if the null hypothesis were true, it would be very unlikely ($p < 0.05$) that we would draw such a sample. Therefore, we reject the null hypothesis and conclude that there is a negative relationship between depression and reading scores within the population of children of immigrants.

- If the null hypothesis were true, we would observe a slope of -5.26 in less than 0.001 of samples. Since our alpha level is 0.05 (and our p-value of 0.001 is less than our alpha level of 0.05), we reject the null hypothesis that there is no relationship between depression and reading in the population of immigrant children, and we conclude that there is a relationship in the population.

- The negative correlation between depression levels and reading scores is statistically significant ($p < 0.05$).

- In the population of U.S. students (in 1988), students who were eligible for free/reduced lunch tended to score lower in reading. In our sample, the average difference in reading scores, 4.8, was statistically significantly different from 0 ($p < 0.001$).

Unit 6 Appendix: Key Terminology

- A sample statistic is statistically significant if, were the population statistic zero, it would be very unlikely to draw a random sample that yielded such a large (or larger) sample statistic. In any data analytic discourse, never use "significant" unless you mean "statistically significant" in which case use "statistically significant."
- In any data analytic discourse, never use "significant" unless you mean "statistically significant" in which case use "statistically significant."
- A sampling distribution is a distribution of statistics taken from many (equal sized, random) samples of the same population. Generally, in life, we get one sample. Nevertheless, it's a useful question to ask what would happen if we took many samples.
- A standard deviation of a sampling distribution has a special name: standard error.
- The null hypothesis is a plausible worst case scenario for the location of our sampling distribution. The null hypothesis (usually) states that there is no relationship in the population. We use the null hypothesis as a baseline, and we ask, "Would our sample be weird if the null hypothesis were true?" If our sample would be "weird," we reject the null hypothesis and conclude that there is a relationship in the population.
- Our definition of "weird" is our alpha level. Here, our alpha level is 0.05. Other common alpha levels are 0.1, 0.01, and 0.001.
- The significance level (or p-value) is the probability that we would draw a sample with a value as extreme or more extreme than our observed value, **if** the null hypothesis were true. We reject the null hypothesis when our p-value is less than our alpha level.

Unit 6 Appendix: Math ("Standard Error" From Wikipedia)

The **standard error of the mean (SEM) is the standard deviation of the** sample mean estimate of a population mean. (It can also be viewed as the standard deviation of the error in the sample mean relative to the true mean, since the sample mean is an unbiased estimator.) SEM is usually estimated by the sample estimate of the population standard deviation (sample standard deviation) divided by the square root of the sample size (assuming statistical independence of the values in the sample):

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where

s is the sample standard deviation (i.e., the sample based estimate of the standard deviation of the population), and

n is the size (number of observations) of the sample.

This estimate may be compared with the formula for the true standard deviation of the mean:

$$SD_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where

σ is the standard deviation of the population.

Note 1: Standard error may also be defined as the standard deviation of the residual error term.[2][3]

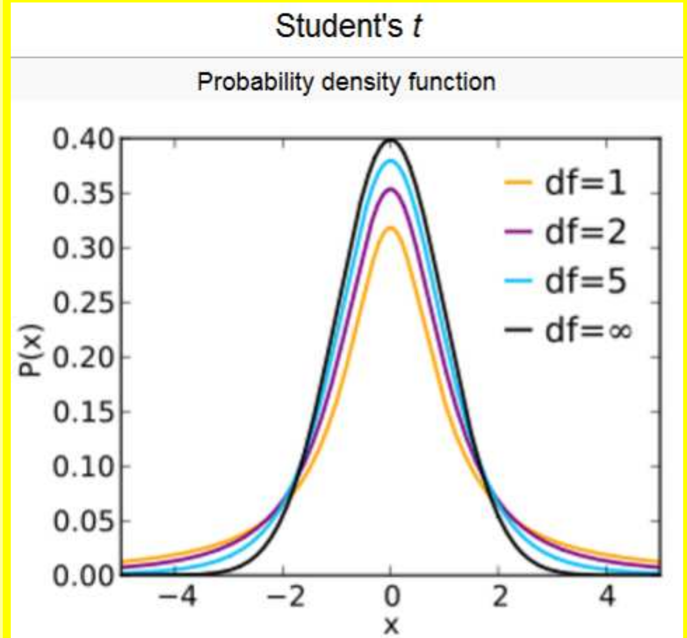
A practical result: Decreasing the uncertainty in your mean value estimate by a factor of two requires that you acquire four times as many observations in your sample. Worse, decreasing standard error by a factor of ten requires a hundred times as many observations.

Sean's notes: This is the standard error for the mean, the most basic of standard errors. In this course, we will be working with the standard error of the slope or the correlation, which we'll look at in more mathematical detail in Unit 7. The key here is that all standard errors have the same structure:

$$\text{Standard Error} = \frac{\text{Guestimation of Population Variation}}{\sqrt{\text{Basically Sample Size (aka Degrees of Freedom)}}} = \frac{\text{Want Small}}{\text{Want Big}} = \text{Want Small}$$

Unit 6 Appendix: Math ("t-distribution" From Wikipedia)

In probability and statistics, **Student's *t*-distribution (or simply the *t*-distribution) is a continuous** probability distribution that arises in the problem of estimating the mean of a normally distributed population when the sample size is small. It is the basis of the popular Student's *t*-tests for the statistical significance of the difference between two sample means, and for confidence intervals for the difference between two population means. The *Student's t-distribution* is a special case of the generalised hyperbolic distribution. The derivation of the *t*-distribution was first published in 1908 by William Sealy Gosset while he worked at a Guinness Brewery in Dublin. Due to proprietary issues, the paper was written under the pseudonym *Student*. The *t*-test and the associated theory became well-known through the work of R.A. Fisher, who called the distribution "Student's distribution".



Sean's notes: "df" stands for degrees of freedom, basically sample size minus a smidge (-1 or -2, in this course). The black line is perfectly normal; even small sample sizes get very close.

Student's distribution arises when (as in nearly all practical statistical work) the population standard deviation is unknown and has to be estimated from the data. Quite often, however, textbook problems will treat the population standard deviation as if it were known and thereby avoid the need to use the Student's *t*-test. These problems are generally of two kinds: (1) those in which the sample size is so large that one may treat a data-based estimate of the variance as if it were certain, and (2) those that illustrate mathematical reasoning, in which the problem of estimating the standard deviation is temporarily ignored because that is not the point that the author or instructor is then explaining.

