# Appendix A: Contingency Table Analysis and The Chi-Square Statistic

## Appendix A Post Hole:

Interpret a contingency table and associated chi-square statistic.

## Appendix A Technical Memo and School Board Memo:

Create a contingency table with an associated chi-square statistic in order to describe the relationship between two categorical predictors (i.e., dichotomous or polychotomous predictors).

## Appendix A Reading:

http://onlinestatbook.com/
Chapter 14, Chi Square

# Appendix A: Road Map (VERBAL)

Outcome Variable (aka Dependent Variable):

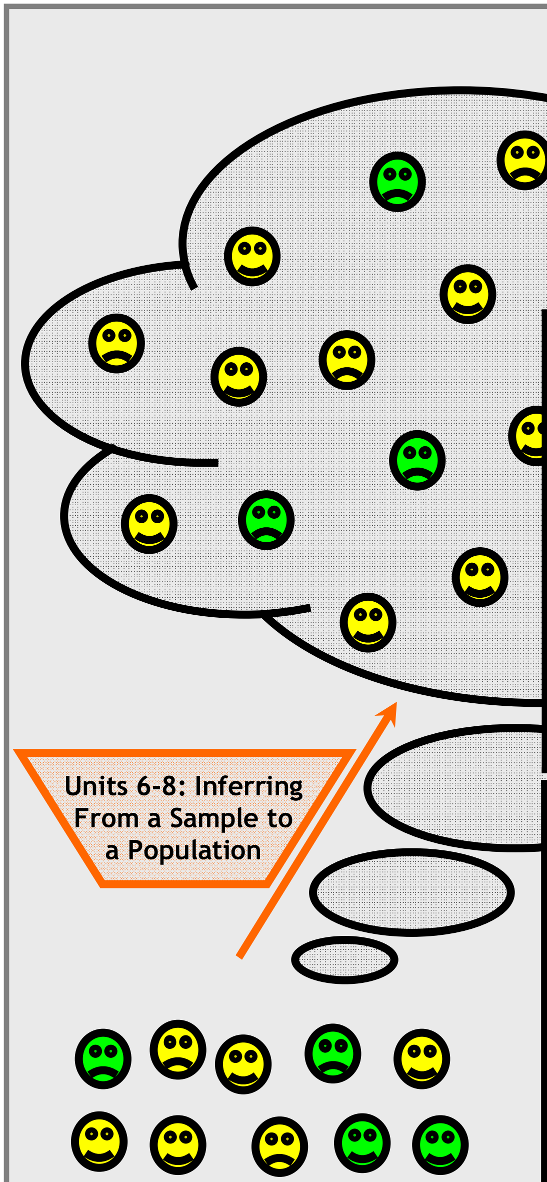
***READING***, a continuous variable, standardized test score, mean = 47 and standard deviation = 9

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? (Perspective I)
- Unit 3: In our sample, what does reading achievement look like? (Perspective II)
- 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 both race and free lunch? (Part I)
- Unit 11: In the population, is there a relationship between reading and both race and free lunch? (Part II)
- Appendix A: In the population, is there a relationship between race and free lunch?

# Appendix A: Road Map (Schematic)




| | | Single Predictor | | |
|---|---|---|---|---|
| | | Continuous | Polychotomous | Dichotomous |
| Outcome | Continuous | Regression | Regression<br>ANOVA | Regression<br>ANOVA<br>T-tests |
| | Polychotomous | Logistic Regression | Chi Squares | Chi Squares |
| | Dichotomous | | Chi Squares | Chi Squares |

| | | Multiple Predictors | | |
|---|---|---|---|---|
| | | Continuous | Polychotomous | Dichotomous |
| Outcome | Continuous | Multiple Regression | Regression<br>ANOVA | Regression<br>ANOVA |
| | Polychotomous | Logistic Regression | Chi Squares | Chi Squares |
| | Dichotomous | | Chi Squares | Chi Squares |

# The story so far…

You have the training and experience to handle a wide range of data analytic challenges. Whenever you are confronted with data, be it your own project or the project of another, ask yourself these three pre-data-analytic questions:

1. What is the theory?
2. What are the research questions?
3. What are the outcome and predictor(s).

Then, you are ready to conduct your data analysis:

1. Exploratory Data Analysis (EDA), Units 1 and 2
2. Descriptive Data Analysis, Units 3, 4, 5
3. Confirmatory Data Analysis, Units 6, 7, 8, 9, 10

If your outcome is continuous, you are golden. Hitherto, everything in this course has been geared toward continuous outcomes.

If your outcome is categorical (i.e., dichotomous or polychotomous), we can learn to deal with it today as long as your predictors are also categorical.

If your outcome is categorical and your predictor is continuous, then you need logistic regression which you can learn in Appendix B.

| | | Single Predictor | | |
|---|---|---|---|---|
| | | Continuous | Polychotomous | Dichotomous |
| Outcome | Continuous | Regression | Regression<br>ANOVA | Regression<br>ANOVA<br>T-tests |
| | Polychotomous | Logistic Regression | Chi Squares | Chi Squares |
| | Dichotomous | | Chi Squares | Chi Squares |

I focused on regression over ANOVA and t-tests, because regression is the most flexible tool. Once you determine that your outcome is continuous, you do not need to fuss about what statistical tool to use—use regression. Of course, you can use the other tools, and you should if your lab uses those tools.

**Our tools are powerful, but limited. Always check your assumptions thoroughly and interpret your results cautiously.**

# Epistemological Minute

In the face of uncertainty, we must continue to make decisions. A dominant goal of this course has been making reasonable decisions about our hypotheses despite uncertainty due to sampling error. We see a relationship in our random sample, but can we draw an inference from the sample to the population?

In elementary decision theory, to make rational decisions, we complete a decision/condition table with costs/benefits for each combination of the decisions and conditions and with probabilities for each condition. We must consider the consequences *and* the probabilities.

| | It Is Going to Rain<br>Probability = p | It is Not Going to Rain<br>Probability = 100% - p |
|---|---|---|
| Conclude<br>"It Is Going to Rain" | Costs/Benefits | Costs/Benefits<br>(False Positive) |
| Conclude<br>"It Is Not Going to Rain" | Costs/Benefits<br>(False Negative) | Costs/Benefits |

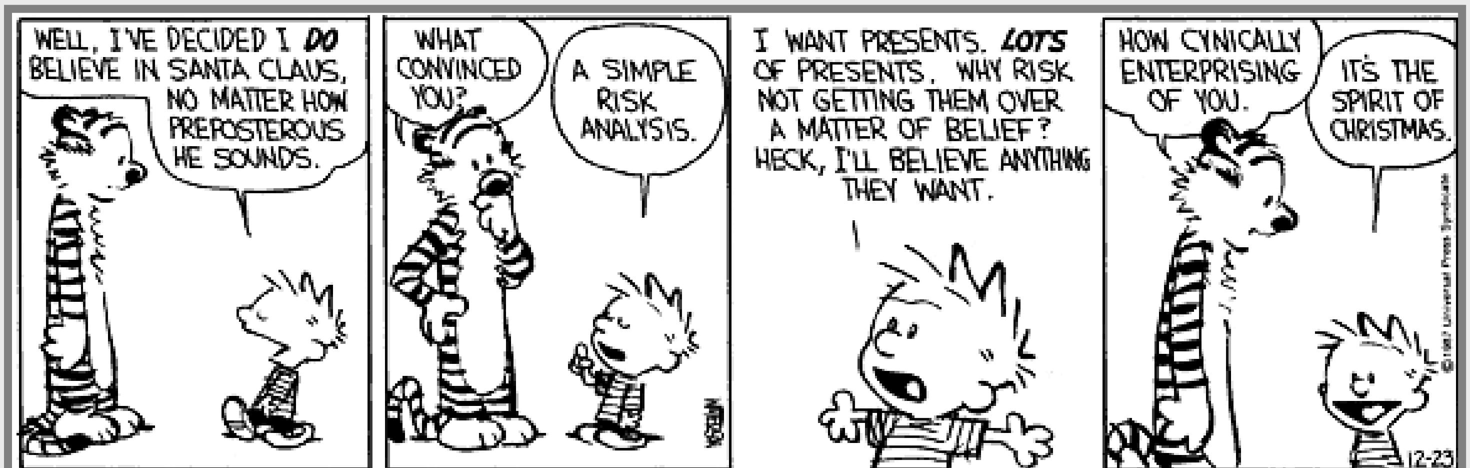
| | It Is Going to Rain<br>Probability = 50% | It is Not Going to Rain<br>Probability = 50% |
|---|---|---|
| Conclude<br>"It Is Going to Rain" | You bring your umbrella, and it keeps you dry. | You bring your umbrella, and it is dead weight. |
| Conclude<br>"It Is Not Going to Rain" | You get soaked on the way to a job interview. | Life is life. |

# Epistemological Minute

In his *Pensées* (1670), Blaise Pascal presents a seminal (and controversial) argument in philosophy, theology and decision theory. This argument is called “Pascal’s Wager”:

| | God Exists<br>Probability > 0%<br>(According to Pascal) | God Does Not Exist<br>Probability < 100%<br>(According to Pascal) |
|---|---|---|
| **Conclude “God Exists”** | From Pascal’s Christian perspective, there are infinite benefits to believing in God when God exists. | You fill in this blank. |
| **Conclude “God Does Not Exist”** | You fill in this blank. | You fill in this blank. |

You decide the probabilities and the costs and benefits, but if you agree with Pascal that there are infinite benefits to believing in God when God exists, and if you agree with Pascal that the probability of God’s existence is greater than zero, even if it is 0.000000001%, then it is reasonable to conclude that God exists.




The special thing about Pascal’s argument is the infinite nature of the true-positive benefits. It makes the probabilities virtually obsolete. Usually, knowing the probabilities is essential, but this is an exception.

