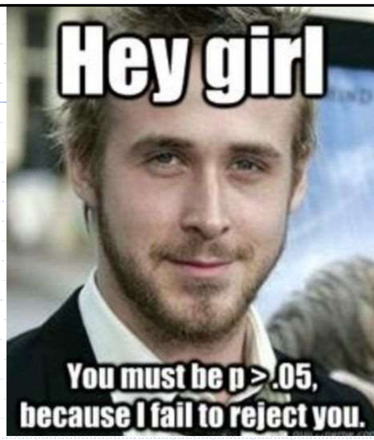


1

SOC497/L: SOCIOLOGY RESEARCH METHODS

The Three Key Interval Tests:

S l o w l y … at first ☺

Ellis Godard

Which is more entertaining?

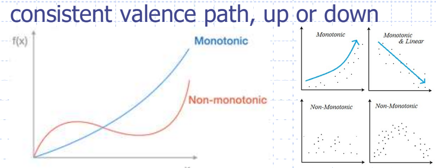
1. Facebook friends 53%
2. Fox & Friends 0%
3. Phil Lesh & Friends 7%
4. Shicker & Friends 7%
5. Thomas & Friends 33%

SOC497 @ CSUN w/ Ellis Godard 3

20

Interlude: Monotonic

- Thanks to… um… who taught me this word? ☺
- In math, it means something that either never increases or never decreases – it has a consistent valence path, up or down



- So, in survey wording, I guess it means that the values are in order?

SOC497 @ CSUN w/ Ellis Godard 4

Q1. How many basic statistical tests are typically used in Sociology?

1. one 0%
2. two 0%
3. three 11%
4. four (x^2, t, F, & r^2) 76%
5. five 14%

SOC497 @ CSUN w/ Ellis Godard 5

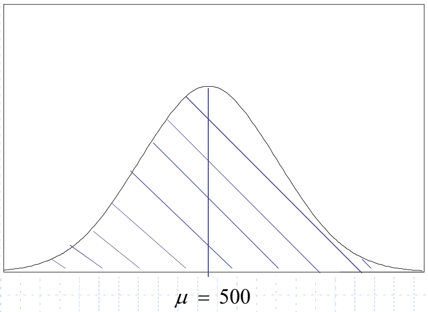
0

Outline for Today

- Hold on tight… ☺
- Review: Areas, Distributions, Errors, & Ps
- Hypotheses & the Other Three Tests
  - t-testing a mean across 2 groups
  - ANOVA (F) – means in more than 2 groups
  - Regression (r^2) – means across interval values
- Substantive vs. Statistical Sig. (time permitting)
- Demo (in lab?) & Lab Exercise

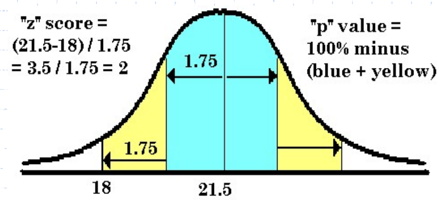
SOC497 @ CSUN w/ Ellis Godard 6

Entire Curve = 100%

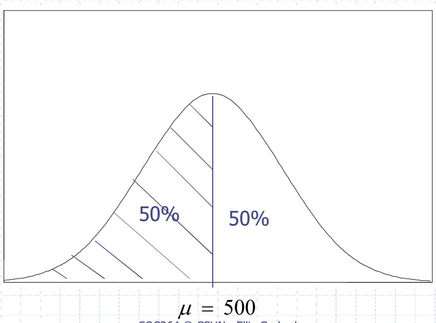


Remembering Z's & p's

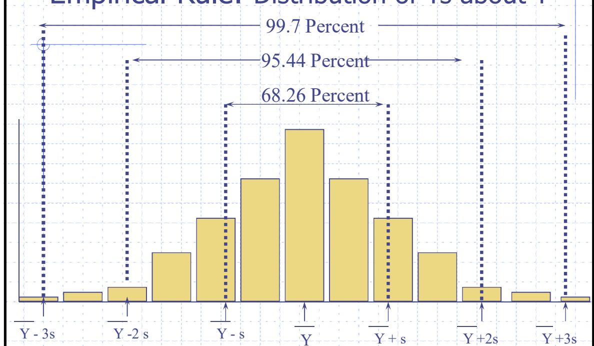
- Someone aged 18 is 2 standard deviations from the mean – and an estimated 95.44% of the values under the curve are in the same interval