Unit 6 Appendix: SPSS Syntax

No New Programming For This Unit

```
*You can use my code by switching out my variables (circled) with your variables.
*You can make a comment by starting with an asterisk and ending with a period.
*SPSS will ignore anything between the asterisk and period.
*SPSS loves/needs to end chunks of command with a period, so if something is
acting funky, make sure that your periods are in order.
********************************************************************************.
*I'm going to create a scatterplot with PERCENT on the x-axis and
SAT on y-axis; the only thing that you can't decipher is the
"/MISSING=LISTWISE" line, but all this does is tell SPSS to ignore
anybody with missing data for the variables at play in this
chunk of code.
********************************************************************************.
GRAPH
  /SCATTERPLOT(BIVAR)=PERCENT WITH SAT
  /MISSING=LISTWISE.
********************************************************************************.
*I'm going to linearly regress SAT on PERCENT.
*NOTE THAT IT IS STUPID TO TAKE SERIOUSLY THE RESULTS SINCE THE RELATIONSHIP IS NONLINEAR.
*Notice our now familiar friend "LISTWISE".
*Notice that, against proper English, I put the last period outside the quotation marks!.
*I didn't want SPSS to "see" a dangling quotation mark and wonder what to do.
*Notice the last two lines; you should be able to decipher a little.
*Ignore the rest for now.
********************************************************************************.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT SAT
  /METHOD=ENTER PERCENT.
```

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)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.731[a]	.534	.526	.80682

a. Predictors: (Constant), Self-disclose to boyfriend

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	43.280	1	43.280	66.487	.000[a]
	Residual	37.756	58	.651		
	Total	81.037	59			

a. Predictors: (Constant), Self-disclose to boyfriend

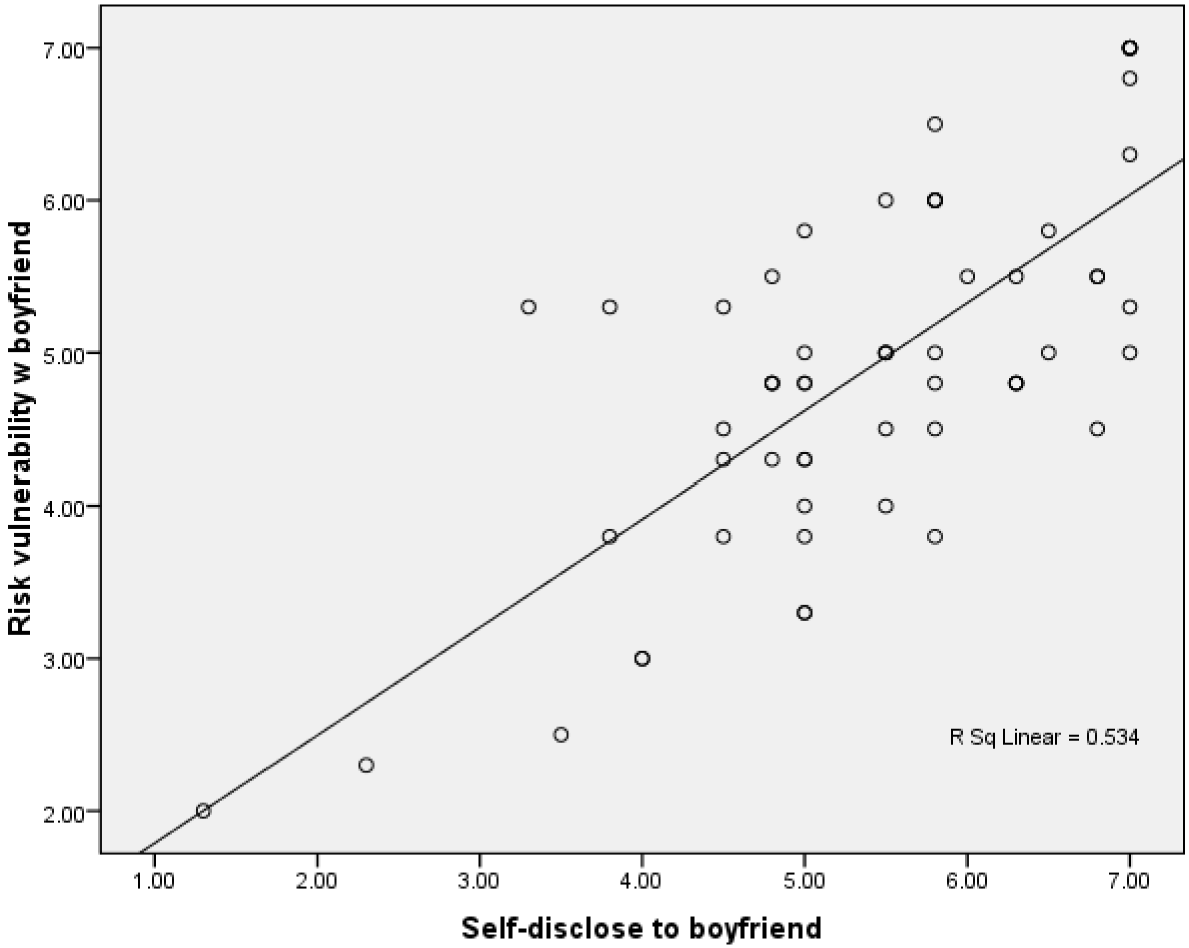
b. Dependent Variable: Risk vulnerability w boyfriend

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.081	.482		2.244	.029
	Self-disclose to boyfriend	.708	.087	.731	8.154	.000

a. Dependent Variable: Risk vulnerability w boyfriend

Perceived Intimacy of Adolescent Girls (Intimacy.sav)



Perceived Intimacy of Adolescent Girls (Intimacy.sav)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.002[a]	.000	-.017	1.19785

a. Predictors: (Constant), Self-disclose to mother

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.000	1	.000	.000	.985[a]
	Residual	83.221	58	1.435		
	Total	83.222	59			

a. Predictors: (Constant), Self-disclose to mother

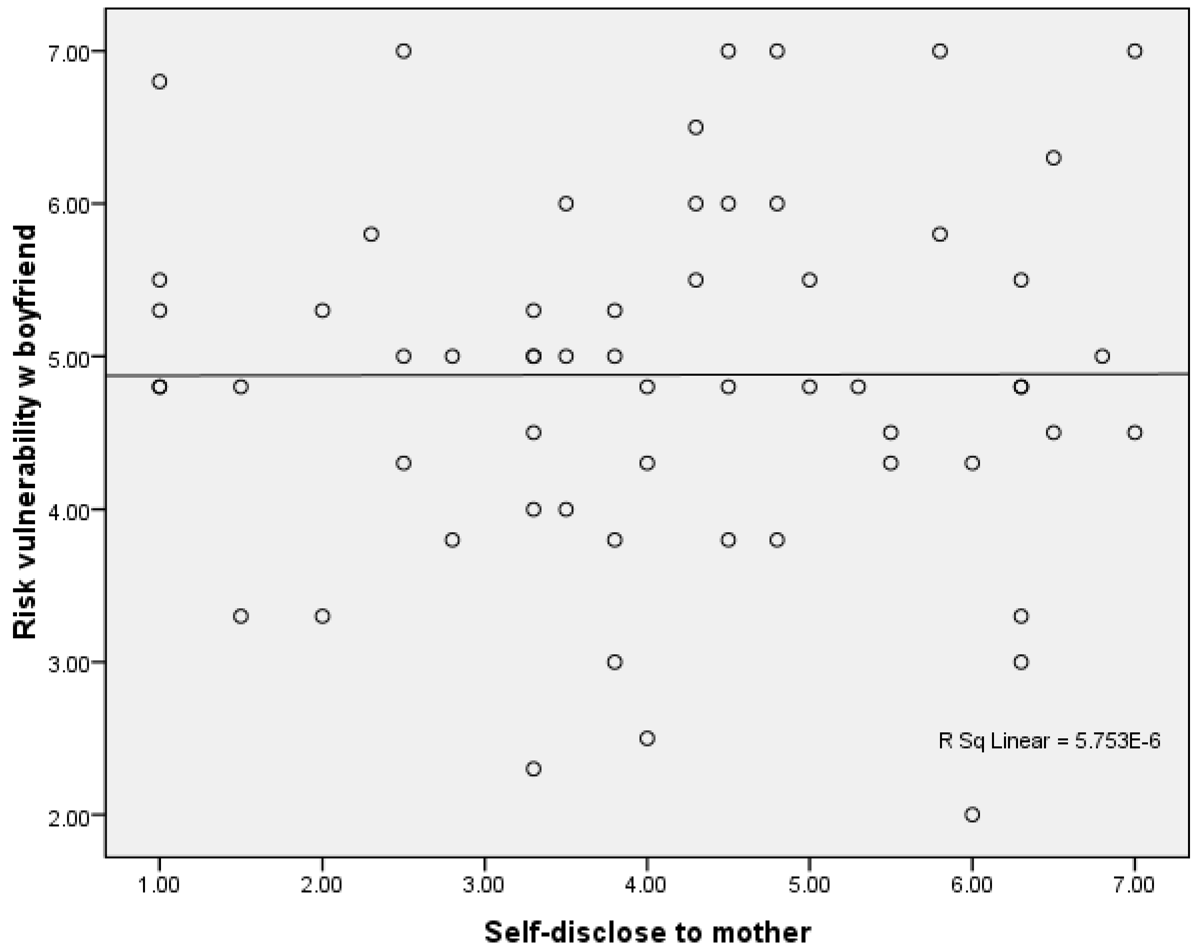
b. Dependent Variable: Risk vulnerability w boyfriend