# Epistemological Minute

Whether you agree with Pascal or not, I want you to understand that, for making decisions, it is essential to take into consideration the various consequences as well as the various probabilities. In statistics, we are presented with this (perhaps less weighty) decision matrix:

| | There Is a Relationship In The Population *If* Probability = 0% | There Is NO Relationship In The Population *If* Probability = 100% |
|---|---|---|
| Conclude "There Is a Relationship In The Population" | Benefits of concluding the truth. | *Conditional* Probability Of Type I Error = Significance Level E.g., 2.1% |
| Do NOT Conclude "There Is a Relationship In The Population" | False Negative - Type II Error Costs of not concluding the truth. | Benefits of not concluding a falsehood. |

Determining the probabilities is challenging business, and the killer here is that many data analysts wrongly think that they know the probabilities. They think that that the p-value (i.e., significance level) associated with a null hypothesis test is either the probability that "There is NO Relationship In The Population" or the probability of Type I Error. Neither interpretation is right. Both are wrong. Rather, the p-value is the probability of Type I Error *when* "There is NO Relationship." Say, $p = .021$; this does not mean that there is a 2.1% chance that "There is NO relationship in the population (i.e., the null is true). Nor does "$p = .021$" mean that there is a 2.1% chance of Type I Error (i.e., a False Positive). Rather, "$p = .021$" means that there is a 2.1% chance of Type I Error *if* there is a 100% chance "There is NO Relationship." The significance level is a *conditional* probability, not the absolute probability we need.

Determining the costs and benefits is also challenging business. For example, I believe that educational researchers give too little weight to the costs of false negatives, Type II Error. There is no magic bullet in education. Effective education is the accumulation of a million good educational effects. If a study examines one educational effect, the effect may be very small. Small effects are hard to detect. Large sample sizes help to detect small effects, but in school-based research, even the largest possible samples (the entire school!) may not be large enough. There are some other things that can help: (1) Measure the specific educational effect, not the general educational effect. Don't use a shotgun when you can use a laser. Don't use the MCAS when you can use a test of the exact skills being taught. The cost is that new measures are difficult to design. (2) Make sure that you measure your outcome reliably. The cost is that reliable measures are difficult to design. (3) Measure a bundle of interventions that combine small effects for a larger effect. The cost is that interventions must be more complex and thus more difficult (and expensive) to implement. (4) Increase your alpha level. The cost is increasing the probability of Type I Error. In sum, false negatives (Type II Errors) are costly, but you *can* decrease their probability.

# Epistemological Minute

If the statistical significance approach leaves you in limbo, you may want to try the confidence interval approach.  The confidence interval approach gives us the probabilities for which we are looking.

| | The Population Magnitude Is Within The Interval Probability = 95% | The Population Magnitude Is NOT Within the Interval Probability = 5% |
|---|---|---|
| Conclude "The Population Magnitude Is Within The Interval" | Benefits of concluding the truth. | Costs of concluding a falsehood. |
| Conclude "The Population Magnitude Is NOT Within The Interval" | Costs of concluding a falsehood. | Benefits of concluding the truth. |

However, the 95% assumes:

1. The confidence interval is the only available evidence about the population magnitude. If there is other evidence, it should be reflected in the probability assignment. For example, if three other studies estimate the population magnitude to be outside your 95% confidence interval, maybe you should not be so confident about your interval.
2. The sample is randomly drawn from the population. Strictly, we can only draw statistically warranted conclusions about the population from which randomly drew our sample. Practically, there is wiggle room for which you must argue: "It is reasonable to treat my convenience sample of 47 Somerville preschoolers as a random sample of United States preschoolers, because… <<Proceed to pull a rabbit out of your hat.>>"
3. The regression assumptions are perfectly met. Regression methods are largely robust to assumption violations. As demonstrated in Unit 8, for example, heteroskedasticity and non-normality distort the 95% but not wildly. Robustness is not an invitation to ignore assumption violations, but it does recommend some tolerance.
4. No math errors. No transcription errors. Our methods are designed to deal with sampling error (and measurement error in the outcome); they are not a panacea for all possible sources of error.

Today we are going to learn about non-parametric tests of statistical significance. They are wonderfully flexible, but not as informative as the parametric methods we have developed hitherto. One weakness of non-parametric methods is that they do not (easily) permit the construction of confidence intervals.

# Appendix A: Research Question I

Theory: Girls develop social aptitudes faster than boys. Therefore, among children of 36 to 60 months of age, girls are more likely than boys to endorse trustworthy informants.

Research Question: Among 36 to 60-month-old children, girls, more than boys, initially endorse trustworthy informants.

Data Set: Trust_and_Testimony.sav

Variables:

Outcome—*InitialEndorsementYesNo* (a categorical variable)

0 = no initial endorsement, 1 = initial endorsement

Predictor—*Female* (a categorical variable)

0 = male, 1 = female

The Children's Trust

# Trust_and_Testimony.sav Codebook

**Trust_and_Testimony.sav**

This is a small data subset from the trust and testimony research program conducted by Paul Harris and Kathleen Corriveau (among others) at the Harvard Graduate School of Education. The purpose of the research program is to answer questions about the role of testimony in cognitive development. Their work is much more thorough than the limited peek I provide here. You may read about it here:

Harris, P.L. (2007). Trust. *Developmental Science*, *10*, 135-138.

Corriveau, K., & Harris, P.L. (in press). Preschoolers continue to trust a more accurate informant 1 week after exposure to accuracy information. *-Developmental Science*.

Sample of 86 preschoolers local to Cambridge, MA.

| Variable Name | Variable Descriptions |
|---|---|
| Age | Age in Months (36-60) |
| Female | 0 = Male, 1 = Female |
| InitialEndorsementYesNo | 1 if the child tended to endorse the trustworthy informant immediately after stimulus, 0 if else |
| LaterEndorsementProportion | Proportion of possible endorsement of the trustworthy informant, 4-7 days after stimulus. |

# Trust and Testimony Data Set

*Trust and Testimony.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Util

47 :

| | Age | Female | InitialEndorsementYesNo |
|---|---|---|---|
| 38 | 51 | 0.00 | 0.00 |
| 39 | 55 | 0.00 | 1.00 |
| 40 | 60 | 0.00 | 1.00 |
| 41 | 44 | 1.00 | 1.00 |
| 42 | 47 | 1.00 | 1.00 |
| 43 | 45 | 1.00 | 1.00 |
| 44 | 48 | 1.00 | 0.00 |

Trust and Testimony.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 | Age | Numeric | 3 | 0 | Age in Months | None | None | 3 | Left | Scale |
| 2 | Female | Numeric | 8 | 2 | Female | {0.00, Boy}... | None | 8 | Right | Nominal |
| 3 | InitialEndorsementYesNo | Numeric | 8 | 2 | Initial Endorsement (Y/N) | {0.00, No Ini... | None | 18 | Right | Nominal |
| 4 | FxIE | Numeric | 8 | 2 | Interaction (FxIE) | None | None | 13 | Right | Scale |
| 5 | LaterEndorsementProportion | Numeric | 8 | 2 | Proportion of Endorsement 4-7 Days Later | None | None | 8 | Right | Scale |

# Getting a Handle on "Expectation"

The 111th United States Senate had 100 members, one of whom is African American, Roland Burris.

How would you describe the representation of African Americans in the U.S. Senate? Overrepresented? Proportionally Represented? Underrepresented?

What would proportional representation look like?

The 111th United States Senate had 17 female senators.

What would proportional representation look like?

We can think of two categorical variables (in this case dichotomous variables): *FEMALE* and *SENATOR*. Our sample is all U.S citizens. *__If there were no relationship__* between *FEMALE* and *SENATOR*, we would *__expect__* proportional representation. We can compare our observation to our expectation.

| | | Senator = 0 (Not a Senator) | Senator = 1 (Senator) | Total |
|---|---|---|---|---|
| Female = 0 (Male) | Observed Count | 150,972,780 | 83 | 150,972,863 |
| | Expected Count | 150,972,813.7 | 49.3 | 150,972,863 |
| Female = 1 (Female) | Observed Count | 155,260,107 | Observation 17 | 155,260,124 |
| | Expected Count | 155,260,073.3 | Vs. 50.7 Expectation | 155,260,124 |
| Total | Observed Count | 306,232,887 | 100 | 306,232,987 |
| | Expected Count | 306,232,887 | 100 | 306,232,987 |

Here, "expected" means the expected frequency if the null hypothesis were true. It is our baseline for comparison.

# Contingency Table Analysis (With $\chi^2$ Statistic)

**Female * Initial Endorsement (Y/N) Crosstabulation**

| | | | Initial Endorsement (Y/N) | | |
|---|---|---|---|---|---|
| | | | No Initial Endorsement | Initial Endorsement Yes | Total |
| Female | Boy | Count | 8 | 32 | 40 |
| | | Expected Count | 8.4 | 31.6 | 40.0 |
| | | Std. Residual | -.1 | .1 | |
| | Girl | Count | 10 | 36 | 46 |
| | | Expected Count | 9.6 | 36.4 | 46.0 |
| | | Std. Residual | .1 | .0 | |
| | Total | Count | 18 | 68 | 86 |
| | | Expected Count | 18.0 | 68.0 | 86.0 |

The observed Count is the number of people in that category.

The Expected Count is the number of people who we would expect in that category if there were no relationship in the population.

The Standardized Residual is the residual (Observed Count minus Expected Count) divided by a standard deviation.

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
|---|---|---|---|---|---|
| Pearson Chi-Square | .039[a] | 1 | .843 | | |
| Continuity Correction[b] | .000 | 1 | 1.000 | | |
| Likelihood Ratio | .039 | 1 | .843 | | |
| Fisher's Exact Test | | | | 1.000 | .528 |
| Linear-by-Linear Association | .039 | 1 | .844 | | |
| N of Valid Cases | 86 | | | | |

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 8.37.

b. Computed only for a 2x2 table

We *DO NOT* reject the null hypothesis that there is no relationship in the population.

If the standardized residual is greater than |2|, there is considerable over/ underrepresentation in that cell.

There is a relationship in our sample (sort of, since the observed counts differ from the expected counts), but we do not want to generalize that relationship to the population, for obvious (here) reasons.