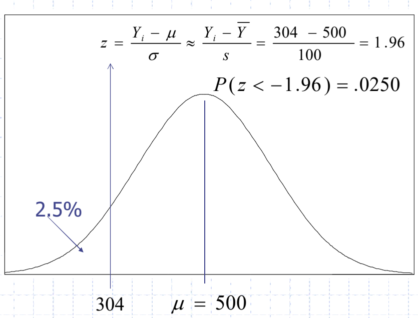
Median/Mean/Mode splits it 50/50



Empirical Rule: Distribution of Ys about Y



There's some p beyond any z (distance, in stdevs)



4 Kinds of Distributions

- ***Population Distribution:*** "Real" arrangement of a variable's data, in the population being studied
- ***Sample Distribution:*** Arrangement (as illustrated in a histogram) of the data actually collected or observed
- ***Normal Distribution:*** An arrangement w/ particular shape characteristics (similar to a "bell curve", but much more specific)
- ***Sampling Distribution:*** The distribution of sample statistics, such as sample means (that is, the collection of all of the means of all possible samples, for a given variable and population)

Illustrate w/ a Small Population

- **3 students as a population**
- **Variable is satisfaction w/ class**
- **N=3, with values of 1, 2, and 3**
  - Mu = 2
- **3 possible samples of 2:** 1,2 1,3 2,3
  - Means of those samples: 1.5 2 2.5
  - Mean of those means: 2

Illustrate w/ Larger Population

- **N=4, with values of 1, 2, 3, and 4**
  - Mu = 2.5
- **6 possible samples of 2:** 12 13 23 14 24 34
  - Means of those samples: **1.5 2 2.5 2.5 3 3.5**
  - Mean of those means: 2.5
- **4 possible samples of 3:** 123 124 134 234
  - Means of those samples: 2 2.33 2.66 3
  - Mean of those means: 2.5
  - *Notice that the sample means are closer together*
    - Smaller std. deviation (which, for sample means, is a **standard error**)

Two *Sampling* Distributions

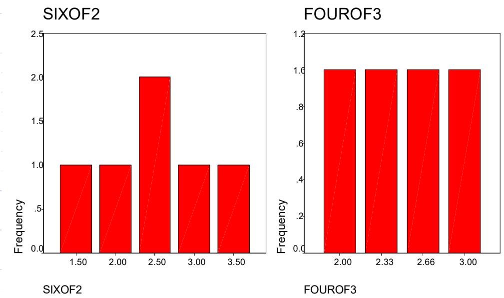
SIXOF2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.50	1	16.7	16.7	16.7
	2.00	1	16.7	16.7	33.3
	2.50	2	33.3	33.3	66.7
	3.00	1	16.7	16.7	83.3
	3.50	1	16.7	16.7	100.0
	Total	6	100.0	100.0	

FOUROF3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2.00	1	16.7	25.0	25.0
	2.33	1	16.7	25.0	50.0
	2.66	1	16.7	25.0	75.0
	3.00	1	16.7	25.0	100.0
	Total	4	66.7	100.0	
Missing	System	2	33.3		
Total		6	100.0		

Sampling Distribution Histograms



Possible # of Samples HUGE!

For a population size (N) o 26, how many different samples of size n=10 are there?

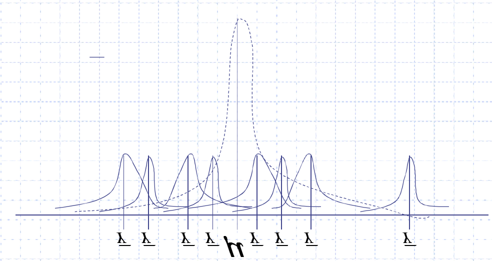
Number of Permutations:

$$\frac{26!}{(26-10)!} = 26 \times 25 \times 24 \times 23 \times 22 \times 21 \times 20 \times 19 \times 18 \times 17 = 1.927 \times 10^{13}$$

Number of Combinations (Samples):

$$\frac{26!}{10!(26-10)!} = 312{,}455$$

Mean of Sampling Means = "Mu"



Q2. An imagined collection of every possible sample is a __ distribution.