Coefficients[a]

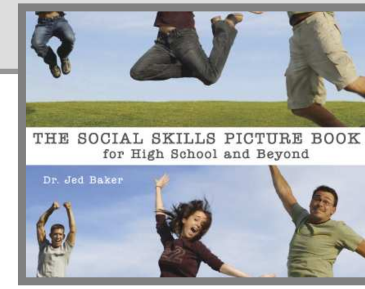
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.872	.404		12.050	.000
	Self-disclose to mother	.002	.091	.002	.018	.985

a. Dependent Variable: Risk vulnerability w boyfriend

Perceived Intimacy of Adolescent Girls (Intimacy.sav)



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:**

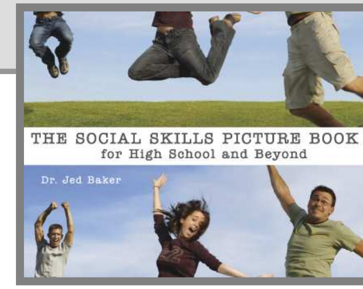
Variables about the student—

(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)



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.440[a]	.193	.192	7.71738

a. Predictors: (Constant), Base Year SES

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	14858.061	1	14858.061	249.473	.000[a]
	Residual	62059.321	1042	59.558		
	Total	76917.382	1043			

a. Predictors: (Constant), Base Year SES

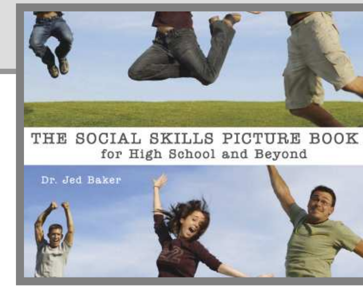
b. Dependent Variable: Base Year Composite Test

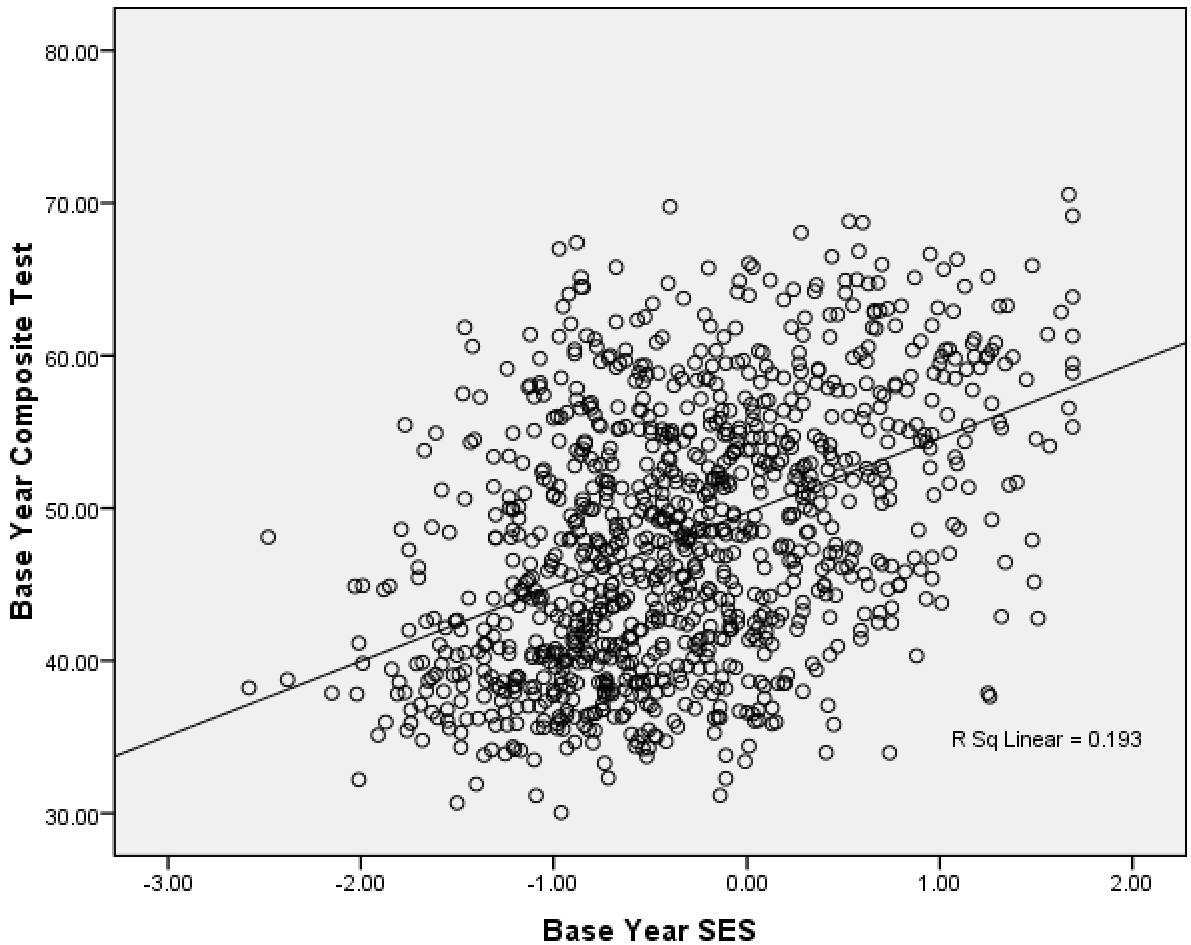
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	49.726	.260		191.448	.000	49.216	50.235
	Base Year SES	4.879	.309	.440	15.795	.000	4.273	5.485

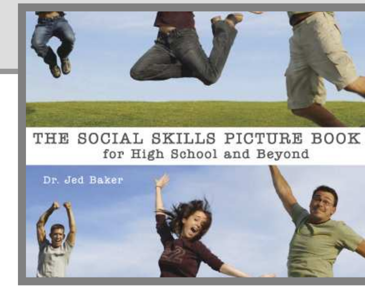
a. Dependent Variable: Base Year Composite Test