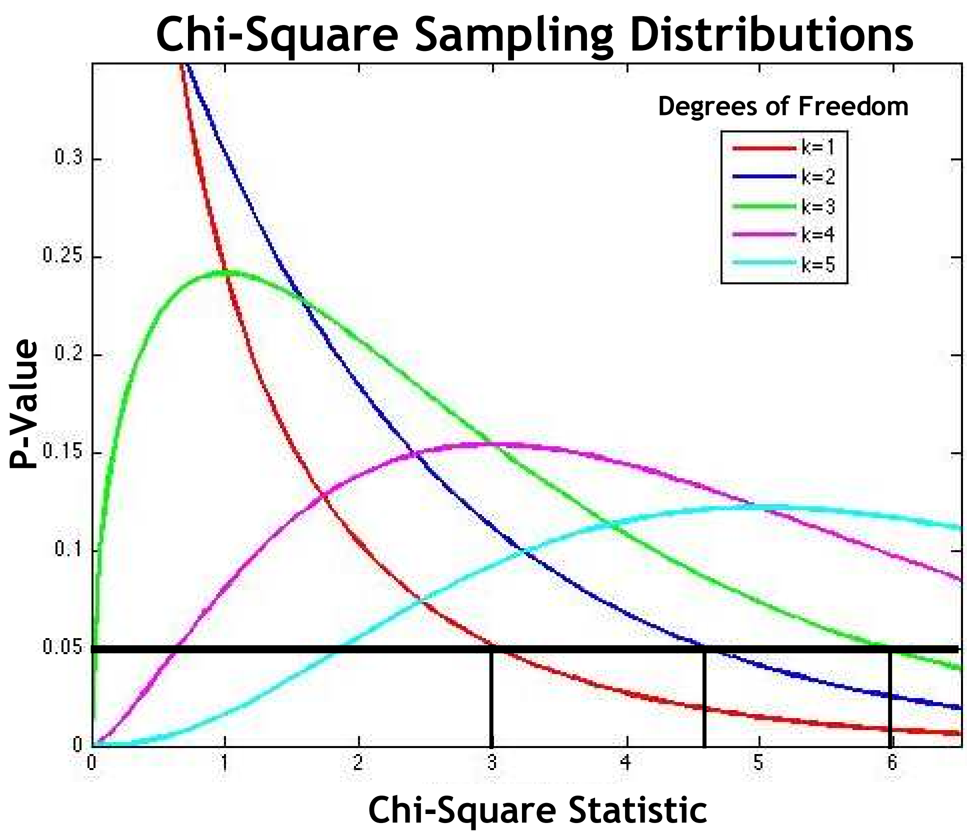
The relationship between sex and initial endorsement is not statistically significant, $\chi^2 = 0.039$, $p = 0.843$.

# The $\chi^2$ Statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

Where :

- O = Observed frequency
- E = Expected frequency
- k = Number of categories or groupings




**We "want" a big chi-square statistic, so we "want" a big difference between observed and expected frequencies. How big is a big chi-square statistic? That depends on the degrees of freedom. As with t-tests and F-tests, we must refer to the sampling distribution to a get a p-value for a given chi-square statistic, and the sampling distributions look different depending on the degrees of freedom. A chi-square statistic of 0.039 is small no matter how you slice it.**

```
CROSSTABS
  /TABLES=Female BY InitialEndorsementYesNo
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT EXPECTED SRESID
  /COUNT ROUND CELL.
```

# The $\chi^2$ Statistic By Hand (And Key Facts)

Table A.X. A contingency table with observed and expected counts for sex and Senate membership in the U.S. population, sample size approximately 306 million.

| | | Senator=0 (Not a Senator) | Senator=1 (Senator) | Total |
|---|---|---|---|---|
| Female = 0 (Male) | Observed | 150,972,780 | 83 | 150,972,863 |
| | Expected | 150,972,813.7 | 49.3 | 150,972,863 |
| Female = 1 (Female) | Observed | 155,260,107 | 17 | 155,260,124 |
| | Expected | 155,260,073.3 | 50.7 | 155,260,124 |
| Total | Observed | 306,232,887 | 100 | 306,232,987 |
| | Expected | 306,232,887 | 100 | 306,232,987 |

Different cells make different chi-square contributions. We might say "the action" is in the cells that make the biggest chi-square contributions.

When you have fewer than 5 expected observations in a cell, your chi-square statistic becomes questionable.

Besides the independence assumption, there are no "HINLO" assumptions to worry about. This is not really a strength! You pay for it in statistical power. Whenever appropriate, work with continuous variables.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

$$= \frac{(150{,}972{,}780 - 150{,}972{,}813.7)^2}{150{,}972{,}813.7} + \frac{(155{,}260{,}107 - 155{,}260{,}073.3)^2}{155{,}260{,}073.3} + \frac{(83-49.3)^2}{49.3} + \frac{(17-50.7)^2}{50.7}$$

$$= \frac{(-33.7)^2}{150{,}972{,}813.7} + \frac{(33.7)^2}{155{,}260{,}073.3} + \frac{(33.7)^2}{49.3} + \frac{(-33.7)^2}{50.7}$$

$$= \frac{1{,}135.69}{150{,}972{,}813.7} + \frac{1{,}135.69}{155{,}260{,}073.3} + \frac{1{,}135.69}{49.5} + \frac{1{,}135.69}{50.7}$$

$$= 0.000008 + 0.000007 + 22.9 + 22.4$$

$$= 45.3$$

The chi-square statistic is a non-parametric statistic; it does not rely on averages, which frees it from assumptions about homoscedasticity, normality, linearity and outliers. Non-parametric statistics instead rely on counts and ranks. The cost of these nearly-assumption-free statistics is that they cannot take advantage of relative differences. If I am in a foot race with nine Olympic runners, non-parametric stats only care that I came in last place. They ignore (perhaps useful) information about how far behind I finished.

# Calculating The Expectations

Note: You can include any expected frequencies that you want! You can test all sorts of null hypotheses. So, if you have information from *outside* your data about expected frequencies, then you can use those expectations.

If you only have information from *inside* your data about expected frequencies, then you can calculate the expected cell frequencies from the marginal frequencies. If you do this, your null hypothesis will be proportional representation.

$$E = \frac{RowTotal * ColumnTotal}{TotalSampleSize} = \boxed{\frac{RowTotal}{TotalSampleSize}} * \boxed{ColumnTotal}$$

| | | Senator=0 (Not a Senator) | Senator=1 (Senator) | Total |
|---|---|---|---|---|
| Female = 0 (Male) | Observed | 150,972,780 | 83 | 150,972,863 |
| | Expected | 150,972,813.7 | 49.3 | 150,972,863 |
| Female = 1 (Female) | Observed | 155,260,107 | 17 | 155,260,124 |
| | Expected | 155,260,073.3 | 50.7 | 155,260,124 |
| Total | Observed | 306,232,887 | 100 | 306,232,987 |
| | Expected | 306,232,887 | 100 | 306,232,987 |

If our null hypothesis is proportional representation, we can grab the proportions from the row total.

We have 150,972,863 men in our sample.

That's .493 or (49.3%) of the total. So, if we "expect" proportional representation, men should be .493 of the non-Senators and men should be .493 of the Senators.

.493 of 306,232,887 is 159,972,813.7

.493 of 100 is 49.3

# Appendix A: Research Question II

Theory: When assessing the magnitude of the Anglo/Latino reading gap, we must control for SES, because SES, as a known correlate with ethnicity, is a potential confound.

Research Question: Are ethnicity and SES correlated?

Data Set: NELSBoys.sav National Education Longitudinal Survey (1988), a subsample of 1820 four-year-college bound boys, of whom 182 are Latino and the rest are Anglo.

Variables:

Outcome—Low SES=1, Mid SES=2, High SES=3 (*SocioeconomicStatus*)
Predictors—Latino = 1, Anglo = 0 (*LATINO*)

# Correlations Among Categorical Variables

**Latino * SocioEconomicStatus Crosstabulation**

| | | | SocioEconomicStatus | | | |
|---|---|---|---|---|---|---|
| | | | Low SES | Mid SES | High SES | Total |
| Latino | 0 | Count | 196 | 482 | 960 | 1638 |
| | | Expected Count | 237.6 | 485.1 | 915.3 | 1638.0 |
| | | Std. Residual | -2.7 | -.1 | 1.5 | |
| | 1 | Count | 68 | 57 | 57 | 182 |
| | | Expected Count | 26.4 | 53.9 | 101.7 | 182.0 |
| | | Std. Residual | 8.1 | .4 | -4.4 | |
| | Total | Count | 264 | 539 | 1017 | 1820 |
| | | Expected Count | 264.0 | 539.0 | 1017.0 | 1820.0 |

**If the relationship is statistically significant, look for large (greater than |2|) standardized residuals to see where the action is. Note that if there is overrepresentation in one cell, there must be underrepresentation in at least two other cells; this is deeply related to the concept of degrees of freedom.**

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 94.863[a] | 2 | .000 |
| Likelihood Ratio | 78.966 | 2 | .000 |
| Linear-by-Linear Association | 85.312 | 1 | .000 |
| N of Valid Cases | 1820 | | |

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 26.40.

```
CROSSTABS
  /TABLES=Latino BY SocioEconomicStatus
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT EXPECTED SRESID
  /COUNT ROUND CELL.
```

We reject the null hypothesis that there is no relationship in the population between ethnicity and socioeconomic status, $p < 0.05$.

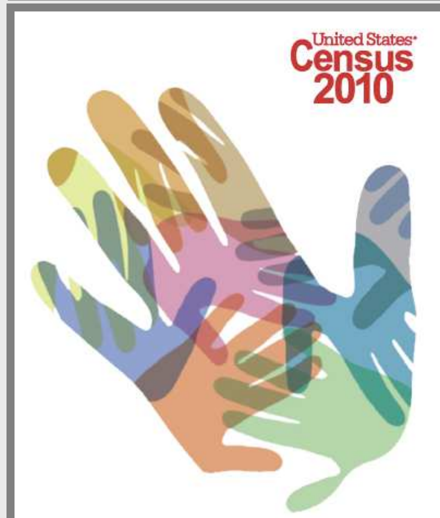
**We find a statistically significant relationship between ethnicity and socioeconomic status, $\chi^2 = 94.9$, $p < 0.001$. Latino students are overrepresented among low SES students and underrepresented among high SES students; 37% of Latino students are low SES as compared to only12% of Anglo students.**

# Interpreting Contingency Tables

- Familiarize yourself with the table by thinking in terms of representation: overrepresentation and underrepresentation. Compare the observed frequencies to the expected frequencies.