1. Normal
   0%
2. Population
   16%
3. Random
   9%
4. Sample
   3%
5. Sampling
   72%

0

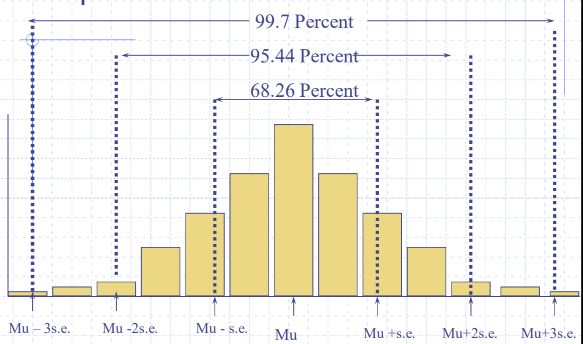
SOC497 @ CSUN w/ Ellis Godard 19

The Purpose of Null Hypotheses

- Research Hypothesis (Ha) is what you expect ("want") to find
- Null hypothesis (Ho) is that the Ha is wrong
  - Univariate: actual value is something else
  - Bivariate: difference between groups (sumsamples) is 0
- Null hypothesis value is a paper tiger – a pretend middle
  - The # in your sample will differ, if only due to sampling variation
- Want to show null *couldn't* be right given what we've seen
  - That is, hoping to find a number very different from the null
- The more different the sample is, the less convincing the null is
  - The farther apart they are, the less chance you could find

SOC497 @ CSUN w/ Ellis Godard 22

Empirical Rule: Distribution of $\bar{Y}$s about "Mu"



The Pattern of Statistical Tests

- All based on some sort of (standard) error

$$\frac{some\ \# - someother\ \#}{someerror}$$

- All generate probability of getting a difference larger than the one observed in the sample
  - Test ("for significance") the difference between 2 #s
    - usually an observed stat & some expected value
    - May be the value for two groups (mean for F, mean for M)
  - Divide by some (standard) error
- All follow same 5 steps, & result in a p-value

SOC497 @ CSUN w/ Ellis Godard 23

Q3. The standard deviation of a sampling distribution is a…

1. Sampling deviation
   0%
2. Sampling error
   2%
3. Standard deviation
   2%
4. Standard error
   95%
5. Something else
   0%

0

SOC497 @ CSUN w/ Ellis Godard 21

Four Basic Statistical Tests

- "Chi-square"
  - DV & IV both categorical (nominal or ordinal)
    - Test doesn't distinguish between DV & IV
    - But convention is DV is rows, IV in columns
  - Also need a PRE/MOA
    - If both ordinal, also use Gamma
    - If either is nominal w/ 3+ groups, also use Lambda
- "T-test"
  - DV is interval
  - IV is categorical, with two categories

SOC497 @ CSUN w/ Ellis Godard 24

Four Basic Statistical Tests, cont'd

- ANOVA ("F test")
  - DV is interval
  - IV is categorical, usually w/ >2 categories
- Regression ("r" correlation) for anything
  - OLS/GLM: several interval variables
  - Multiple regression:
    - no theoretical limit (but practical & empirical)
  - Categoricals – easiest as "dummy" variables
    - Other forms: logistic, et al.
  - Interaction terms

Recall Testing a Mean…

- **Z & T use same basic test score formula**

$$\frac{ObsMean - NullMean}{stderror} = \frac{\overline{Y} - \mu_{H_0}}{\hat{\sigma}_{\overline{Y}}} = \frac{\overline{Y} - \mu_{H_0}}{\sigma_Y / \sqrt{n}}$$

  - Use t if n<30; use z if n>30
  - t distribution approximately normal (z) at 30
    - higher kurtosis below 30 - e.g. df=6 or df=2
- **Is this class significantly older than 21, on avg.?**
  - 22.2 – 21 / 2.449 = 1.2 / 2.449 = .49 stderrors
  - That's the *calculated t* – how far observation is from null
  - There's some p-value associated with it – area beyond curve
  - SPSS will give it to you, or
    Compare to *tabular/critical t* – # stderrors assoc w/ p you need

The t-distribution

If we assume that the *population* distribution of a variable is normal, then for a random sample of *any* size n,

$$t = \frac{\overline{Y} - \hat{\mu}}{\hat{\sigma}_{\overline{Y}}}$$

is called the (Student) t distribution, with (n - 1) degrees of freedom.