High School and Beyond (HSB.sav)





High School and Beyond (HSB.sav)



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.429[a]	.184	.184	7.75965

a. Predictors: (Constant), BY SES, School Avg

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	14176.284	1	14176.284	235.439	.000[a]
	Residual	62741.098	1042	60.212		
	Total	76917.382	1043			

a. Predictors: (Constant), BY SES, School Avg

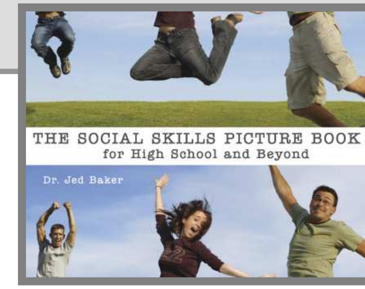
b. Dependent Variable: Base Year Composite Test

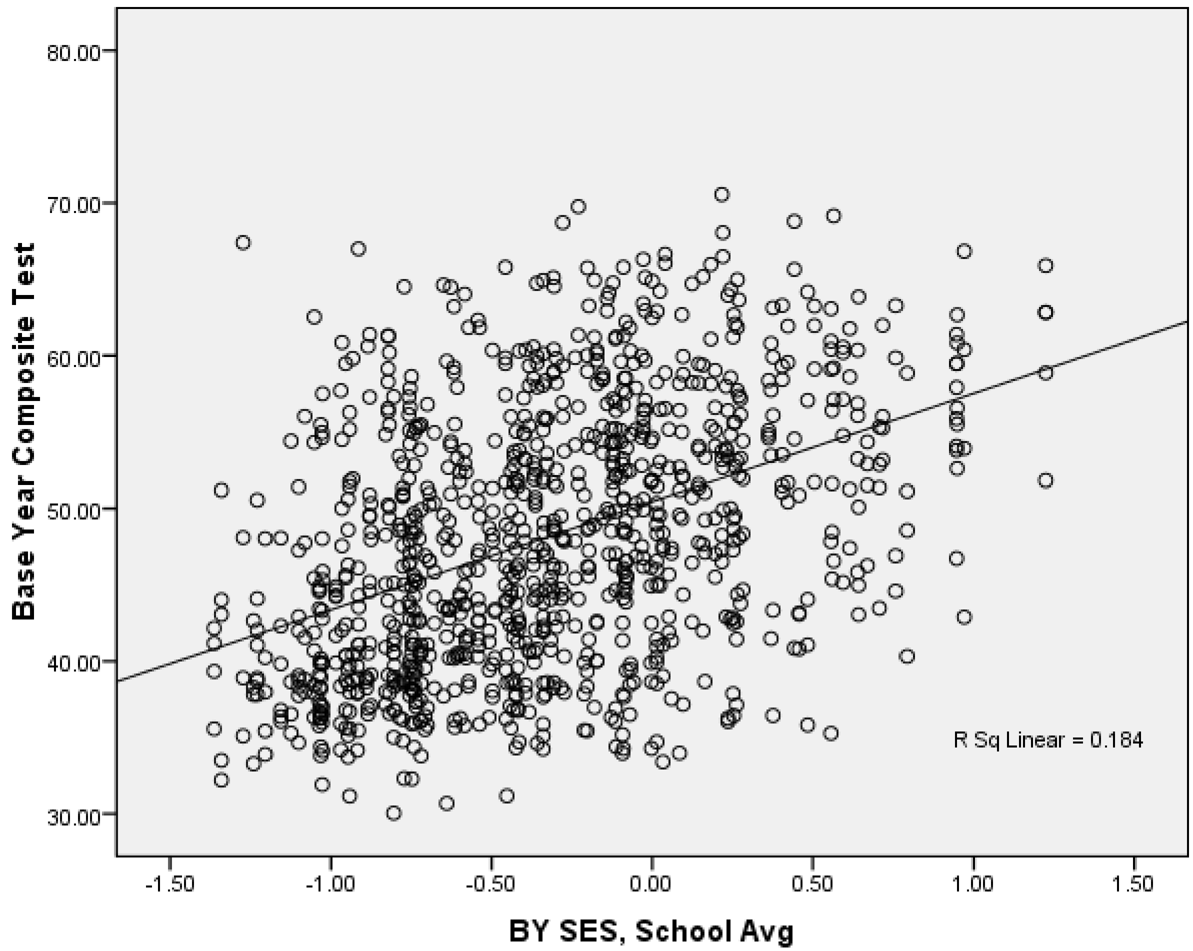
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	50.451	.284		177.397	.000	49.893	51.009
	BY SES, School Avg	7.075	.461	.429	15.344	.000	6.171	7.980

a. Dependent Variable: Base Year Composite Test

High School and Beyond (HSB.sav)





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)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.824[a]	.679	.678	.58181

a. Predictors: (Constant), General Reasoning

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	136.226	1	136.226	402.433	.000[a]
	Residual	64.316	190	.339		
	Total	200.542	191			

a. Predictors: (Constant), General Reasoning

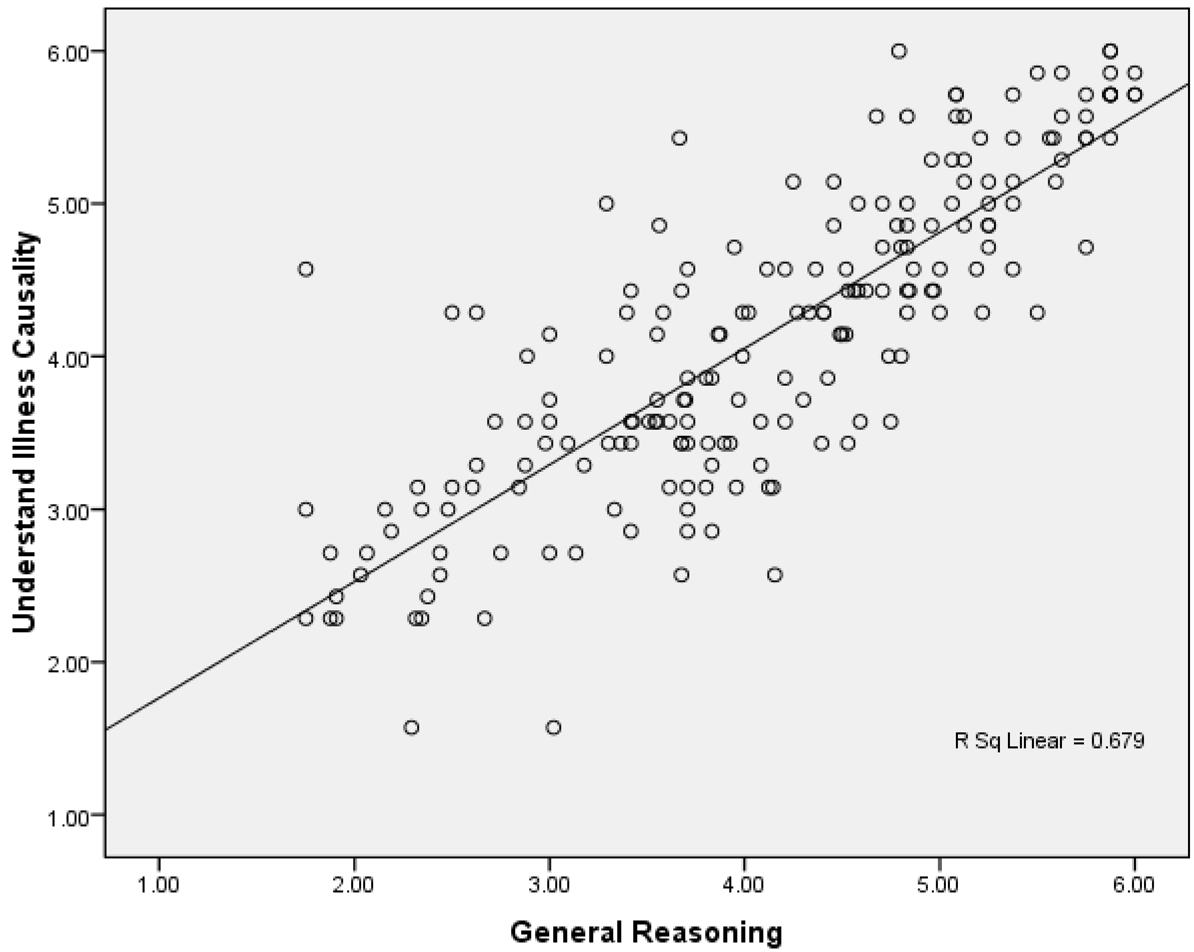
b. Dependent Variable: Understand Illness Causality