- Ask whether the differences between your observations and expectations are statistically significant.  Use the chi-square statistic (and associated p-value) to test the null hypothesis that there is no relationship in the population. In other words, test the null hypothesis that, in the population, the observed and expected frequencies are perfectly equal, and thus the over/underrepresentation in your sample is merely an artifact of sampling error. If, based on a p-value of less than .05 you reject the null, then conclude that there is a relationship in the population.

- Use standardized residuals ( greater than about 2) to determine where the action is.


Now, you have what you need for Post Hole A. Practice is in back.

# Dig the Post Hole (SPSS)

## Appendix A Post Hole:

Interpret a contingency table and associated chi-square statistic.

**Evidentiary material: contingency table with tests.**
**(From http://benbaab.com/salkind/ChiSquare.html.)**

Type * Style Crosstabulation

| | | | Style | | | | |
|---|---|---|---|---|---|---|---|
| | | | Lecture/ Discussion | Collaborative/ Cooperative | Hands-on/ Interactive | Individualize d/Self-paced | Total |
| Type | Art | Count | 5 | 5 | 9 | 11 | 30 |
| | | Expected Count | 9.2 | 5.0 | 9.6 | 6.2 | 30.0 |
| | | Residual | -4.2 | .0 | -.6 | 4.8 | |
| | Humanities | Count | 23 | 9 | 10 | 8 | 50 |
| | | Expected Count | 15.3 | 8.3 | 16.0 | 10.3 | 50.0 |
| | | Residual | 7.7 | .7 | -6.0 | -2.3 | |
| | Math | Count | 9 | 6 | 8 | 7 | 30 |
| | | Expected Count | 9.2 | 5.0 | 9.6 | 6.2 | 30.0 |
| | | Residual | -.2 | 1.0 | -1.6 | .8 | |
| | Science | Count | 9 | 5 | 21 | 5 | 40 |
| | | Expected Count | 12.3 | 6.7 | 12.8 | 8.3 | 40.0 |
| | | Residual | -3.3 | -1.7 | 8.2 | -3.3 | |
| Total | | Count | 46 | 25 | 48 | 31 | 150 |
| | | Expected Count | 46.0 | 25.0 | 48.0 | 31.0 | 150.0 |

Chi-Square Tests

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 20.739[a] | 9 | .014 |
| Likelihood Ratio | 19.785 | 9 | .019 |
| Linear-by-Linear Association | .043 | 1 | .835 |
| N of Valid Cases | 150 | | |

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.00.

**Here is the answer blank:**

**Yakkity yak yak yak, $\chi^2$(df) = xx.x, = .xxx.**

**Steps:**

1) Check the Pearson Chi-Square Statistic.
   1) Keep in mind the null: No relationship (i.e., proportional representation) in the population.
   2) Reject the null if p < .05.
   3) If you reject, the relationship in your sample is statistically significant.
   4) You can draw an inference from the sample to the population.
2) Look for residuals greater than +/- 2 to see where the action is.
3) Check your assumptions.
   1) Independence.
   2) Expected counts are 5 or greater.

**Here is my answer:**

**There is a statistically significant relationship between course type and course style, $\chi^2$(9) = 20.7, = .014.**

**Hands-on/interactive courses are over-represented in science courses and under-represented in humanities courses. Lecture/discussion courses are over-represented in humanities courses.**

**Independence seems a reasonable assumption. All the expected cell counts are 5 or greater.**

# Future Directions: Other Non-Parametric Statistical Tests

**The chi-square test is one of many non-parametric statistical tests. It is the most flexible, but the least powerful. (If I could only teach one statistical test, perhaps it would be the chi-square test, because of its flexibility.) In this slide, I would like to introduce two other non-parametric tests: the Spearman's Rank Correlation and the Wilcoxon Rank Sum Test (aka the Mann-Whitney U Test).**

**There are two types of polychotomies: nominal and ordinal. Nominal polychotomies represent categories with no natural ranking implied. RACE/ETHNICITY is a nominal polychotomy. For data analytic purposes, we may assign White=1, Black=2, Latino=3 and Asian=4, but those numbers are merely numerical labels, not ranks. Other nominal polychotomies are RELIGION (Christian, Jewish, Muslim…), MARITALSTATUS (Single, Married, Divorced, Widowed), MUSICGENRE (Rock, Rap, Country, Classical, Jazz, R&B).**

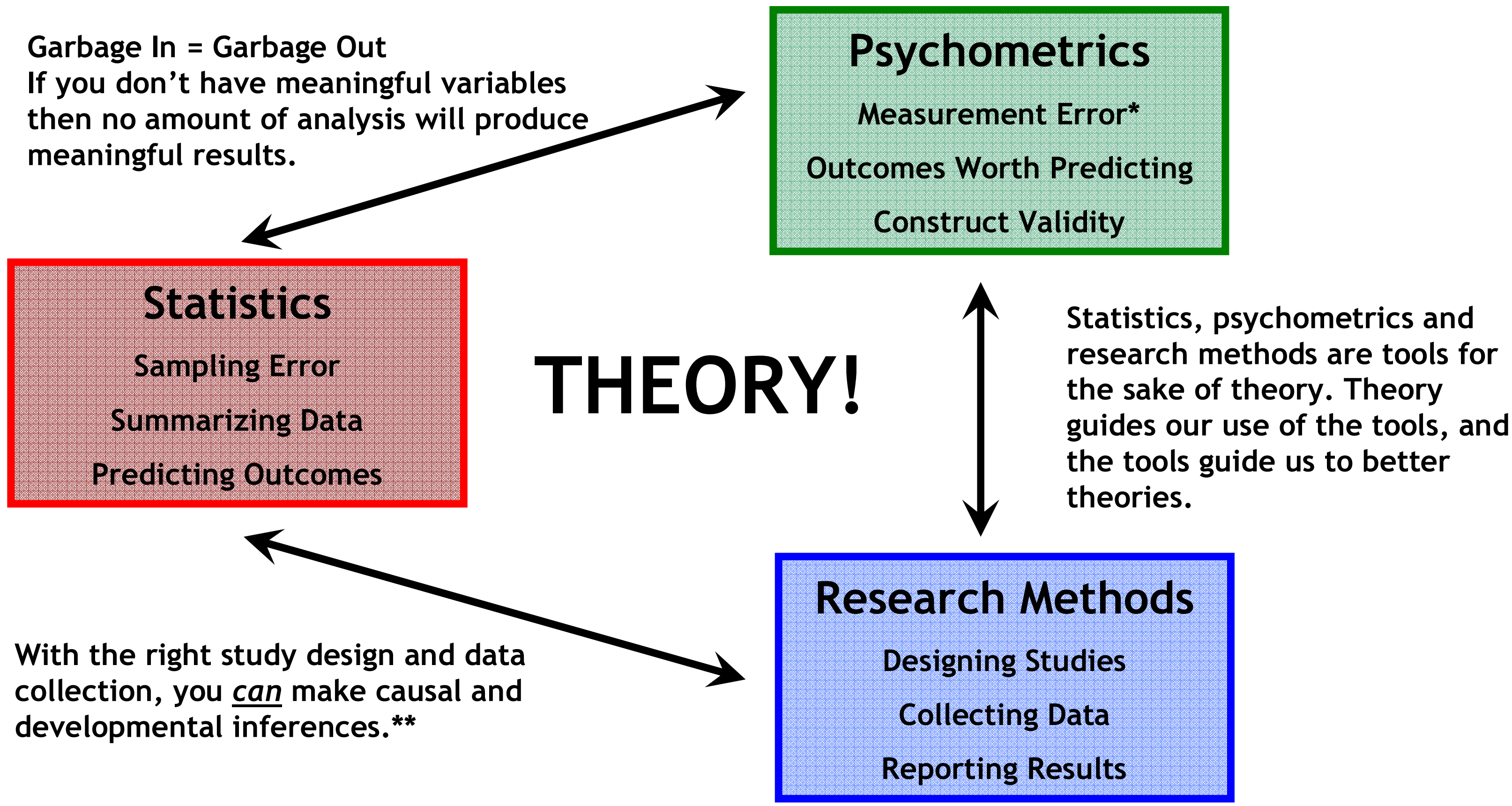
**Ordinal polychotomies, on the other hand, do imply a natural ranking. EDUCATIONAL_ATTAINMENT is an ordinal polychotomy. For data analytic purposes, we may assign a 1 to middle school dropouts, a 2 to high school dropouts, a 3 to high school graduates (but no college), a 4 to some subject with some college but no diploma, a 5 to subjects with a two-year degree, a 6 to subjects with a four-year degree, and a 7 to subjects with a graduate degree. These numbers represent a ranking. A greater number means more of the construct. However, it is ONLY a ranking. We don't claim that a one unit difference means the same amount of construct at each level of the scale. We don't claim that the difference between a 1 and 2 is the same difference as between a 6 and 7 in terms of amount of educational attainment. In other words, we do not claim that the scale is interval. (If the scale were interval, we would could use parametric tests such as regression, ANOVA or T-tests!) Other nominal polychotomies include single-item survey response where 1=Never, 2=Rarely, 3=Often and 4=Always or where 1=Completely Agree, 2=Somewhat Agree and 3=Completely Disagree.**

| Two Ordinal Polychotomies | One Ordinal, One Dichotomy | Two Nominal Polychotomies |
|---|---|---|
| **Spearman's Rank Correlation**<br>This is very much analogous to the Pearson Product Moment Correlation (i.e., the r statistic). | **Wilcoxon Rank Sum Test**<br>This is very much analogous to a two-sample t-test (i.e., regression of a continuous variable on a dichotomy). | **Chi-Square Test** |

**Interval scales provide more information than ordinal scales which in turn provide more information than nominal scales. Tests that incorporate more information will have more statistical power. Statistical power is what you "buy" when you increase your sample size, but you can also "buy" statistical power by using scales that are interval (and reliable).**

# Future Directions: Companion Courses

*Measurement error (aka unreliability) in our outcome adds residual/error variation that we can never predict, but we've been dealing with that all along. Measurement error in our outcome decreases our Pearson correlations, $R^2$ statistics, t-statistics and F-statistics and increases our standard errors. In other words, measurement error in our outcome decreases our statistical power. We have been dealing with measurement error in our outcome, but we have not been dealing with measurement error in our predictors; instead we've been assuming that it has not been enough to make a difference. However, when it does make a difference, and it is hard to tell when it makes a difference, it can bias our results, and it is very difficult to fix. Whereas we want reliable outcomes, we need reliable predictors.