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	1.004	.162		6.204	.000	.685	1.323
	General Reasoning	.762	.038	.824	20.061	.000	.687	.837

a. Dependent Variable: Understand Illness Causality

Understanding Causes of Illness (ILLCAUSE.sav)



Understanding Causes of Illness (ILLCAUSE.sav)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.440[a]	.194	.189	.94848

a. Predictors: (Constant), 1 = Asthmatic, 0 = Healthy

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	34.383	1	34.383	38.219	.000[a]
	Residual	143.040	159	.900		
	Total	177.423	160			

a. Predictors: (Constant), 1 = Asthmatic, 0 = Healthy

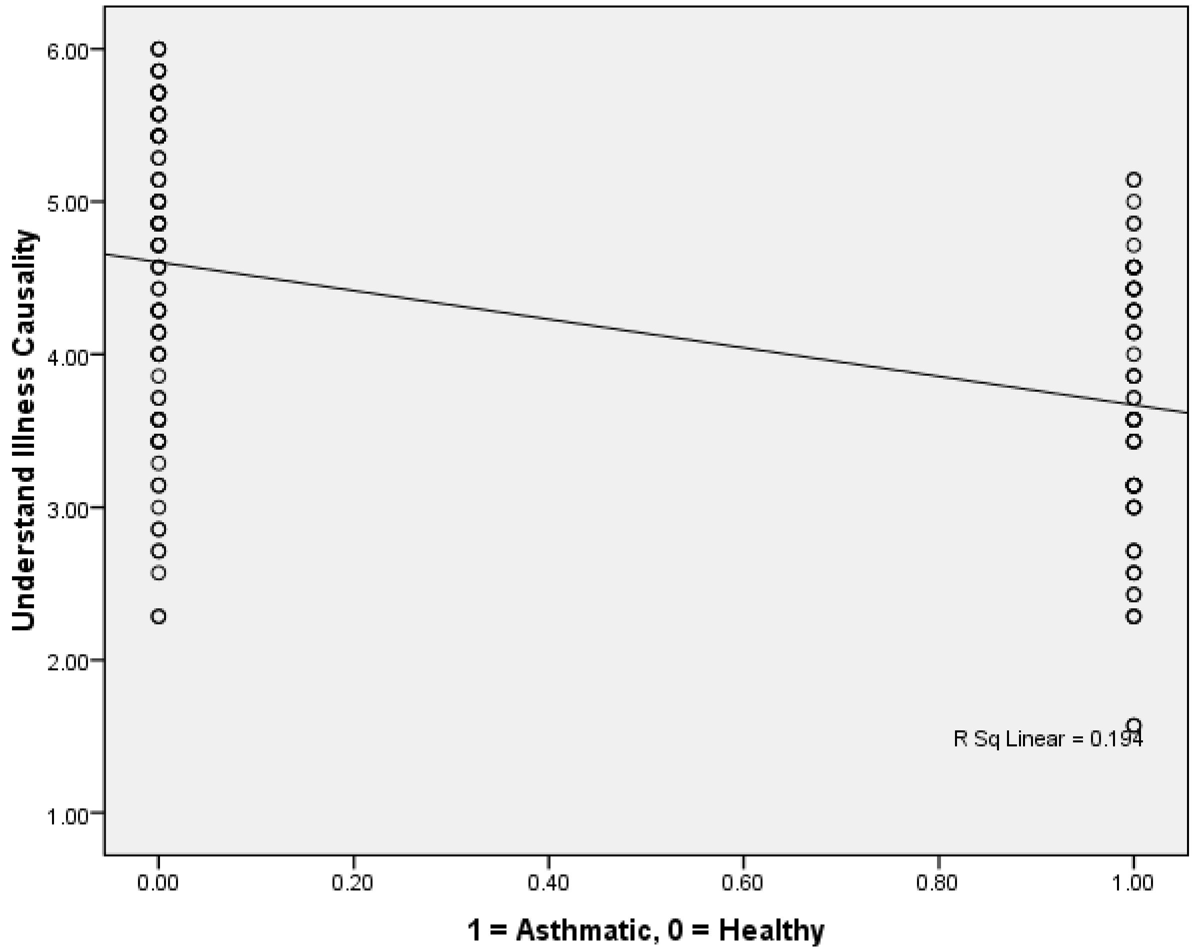
b. Dependent Variable: Understand Illness Causality

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	4.604	.098		46.807	.000	4.409	4.798
	1 = Asthmatic, 0 = Healthy	-.936	.151	-.440	-6.182	.000	-1.234	-.637

a. Dependent Variable: Understand Illness Causality

Understanding Causes of Illness (ILLCAUSE.sav)



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)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.353[a]	.125	.124	35.624

a. Predictors: (Constant), % of Students in Child's School Eligible for Free Lunch

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	158680.746	1	158680.746	125.040	.000[a]
	Residual	1114213.431	878	1269.036		
	Total	1272894.177	879			

a. Predictors: (Constant), % of Students in Child's School Eligible for Free Lunch