**Experimental data support causal conclusions. Longitudinal data support developmental inferences. Although observational, cross-sectional data (generally) support only correlational inferences, those inferences can be very powerful as part of a healthy research program that feeds on evidence for the sake of dialectical proofs and refutations.

# Future Directions: Dealing With Assumption Violations

- **Homoscedasticity**
  - T-Tests Equal Varainces Not Assumed (Intro Stats-We covered this very briefly.)
  - Robust Standard Errors (Intermediate Stats)
- **Independence**
  - Paired-Samples T-Tests (Intro Stats—We covered this very briefly.)
  - Within Subjects ANOVA (Intermediate Stats)
  - Multi-Level Regression (Intermediate/Advanced Stats)
- **Normality**
  - Non-Linear Transformations (Intermediate Stats)
- **Linearity**
  - Non-Linear Transformations (Intermediate Stats)
  - Non-Linear Regression (Intermediate/Advanced Stats)
- **Outliers**
  - Sensitivity Analysis (Intermediate Stats)

# Future Directions: Multiple Regression

$$READING = \beta_0 + \beta_1 ASIAN + \beta_2 BLACK + \beta_3 LATINO + \beta_4 L2HOMEWORKP1 + \beta_5 ESL + \beta_6 FREELUNCH + \beta_7 ESLxASIAN + \beta_8 ESLxBLACK + \beta_9 ESLxLATINO + \beta_{10} FREELUNCHxASIAN + \beta_{11} FREELUNCHxBLACK + \beta_{12} FREELUNCHxLATINO + \varepsilon$$

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | | 95% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. | Lower Bound | Upper Bound |
| 1 | (Constant) | 48.338 | .110 | | 438.242 | .000 | 48.122 | 48.554 |
| | ASIAN | 1.034 | .383 | .030 | 2.697 | .007 | .283 | 1.786 |
| | BLACK | -4.889 | .339 | -.161 | -14.423 | .000 | -5.554 | -4.225 |
| | LATINO | -4.418 | .306 | -.161 | -14.447 | .000 | -5.017 | -3.818 |
| 2 | (Constant) | 43.878 | .280 | | 156.558 | .000 | 43.328 | 44.427 |
| | ASIAN | .727 | .377 | .021 | 1.929 | .054 | -.012 | 1.465 |
| | BLACK | -4.796 | .333 | -.158 | -14.412 | .000 | -5.448 | -4.144 |
| | LATINO | -4.123 | .301 | -.151 | -13.715 | .000 | -4.712 | -3.534 |
| | L2HOMEWORKP1 | 1.766 | .102 | .188 | 17.254 | .000 | 1.565 | 1.967 |
| 3 | (Constant) | 45.381 | .284 | | 159.528 | .000 | 44.823 | 45.938 |
| | ASIAN | .461 | .441 | .013 | 1.045 | .296 | -.404 | 1.325 |
| | BLACK | -3.622 | .331 | -.119 | -10.956 | .000 | -4.270 | -2.974 |
| | LATINO | -3.311 | .366 | -.121 | -9.035 | .000 | -4.029 | -2.592 |
| | L2HOMEWORKP1 | 1.603 | .100 | .170 | 15.974 | .000 | 1.406 | 1.799 |
| | ESL | .218 | .363 | .009 | .600 | .548 | -.494 | .930 |
| | FREELUNCH | -3.867 | .199 | -.213 | -19.452 | .000 | -4.256 | -3.477 |
| 4 | (Constant) | 45.358 | .288 | | 157.560 | .000 | 44.794 | 45.923 |
| | ASIAN | -.377 | .668 | -.011 | -.564 | .573 | -1.687 | .933 |
| | BLACK | -3.447 | .498 | -.113 | -6.922 | .000 | -4.423 | -2.471 |
| | LATINO | -2.779 | .517 | -.102 | -5.371 | .000 | -3.793 | -1.765 |
| | L2HOMEWORKP1 | 1.591 | .100 | .169 | 15.866 | .000 | 1.394 | 1.788 |
| | ESL | -.876 | .638 | -.035 | -1.373 | .170 | -2.126 | .374 |
| | FREELUNCH | -3.574 | .235 | -.197 | -15.208 | .000 | -4.035 | -3.113 |
| | ESLxASIAN | 3.245 | .999 | .080 | 3.249 | .001 | 1.287 | 5.202 |
| | ESLxBLACK | 5.872 | 1.885 | .036 | 3.115 | .002 | 2.177 | 9.568 |
| | ESLxLATINO | .446 | .858 | .013 | .520 | .603 | -1.235 | 2.127 |
| | FREELUNCHxASIAN | -2.769 | .853 | -.041 | -3.245 | .001 | -4.442 | -1.096 |
| | FREELUNCHxBLACK | -.751 | .666 | -.019 | -1.127 | .260 | -2.058 | .555 |
| | FREELUNCHxLATINO | -.437 | .604 | -.012 | -.724 | .469 | -1.622 | .747 |

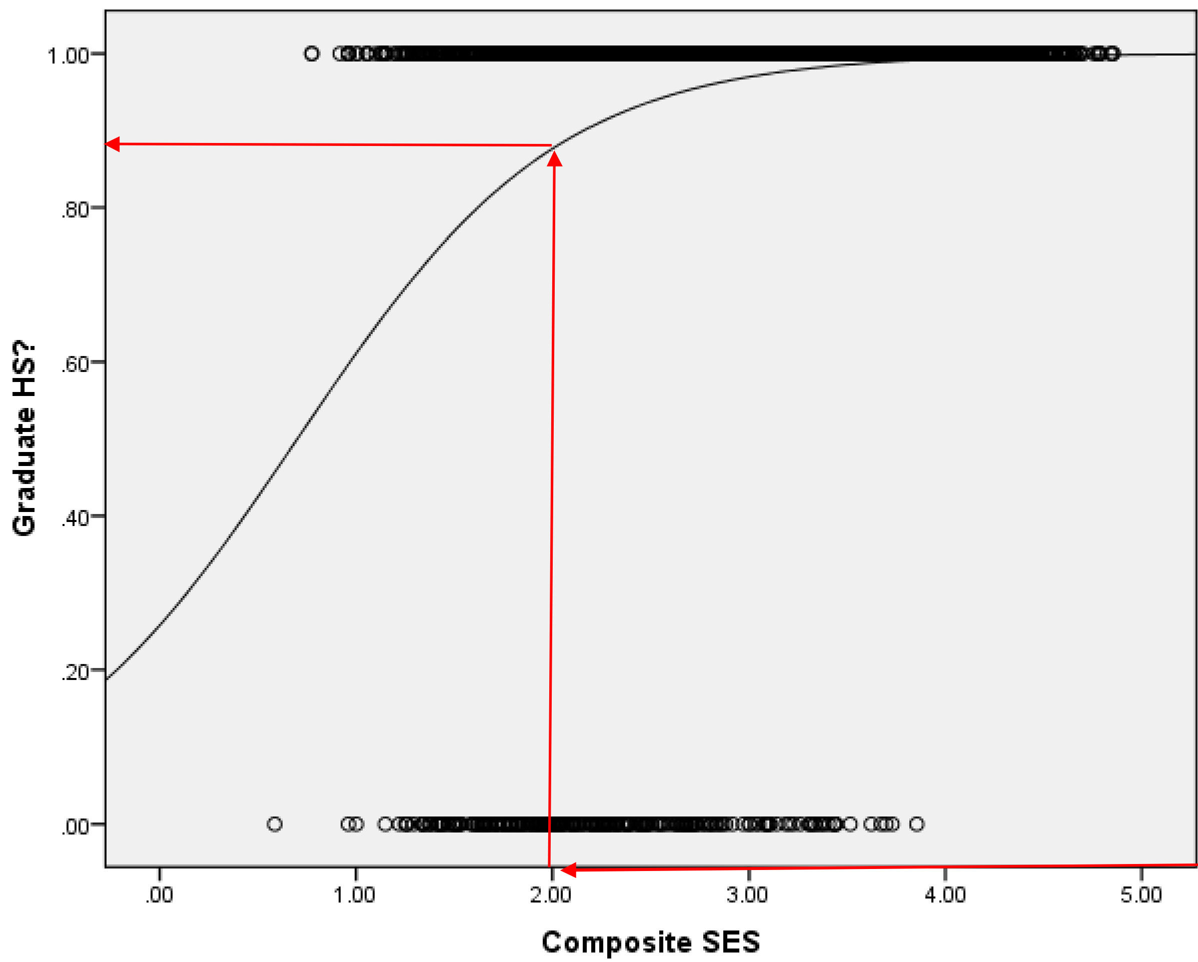
a. Dependent Variable: READING

**You are primed for multiple regression. Your exploratory skills and assumption checking will serve you well in your efforts to carefully construct sensible models. In order to interpret your fitted models, you must use graphs, so all your graph work will pay off. This course was driven to get you there.**

**Race/ethnicity, homework hours, ESL status, and free lunch eligibility (with appropriate interactions) predict 13% of the variation in reading scores.**

**We linearized the (otherwise non-linear) relationship between READING and HOMEWORK by using a logarithmic transformation.**

# Future Directions: Categorical Outcomes with Continuous Predictors




Graph A.X. A bivariate scatterplot of high school graduation versus composite SES with a fitted logistic trend line (n = 4,777).