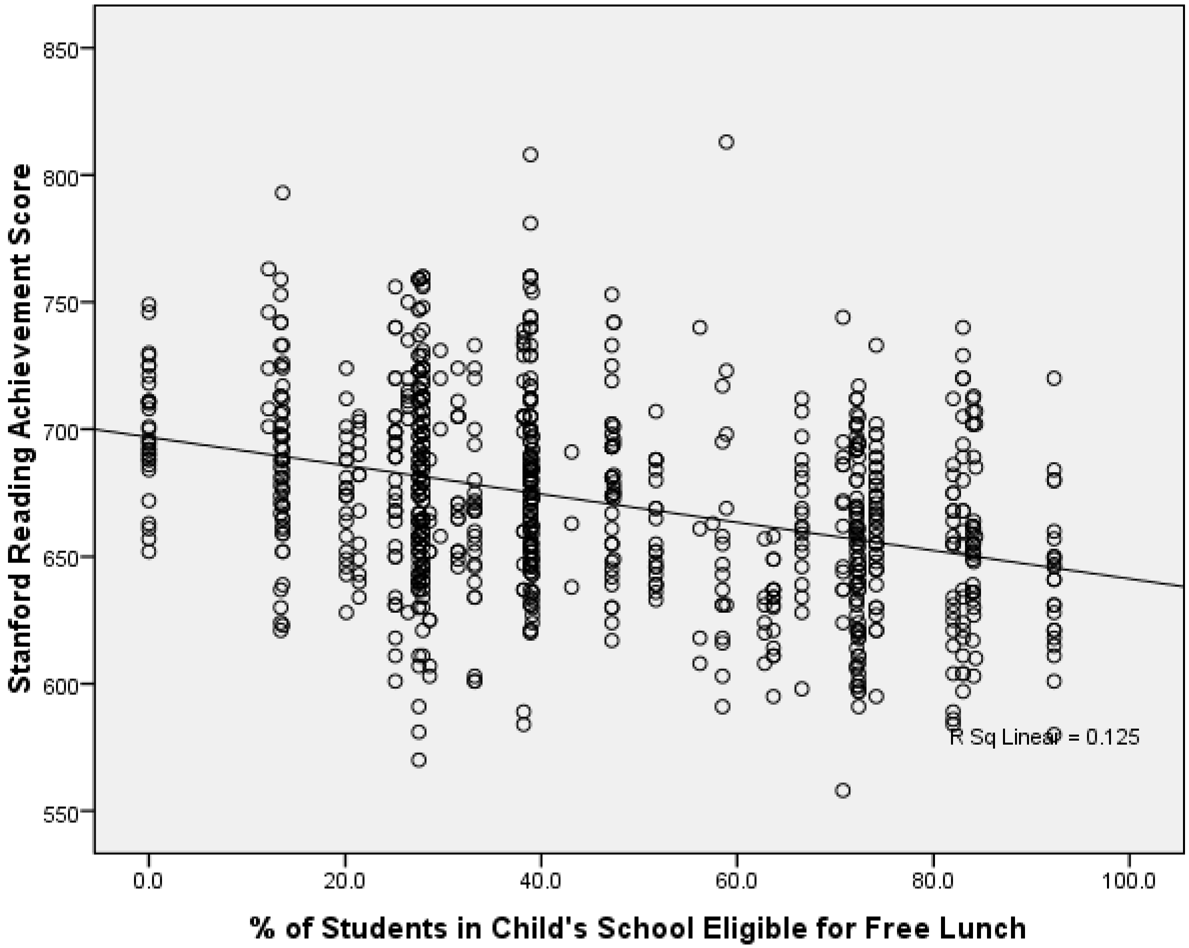
b. Dependent Variable: Stanford Reading Achievement Score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	696.847	2.540		274.325	.000
	% of Students in Child's School Eligible for Free Lunch	-.555	.050	-.353	-11.182	.000

a. Dependent Variable: Stanford Reading Achievement Score

Children of Immigrants (ChildrenOfImmigrants.sav)



Children of Immigrants (ChildrenOfImmigrants.sav)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.404[a]	.163	.162	34.837

a. Predictors: (Constant), Composite Family SES Score

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	207358.576	1	207358.576	170.863	.000[a]
	Residual	1065535.601	878	1213.594		
	Total	1272894.177	879			

a. Predictors: (Constant), Composite Family SES Score

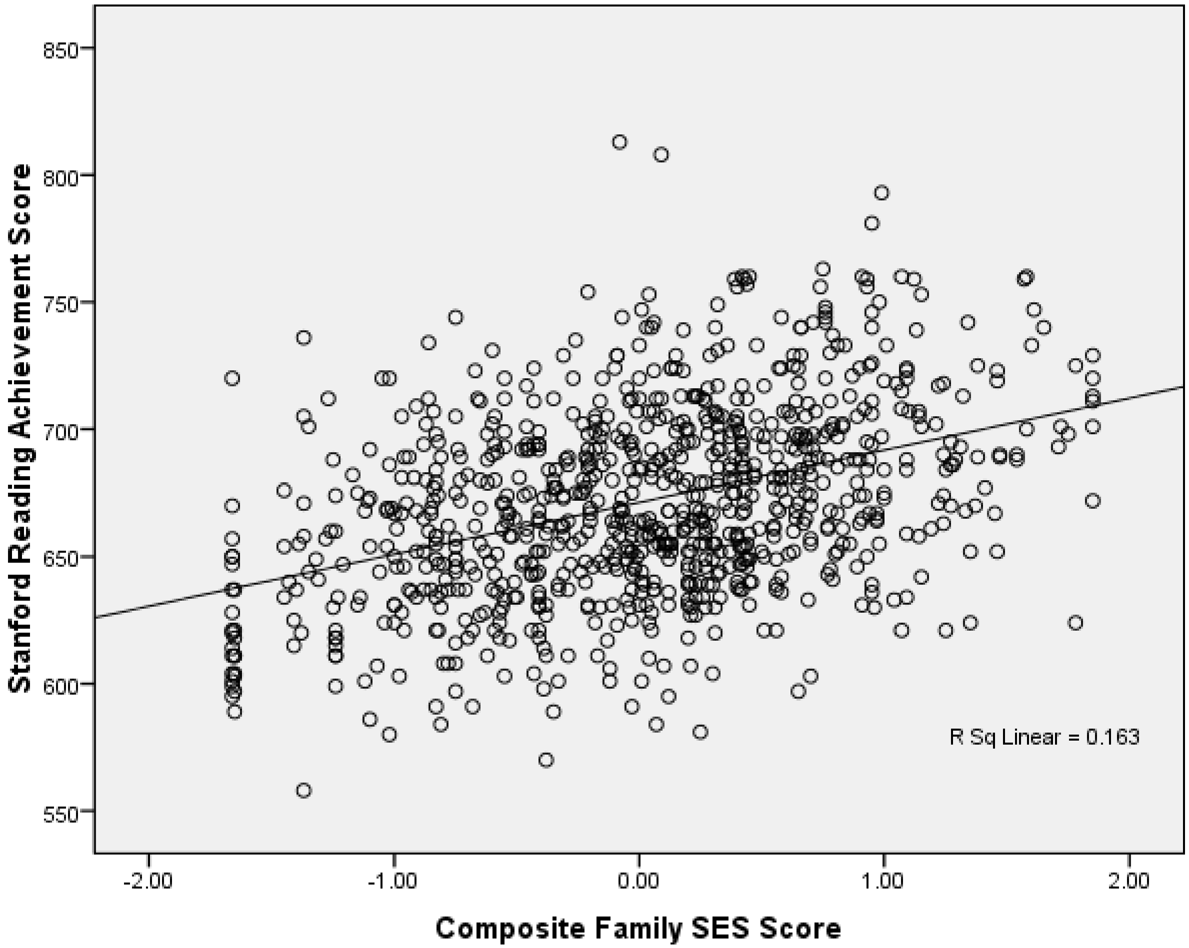
b. Dependent Variable: Stanford Reading Achievement Score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	671.350	1.175		571.418	.000
	Composite Family SES Score	20.418	1.562	.404	13.071	.000

a. Dependent Variable: Stanford Reading Achievement Score

Children of Immigrants (ChildrenOfImmigrants.sav)



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)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.382[a]	.146	.143	.91099

a. Predictors: (Constant), Collective efficacy

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	48.191	1	48.191	58.068	.000[a]
	Residual	282.170	340	.830		
	Total	330.361	341			

a. Predictors: (Constant), Collective efficacy

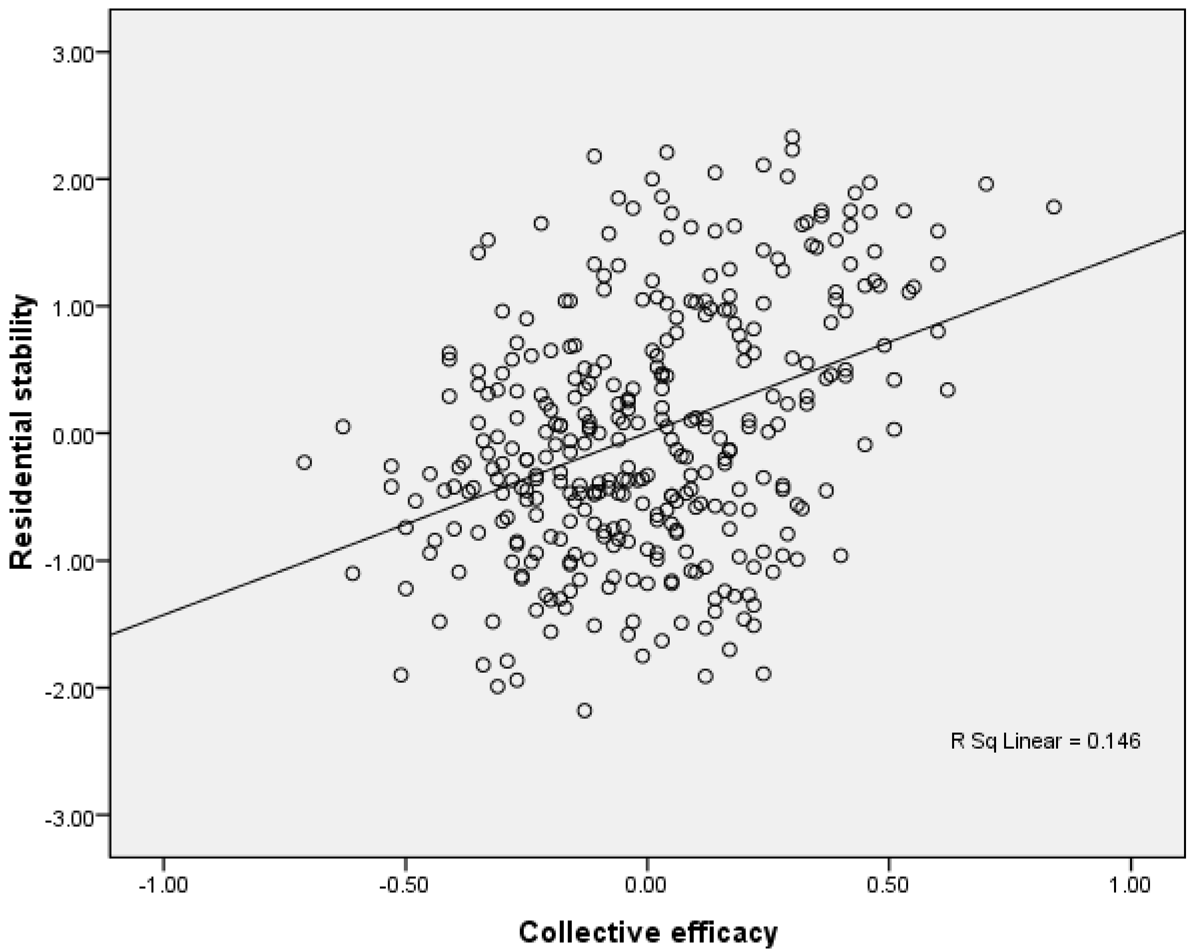
b. Dependent Variable: Residential stability

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.002	.049		.050	.961	-.094	.099
	Collective efficacy	1.429	.187	.382	7.620	.000	1.060	1.797

a. Dependent Variable: Residential stability

Human Development in Chicago Neighborhoods (Neighborhoods.sav)



Human Development in Chicago Neighborhoods (Neighborhoods.sav)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.147[a]	.022	.019	.97506

a. Predictors: (Constant), Homicide rate 1988-90

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7.112	1	7.112	7.480	.007[a]
	Residual	323.249	340	.951		
	Total	330.361	341			

a. Predictors: (Constant), Homicide rate 1988-90

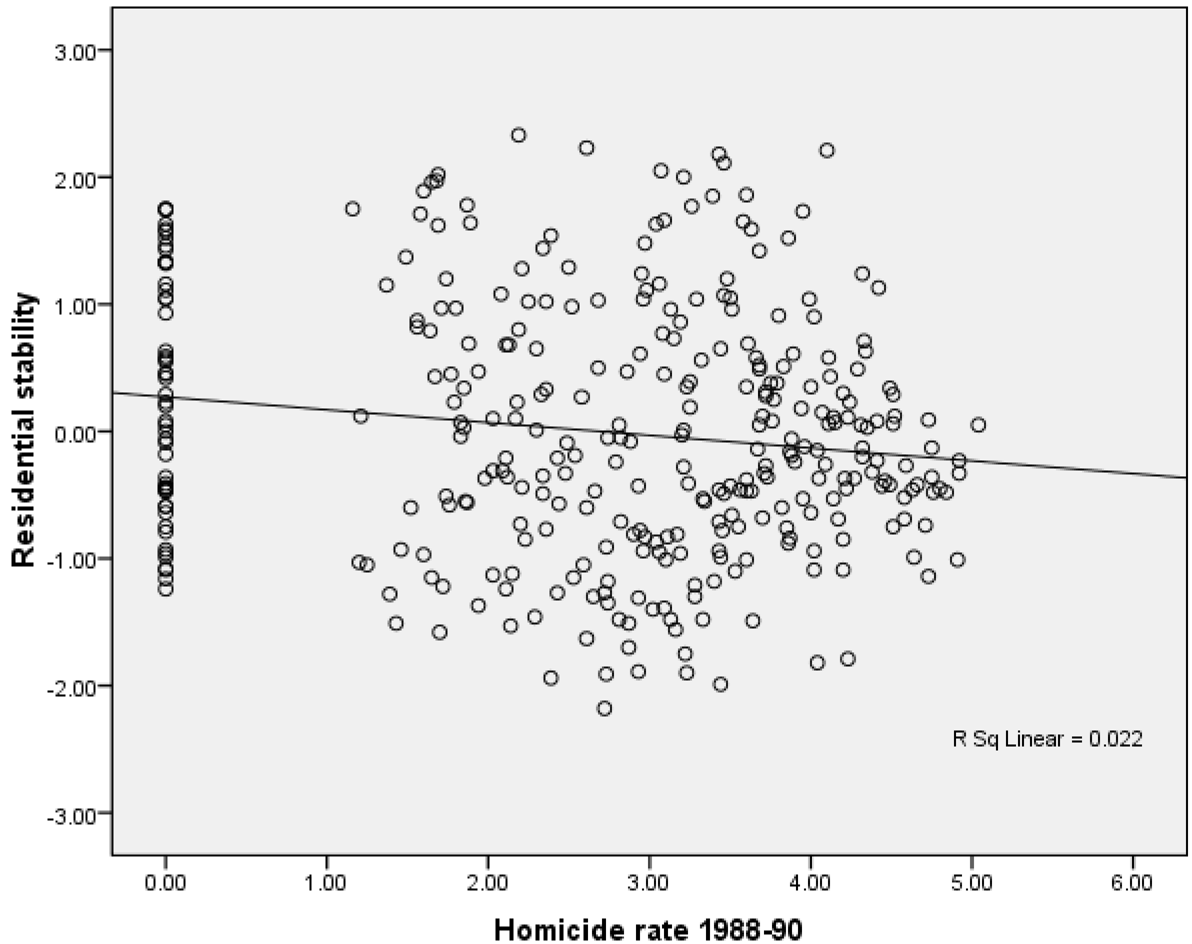
b. Dependent Variable: Residential stability