**You could handle this data by turning SES into a categorical variable and using contingency tables and chi-square tests.**

**Recall that the average of a dichotomous (0/1) variable is the proportion of 1s. Also recall that, with our statistical tools, we are ultimately predicting averages. Thus, when we have a dichotomous outcome and a continuous predictor, we are predicting conditional proportions (or probabilities).**

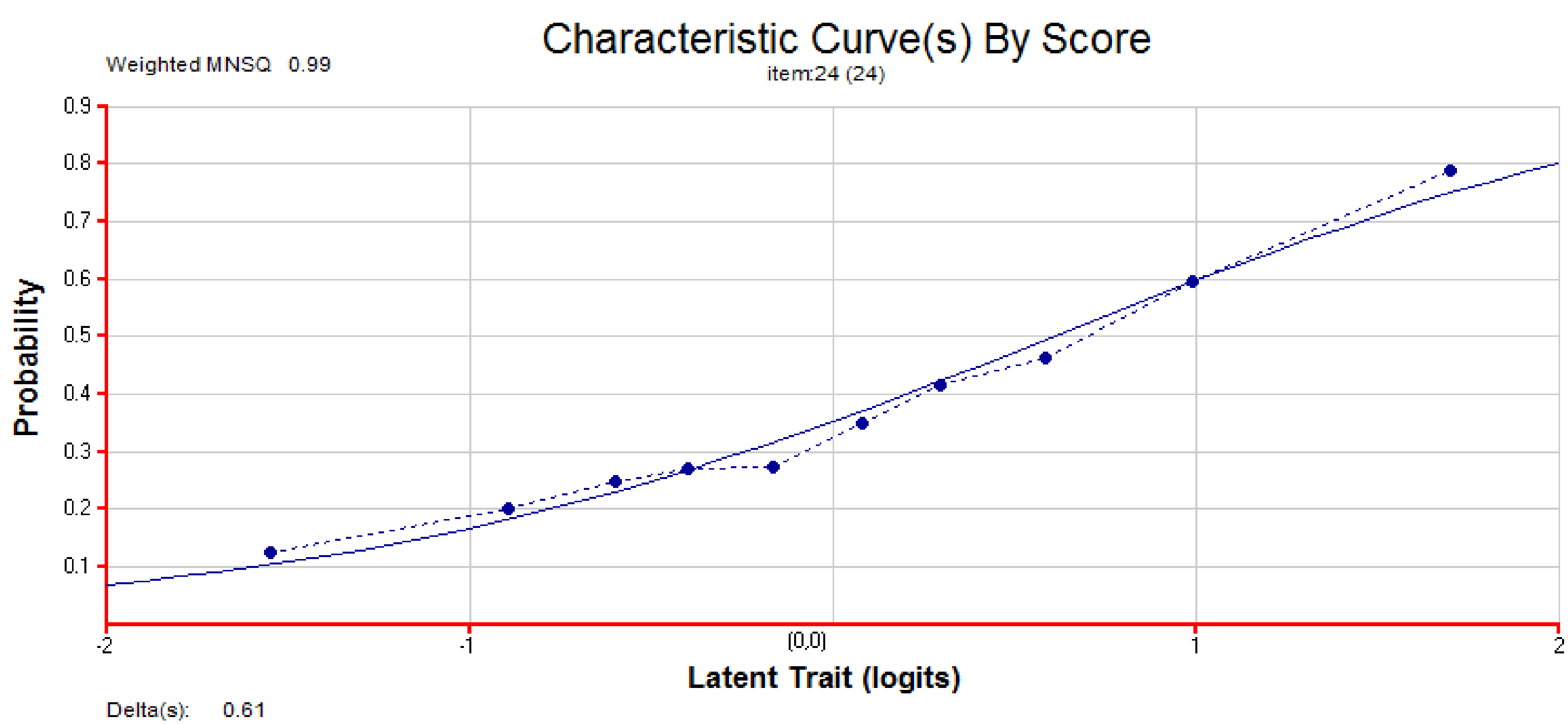
**We want a curve that never goes above Y=1 or below Y=0. The logistic curve (among others) fits the bill.**

**For students of SES=2, what is our predicted graduation rate? 0.88**

**We can use multiple logistic regression to ask if this curve looks different for girls or boys, Black students or White students, Head Start students etc.**

# Future Directions: Item Response Theory (IRT)

This course will largely prepare you for a good class in psychometrics, because most psychometric statistics are regression/correlation based. One use of logistic regression in testing is to map item responses to proficiency levels. Below is an empirical item characteristic curve (dotted line) and an IRT modeled item characteristic curve (smooth line) for Item #24 on an eighth grade algebra test. The latent trait being measured is algebra proficiency. The curves show use that students with a higher proficiency have a higher probability of answering the item correctly. The shape of the curves tell use several things, on of which is the difficulty of the item. We can use the difficulty information for many items to create tests of different items but of the same difficulty.  This "vertical equatability" is very important for longitudinal studies where we want to track growth over time in order to draw developmental conclusions.



# Top Threes For This Course

**Top Three Strengths**

- **Data Analytic Strategy (At Least For Continuous Outcomes)**
  - Theory > Research Questions > Variables (Outcome and Predictors)
  - Exploratory Data Analysis (EDA) > Descriptive Data Analysis > Confirmatory Data Analysis
- **Technical *and* Practical Interpretation**
  - Technical Memos and School Board Memos
- **Visual Methods For EDA, Interpretation and Assumption Checking**
  - Visual methods are easier for many people, but, more importantly, they are often unified and comprehensive (e.g., in simple linear regression one scatterplot does it all).

**Top Three Weaknesses** (Everything you need to know (from an introductory standpoint) is in the slides, but we just did not practice these three things enough. You will be (more than) fine in my next level, but you may have to play catch up in other intermediate statistics courses.)

- **Contingency Table Analysis, Chi-Square Statistics and Other Non-Parametric Stats—these** slides cover the basics. One trick, when you have a larger contingency table (e.g., 2x3), is to break it down by temporarily excluding subgroups from your analysis (e.g., excluding low SES students to make a 2x2 table).
- **Hand Calculations—In** practice, everybody uses software, but teachers like to lean on the pencil.
- **ANOVA: Planned Contrasts and Post Hoc Comparisons—**This is a real weakness. I focused on helping you see the need to dig deeper (the hard part), and I glossed over the methods for digging deeper (the easy part). If, before you begin your analysis, you foresee that you will want to dig deeper, set up planned contrasts, as per the slides. If afterwards, use post hoc comparisons and adjust your alpha level to compensate for all the explicit and implicit statistical tests that you are conducting.

**Top Three Concepts**

- **Sampling Error—**This is *THE* reason for confidence intervals and statistical tests. We recognize that if we took another random sample, we would get different results. We imagine a sampling distribution to get a handle on how different those results might be.
- **Correlation—**Correlation implies neither causation nor development. Rather, correlation implies only that knowing one correlate helps you predict the other correlate.
- **Averages—**We only predict averages, not individuals. Make friends with "tends" and "trends." Report the magnitude of average differences.

**I can't promise that this course will take you every step of the way for every data analytic journey, but I can promise that you will get off on the right foot for most projects. Furthermore, when you come to road blocks, you will be able to communicate effectively with data analysts. If you ask a data analyst to analyze your data, she will roll her eyes. If you ask a data analyst to help you with your heteroskedasticity problem, her eyes will light up.**

# Answering our Roadmap Question

## Appendix A: In the population, is there a relationship between race and free lunch?

**FREELUNCH * R^S RACE/ETHNIC BACKGROUND Crosstabulation**

| | | | R^S RACE/ETHNIC BACKGROUND | | | | |
|---|---|---|---|---|---|---|---|
| | | | Asian | Latino | Black | White | Total |
| FREELUNCH | 0 | Count | 391 | 400 | 279 | 4114 | 5184 |
| | | Expected Count | 344.3 | 570.9 | 451.9 | 3816.9 | 5184.0 |
| | | Std. Residual | 2.5 | -7.2 | -8.1 | 4.8 | |
| | 1 | Count | 127 | 459 | 401 | 1629 | 2616 |
| | | Expected Count | 173.7 | 288.1 | 228.1 | 1926.1 | 2616.0 |
| | | Std. Residual | -3.5 | 10.1 | 11.5 | -6.8 | |
| | Total | Count | 518 | 859 | 680 | 5743 | 7800 |
| | | Expected Count | 518.0 | 859.0 | 680.0 | 5743.0 | 7800.0 |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 4.377E2 | 3 | .000 |
| Likelihood Ratio | 417.313 | 3 | .000 |
| Linear-by-Linear Association | 92.880 | 1 | .000 |
| N of Valid Cases | 7800 | | |

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 173.73.

**In our nationally representative sample of 7,800 8th graders, there is a statistically significant relationship between race/ethnicity and free lunch eligibility $\chi^2(3)=437.7$, $p < .001$. White and Asian students are underrepresentated among free lunch students, and Black and Latino students are over-represented.**

# Appendix A Appendix: Key Interpretations

- The relationship between sex and initial endorsement is not statistically significant, $\chi^2 = 0.039$, $p = 0.843$.
- We find a statistically significant relationship between ethnicity and socioeconomic status, $\chi^2 = 94.9$, $p < 0.001$. Latino students are overrepresented among low SES students and underrepresented among high SES students; 37% of Latino students are low SES as compared to only12% of Anglo students.

# Appendix A Appendix : Key Concepts

- If the standardized residual is greater than |2|, there is considerable over/ underrepresentation in that cell.
- Different cells make different chi-square contributions. We might say “the action” is in the cells that make the biggest chi-square contributions.
- When you have fewer than 5 expected observations in a cell, your chi-square statistic becomes questionable.
- Besides the independence assumption, there are no “HINLO” assumptions to worry about. This is not really a strength! You pay for it in statistical power. Whenever appropriate, work with continuous variables.
- If the relationship is statistically significant, look for large (greater than |2|) standardized residuals to see where the action is. Note that if there is overrepresentation in one cell, there must be underrepresentation in at least two other cells; this is deeply related to the concept of degrees of freedom.
- If you have a categorical outcome, you can deal with it by transforming your predictors into categorical variables.