Coefficients[a]

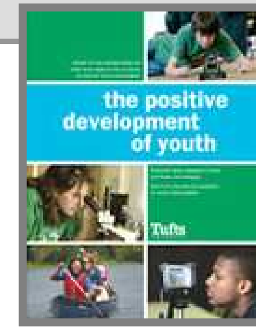
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.270	.111		2.432	.016	.052	.489
	Homicide rate 1988-90	-.100	.037	-.147	-2.735	.007	-.173	-.028

a. Dependent Variable: Residential stability

Human Development in Chicago Neighborhoods (Neighborhoods.sav)



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

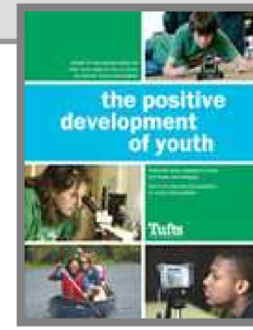


- **4-H Study of Positive Youth Development**
- **Source: Subset of data from IARYD, Tufts University**

- **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)



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.559[a]	.313	.311	.50341

a. Predictors: (Constant), Depression

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	46.912	1	46.912	185.115	.000[a]
	Residual	103.141	407	.253		
	Total	150.053	408			

a. Predictors: (Constant), Depression

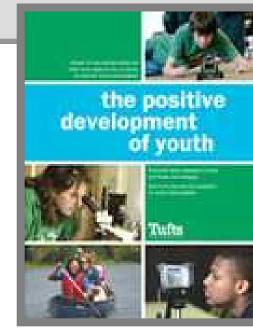
b. Dependent Variable: Self-Worth

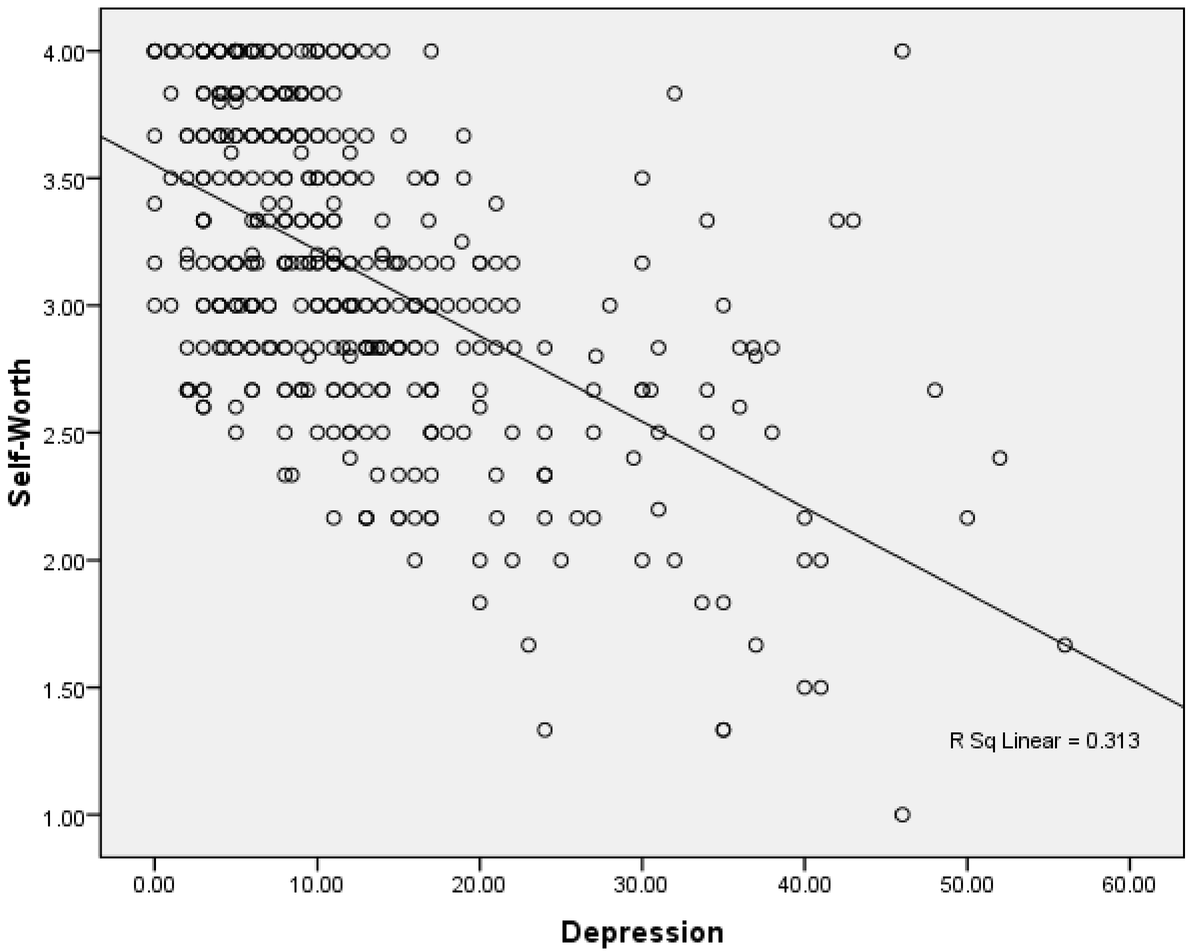
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.552	.040		88.146	.000
	Depression	-.034	.002	-.559	-13.606	.000

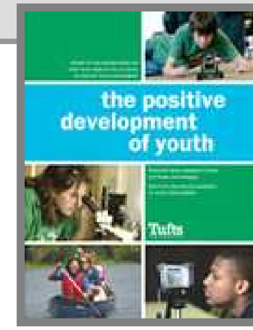
a. Dependent Variable: Self-Worth

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





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



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.504[a]	.254	.252	.52460

a. Predictors: (Constant), Depressed = 1, Not Depressed = 0

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	38.046	1	38.046	138.247	.000[a]
	Residual	112.007	407	.275		
	Total	150.053	408			

a. Predictors: (Constant), Depressed = 1, Not Depressed = 0

b. Dependent Variable: Self-Worth

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.307	.030		108.824	.000
	Depressed = 1, Not Depressed = 0	-.686	.058	-.504	-11.758	.000

a. Dependent Variable: Self-Worth

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