# Appendix A Appendix : Key Terminology

- In contingency tables, “expected” (by default) means the expected frequency if the null hypothesis were true. It is our baseline for comparison. However, you can change expected frequency to whatever you want.
- The observed Count is the number of people in that category.
- The Expected Count is the number of people who  we would expect in that category if there were no relationship in the population.
- The Standardized Residual is the residual (Observed Count minus Expected Count) divided by a standard deviation.
- The chi-square statistic is a non-parametric statistic; it does not rely on averages, which frees it from assumptions about homoscedasticity, normality, linearity and outliers. Non-parametric statistics instead rely on counts and ranks. The cost of these nearly-assumption-free statistics is that they cannot take advantage of relative differences. If I am in a foot race with nine Olympic runners, non-parametric stats only care that I came in last place. They ignore (perhaps useful) information about how far behind I finished.
- There are two types of polychotomies: nominal and ordinal. Nominal polychotomies represent categories with no natural ranking implied.  RACE/ETHNICITY is a nominal polychotomy. For data analytic purposes, we may assign White=1, Black=2, Latino=3 and Asian=4, but those numbers are merely numerical labels, not ranks. Other nominal polychotomies are RELIGION (Christian, Jewish, Muslim…), MARITALSTATUS (Single, Married, Divorced, Widowed), MUSICGENRE (Rock, Rap, Country, Classical, Jazz, R&B).
- Ordinal polychotomies, on the other hand, do imply a natural ranking. EDUCATIONAL_ATTAINMENT is an ordinal polychotomy. For data analytic purposes, we may assign a 1 to middle school dropouts, a 2 to high school dropouts, a 3 to high school graduates (but no college), a 4 to some subject with some college but no diploma, a 5 to subjects with a two-year degree, a 6 to subjects with a four-year degree, and a 7 to subjects with a graduate degree. These numbers represent a ranking. A greater number means more of the construct. However, it is ONLY a ranking.  We don’t claim that a one unit difference means the same amount of construct at each level of the scale. We don’t claim that the difference between a 1 and 2 is the same difference as between a 6 and 7 in terms of amount of educational attainment. In other words, we do not claim that the scale is interval. (If the scale were interval, we would could use parametric tests such as regression, ANOVA or T-tests!) Other nominal polychotomies include single-item survey response where 1=Never, 2=Rarely, 3=Often and 4=Always or where 1=Completely Agree, 2=Somewhat Agree and 3=Completely Disagree.

# Appendix A Appendix: SPSS Syntax

```
**************************************************************************.
*Contingency Table with Chi Square Test.
**************************************************************************.
CROSSTABS
  /TABLES=Female BY InitialEndorsementYesNo
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT EXPECTED SRESID
  /COUNT ROUND CELL.
```

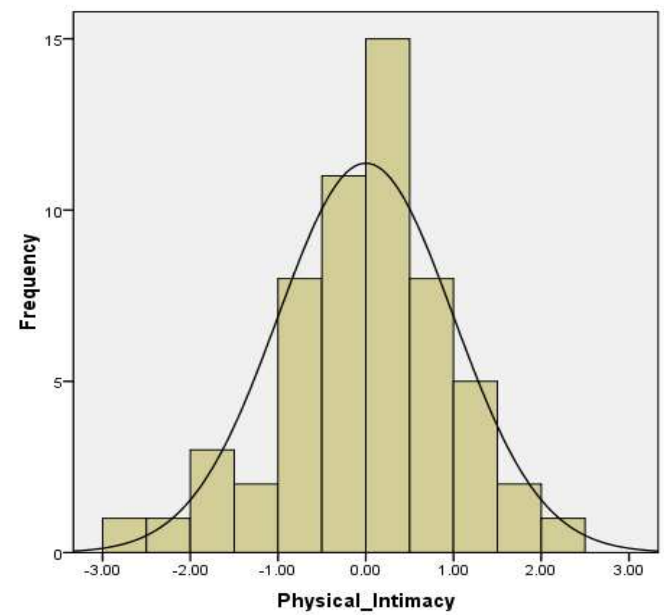
# Perceived Intimacy of Adolescent Girls (Intimacy.sav)

- **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.**
- **Note on Physical_Intimacy (with boyfriend): This is a composite variable based on a principle components analysis. Girls who score high on Physical_Intimacy scored high on (1) Physical Affection and (2) Mutual Caring, but low on (3) Risk Vulnerability and (4) Resolve Conflicts, regardless of (5) Trust and (6) Self Disclosure.**
- **Variables:**

(Physical_Intimacy)
Physical Intimacy With Boyfriend—see above

(RiskVulnerabilityWMom)
1=Tend to Risk Vulnerability with Mom, 0=Not

(ResolveConflictWMom)
1=Tend to Resolve Conflict with Mom, 0=Not



# Perceived Intimacy of Adolescent Girls (Intimacy.sav)

**RiskVulnerabilityWMom * ResConflictWMom Crosstabulation**

| | | | ResConflictWMom | | |
|---|---|---|---|---|---|
| | | | 0 | 1 | Total |
| RiskVulnerabilityWMom | 0 | Count | 21 | 11 | 32 |
| | | Expected Count | 15.0 | 17.0 | 32.0 |
| | | Std. Residual | 1.5 | -1.5 | |
| | 1 | Count | 9 | 23 | 32 |
| | | Expected Count | 15.0 | 17.0 | 32.0 |
| | | Std. Residual | -1.5 | 1.5 | |
| | Total | Count | 30 | 34 | 64 |
| | | Expected Count | 30.0 | 34.0 | 64.0 |

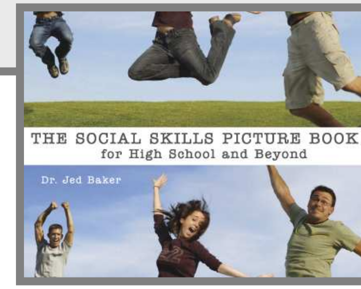
**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
|---|---|---|---|---|---|
| Pearson Chi-Square | 9.035[a] | 1 | .003 | | |
| Continuity Correction[b] | 7.592 | 1 | .006 | | |
| Likelihood Ratio | 9.265 | 1 | .002 | | |
| Fisher's Exact Test | | | | .005 | .003 |
| Linear-by-Linear Association | 8.894 | 1 | .003 | | |
| N of Valid Cases | 64 | | | | |

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 15.00.

b. Computed only for a 2x2 table

# High School and Beyond (HSB.sav)

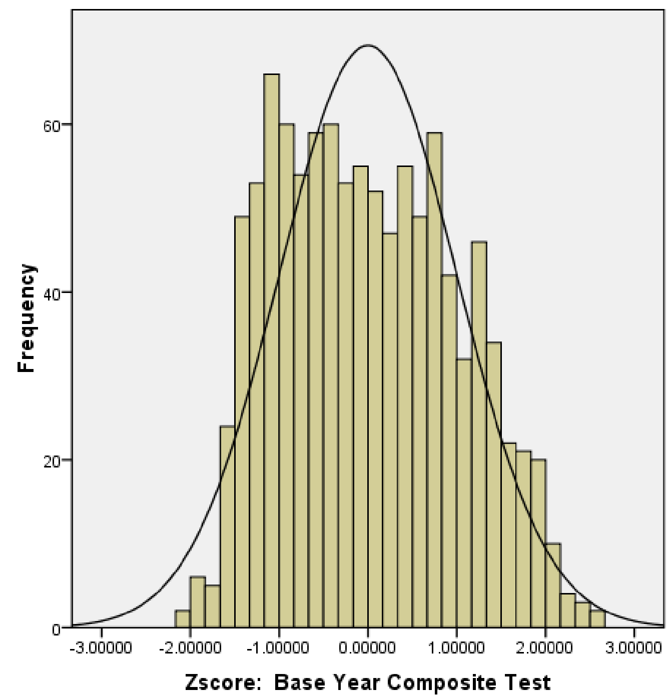



- **Overview: High School & Beyond – Subset of data focused on selected student and school characteristics as predictors of academic achievement.**
- **Source: Subset of data graciously provided by Valerie Lee, University of Michigan.**
- **Sample: This subsample has 1044 students in 205 schools. Missing 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:**

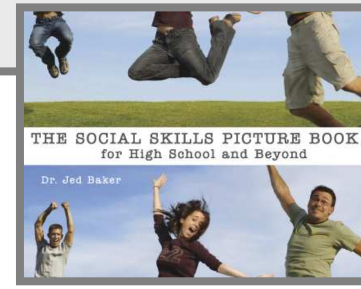
| | |
|---|---|
| (ZBYTest) | Standardized Base Year Composite Test Score |
| (Sex) | 1=Female, 0=Male |
| (RaceEthnicity) | Students Self-Identified Race/Ethnicity<br>1=White/Asian/Other, 2=Black, 3=Latino/a |

Dummy Variables for RaceEthnicity:

| | |
|---|---|
| (Black) | 1=Black, 0=Else |
| (Latin) | 1=Latino/a, 0=Else |



# High School and Beyond (HSB.sav)




**1 = Female, 0 = Other * RaceEthnicity Crosstabulation**

| | | | RaceEthnicity | | | |
|---|---|---|---|---|---|---|
| | | | White/Asian/Other | Black | Latino/a | Total |
| 1 = Female, 0 = Other | Male | Count | 202 | 113 | 150 | 465 |
| | | Expected Count | 193.3 | 133.2 | 138.5 | 465.0 |
| | | Std. Residual | .6 | -1.7 | 1.0 | |
| | Female | Count | 232 | 186 | 161 | 579 |
| | | Expected Count | 240.7 | 165.8 | 172.5 | 579.0 |
| | | Std. Residual | -.6 | 1.6 | -.9 | |
| | Total | Count | 434 | 299 | 311 | 1044 |
| | | Expected Count | 434.0 | 299.0 | 311.0 | 1044.0 |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 7.932[a] | 2 | .019 |
| Likelihood Ratio | 7.996 | 2 | .018 |
| Linear-by-Linear Association | .043 | 1 | .836 |
| N of Valid Cases | 1044 | | |

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 133.18.

# 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) | A Measure of Understanding of Illness Causality |
| (SocioEconomicStatus) | 1=Low SES, 2=Lower Middle, 3=Upper Middle 4 = High SES |
| (HealthStatus) | 1=Healthy, 2=Asthmatic 3=Diabetic |

Dummy Variables for SocioEconomicStatus:

| | |
|---|---|
| (LowSES) | 1=Low SES, 0=Else |
| (LowerMiddleSES) | 1=Lower MiddleSES, 0=Else |
| (HighSES) | 1=High SES, 0=Else |

*Note that we will use SocioEconomicStatus=3, Upper Middle SES, as our reference category.

Dummy Variables for HealthStatus:

| | |
|---|---|
| (Asthmatic) | 1=Asthmatic, 0=Else |
| (Diabetic) | 1=Diabetic, 0=Else |

*Note that we will use HealthStatus=1, Healthy, as our reference category.

# Understanding Causes of Illness (ILLCAUSE.sav)

**SocioEconomicStatus * HealthStatus Crosstabulation**

| | | | HealthStatus | | | |
|---|---|---|---|---|---|---|
| | | | Healthy | Asthmatic | Diabetic | Total |
| SocioEconomicStatus | Low SES | Count | 1 | 15 | 9 | 25 |
| | | Expected Count | 11.7 | 8.9 | 4.4 | 25.0 |
| | | Std. Residual | -3.1 | 2.0 | 2.2 | |
| | Lower Middle SES | Count | 16 | 24 | 10 | 50 |
| | | Expected Count | 23.4 | 17.8 | 8.8 | 50.0 |
| | | Std. Residual | -1.5 | 1.5 | .4 | |
| | Upper Midde SES | Count | 40 | 31 | 14 | 85 |
| | | Expected Count | 39.8 | 30.3 | 14.9 | 85.0 |
| | | Std. Residual | .0 | .1 | -.2 | |
| | High SES | Count | 39 | 3 | 3 | 45 |
| | | Expected Count | 21.1 | 16.0 | 7.9 | 45.0 |
| | | Std. Residual | 3.9 | -3.3 | -1.7 | |
| | Total | Count | 96 | 73 | 36 | 205 |
| | | Expected Count | 96.0 | 73.0 | 36.0 | 205.0 |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 52.436[a] | 6 | .000 |
| Likelihood Ratio | 60.627 | 6 | .000 |
| Linear-by-Linear Association | 38.137 | 1 | .000 |
| N of Valid Cases | 205 | | |

a. 1 cells (8.3%) have expected count less than 5. The minimum expected count is 4.39.

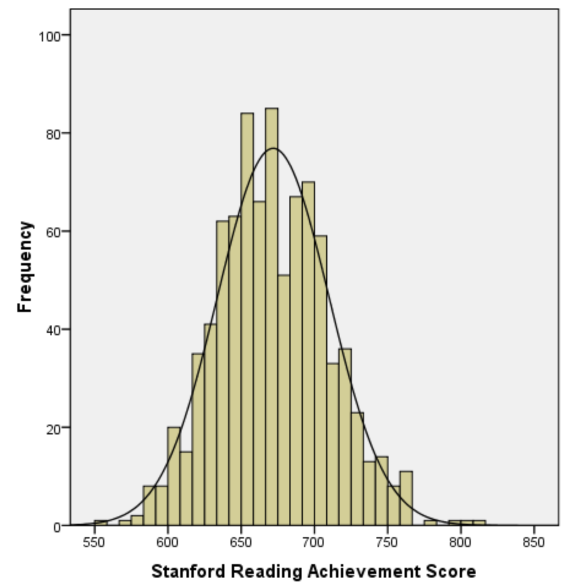
# Children of Immigrants (ChildrenOfImmigrants.sav)

- **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 Scores |
| (Depressed) | 1=The Student is Depressed, 0=Not Depressed |
| (SESCat) | A Relative Measure Of Socio-Economic Status |
| | 1=Low SES, 2=Mid SES, 3=High SES |

Dummy Variables for SESCat:

| | |
|---|---|
| (LowSES) | 1=Low SES, 0=Else |
| (MidSES) | 1=Mid SES, 0=Else |
| (HighSES) | 1=High SES, 0=Else |



# Children of Immigrants (ChildrenOfImmigrants.sav)

**Depressed * SESCat Crosstabulation**

| | | | SESCat | | | |
|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | Total |
| Depressed | 0 | Count | 113 | 538 | 115 | 766 |
| | | Expected Count | 121.0 | 531.8 | 113.2 | 766.0 |
| | | Std. Residual | -.7 | .3 | .2 | |
| | 1 | Count | 26 | 73 | 15 | 114 |
| | | Expected Count | 18.0 | 79.2 | 16.8 | 114.0 |
| | | Std. Residual | 1.9 | -.7 | -.4 | |
| | Total | Count | 139 | 611 | 130 | 880 |
| | | Expected Count | 139.0 | 611.0 | 130.0 | 880.0 |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 4.857[a] | 2 | .088 |
| Likelihood Ratio | 4.456 | 2 | .108 |
| Linear-by-Linear Association | 3.186 | 1 | .074 |
| N of Valid Cases | 880 | | |

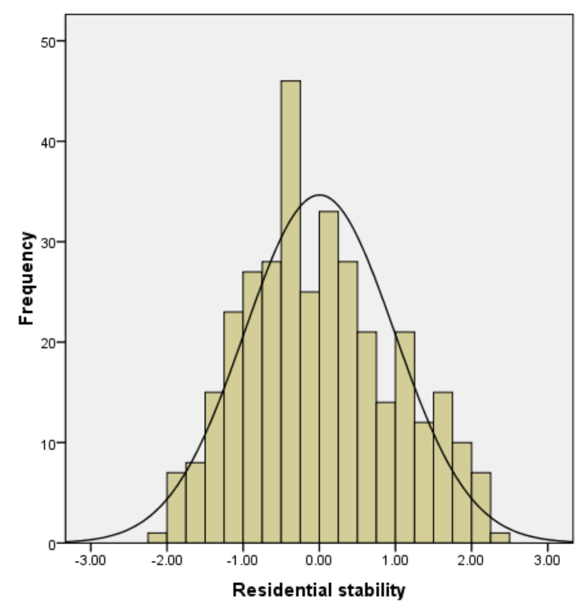
a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 16.84.

# 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:**

(ResStab) Residential Stability, A Measure Of Neighborhood Flux
(NoMurder95) 1=No Murders in 1995, 0=At Least One Murder in 1995
(SES) A Relative Measure Of Socio-Economic Status
1=Low SES, 2=Mid SES, 3=High SES

Dummy Variables for MothEdCat:
(LowSES) 1=Low SES, 0=Else
(MidSES) 1=Mid SES, 0=Else
(HighSES) 1=High SES, 0=Else



# Human Development in Chicago Neighborhoods (Neighborhoods.sav)

**NoMurder95 * SES Crosstabulation**

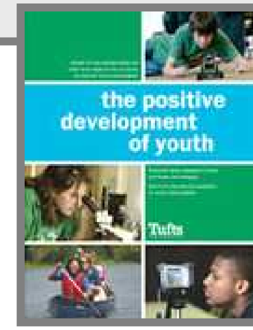
| | | | SES | | | |
|---|---|---|---|---|---|---|
| | | | Low SES | Mid SES | High SES | Total |
| NoMurder95 | At Least One Murder in 1995 | Count | 85 | 99 | 48 | 232 |
| | | Expected Count | 66.5 | 82.1 | 83.4 | 232.0 |
| | | Std. Residual | 2.3 | 1.9 | -3.9 | |
| | No Murders in 1995 | Count | 13 | 22 | 75 | 110 |
| | | Expected Count | 31.5 | 38.9 | 39.6 | 110.0 |
| | | Std. Residual | -3.3 | -2.7 | 5.6 | |
| | Total | Count | 98 | 121 | 123 | 342 |
| | | Expected Count | 98.0 | 121.0 | 123.0 | 342.0 |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 73.680[a] | 2 | .000 |
| Likelihood Ratio | 73.624 | 2 | .000 |
| Linear-by-Linear Association | 60.708 | 1 | .000 |
| N of Valid Cases | 342 | | |

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 31.52.

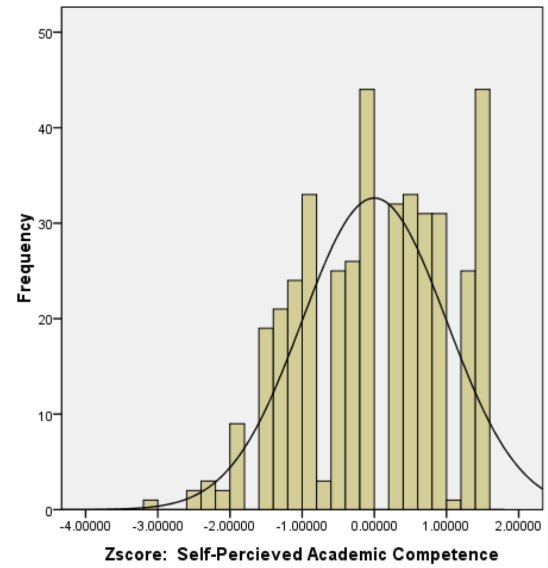
# 4-H Study of Positive Youth Development (4H.sav)




- **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:**

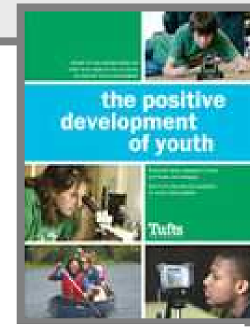
| | |
|---|---|
| (ZAcadComp) | Standardized Self-Perceived Academic Competence |
| (SexFem) | 1=Female, 0=Male |
| (MothEdCat) | Mother's Educational Attainment Category<br>1=High School Dropout, 2=High School Graduate,<br>3 =Up To 3 Years of College, 4 = 4-Plus Years of College |




Dummy Variables for MothEdCat:

| | |
|---|---|
| (MomHSDropout) | 1=High School Dropout, 0=Else |
| (MomHSGrad) | 1=High School Graduate, 0=Else |
| (MomUpTo3YRSCollege) | 1=Up To 3 Years of College, 0=Else |
| (Mom4plusYRSCollege) | 1=4-Plus Years of College, 0=Else |

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




**Female = 1, Male = 0 * MothEdCat Crosstabulation**

| | | | MothEdCat | | | | |
|---|---|---|---|---|---|---|---|
| | | | Mom HS Dropout | Mom HS Grad | Mom Up to 3 YRS College | Mom 4+ YRS College | Total |
| Female = 1, Male = 0 | Boy | Count | 4 | 33 | 68 | 60 | 165 |
| | | Expected Count | 7.3 | 44.4 | 62.9 | 50.4 | 165.0 |
| | | Std. Residual | -1.2 | -1.7 | .6 | 1.3 | |
| | Girl | Count | 14 | 77 | 88 | 65 | 244 |
| | | Expected Count | 10.7 | 65.6 | 93.1 | 74.6 | 244.0 |
| | | Std. Residual | 1.0 | 1.4 | -.5 | -1.1 | |
| | Total | Count | 18 | 110 | 156 | 125 | 409 |
| | | Expected Count | 18.0 | 110.0 | 156.0 | 125.0 | 409.0 |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 11.074[a] | 3 | .011 |
| Likelihood Ratio | 11.402 | 3 | .010 |
| Linear-by-Linear Association | 10.225 | 1 | .001 |
| N of Valid Cases | 409 | | |

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 7.26.