Testing a *Difference* of Means

- **Same idea as other tests**
  - Test the null of no difference
  - Get a t value associated w/ some p value
  - P = probability of being wrong in rejecting the null
  - We want that probability to be a small as possible
    - We want to reject the null hypothesis
    - We don't want to be wrong when we do that
    - So we want the risk of being wrong to be really small
- **One statistical difference:**
  - Sampling distribution is of *differences in means*
    - Rather than a sampling distribution of means (t or z)
    - Table in the book is like the chi-square table (later)

Diff. of Means Hypotheses

Usually, we are interested in whether or not difference between 2 parameters = 0.

For a two-tail test,

$$H_o: \mu_1 = \mu_2 \qquad H_a: \mu_1 \neq \mu_2 \qquad \text{so} \qquad H_o: \mu_1 - \mu_2 = 0 \qquad H_a: \mu_1 - \mu_2 \neq 0$$

For a one-tail test,

$$H_o: \mu_1 = \mu_2 \qquad H_a: \mu_1 < \mu_2 \qquad \text{so} \qquad H_o: \mu_1 - \mu_2 = 0 \qquad H_a: \mu_1 - \mu_2 < 0$$

Example: Educ. & Children

Would you expect college-educated adults to have more or fewer children than adults without a college education? Why?

$H_o: \mu_{Coll} = \mu_{noColl}$ **(No difference in family size)**

$H_a: \mu_{Coll} < \mu_{noColl}$ **(College-educated have fewer children)**

GSS Results

(College Educated = 16 or more years of schooling)

	SampleSize	Mean	StdDev	$(EstStdErr)^2$
Not College	1168	2.05	1.800	$(1.800)^2 / 1168$
College	334	1.35	1.487	$(1.487)^2 / 334$
Difference		0.7045		$\frac{(1.800)^2}{1168} + \frac{(1.487)^2}{334}$

Test of the Difference

$$\text{standard error} = \sigma_{\overline{Y_2} - \overline{Y_1}} = \sqrt{\frac{(1.800)^2}{1168} + \frac{(1.487)^2}{334}} = 0.096$$

test statistic

$$z = \frac{\overline{Y_2} - \overline{Y_1}}{\sigma_{\overline{Y_2} - \overline{Y_1}}} = \frac{0.7045}{0.096} = 7.27$$

The p-value that corresponds to this test is so small that it is not even listed in Table A. We can reject H_0 and find support for the claim that college-educated individuals have fewer children than non-college educated individuals.

T-test in SPSS

- **Use Independent Samples T-test**
  - You're testing for a difference of means (in the "testing" variable) between 2 groups (of the "grouping" variable)
  - You'll need to tell SPSS which groups (M&F) and it only knows numbers (i.e. doesn't know what "men" means")
- **Levene's Test – the 1st p is not the p for "t" ☺**
  - There are 2 ways to calculate t, so 2 rows to the table
  - Levene's tells which formula (which row) to use
  - If the Levene's p is above .05, use the above row; if it's below .05, use the below row
- **When you have the right row, find the t and the p ("sig")**

Introduction to ANOVA

- Could use T-tests (Diff Means) for each pair
  - E.g. compare Family A&B, A&C, and B&C
  - But sampling error occurs with each comparison
    - Inflated probability of Type I error (falsely rejecting null)
    - Each comparison increases discrepency, =*additive error*
    - Larger S.E.→smaller test stat.→larger p→harder to reject null
- 3 advantages to the ANOVA alternative
  - Compares means *indirectly*, with variances
    - Uses "F" distribution, instead of "t" distribution
  - Compares two or more groups (if 2, then F = t^2)
  - F is an "omnibus test", overall test of the "model"

What ANOVA Examines

- Recall Variance
  - = sum of squared deviations from the mean

$$\sum \frac{(Y_i - \overline{Y})^2}{n-1}$$

- WSS & BSS are variations on that idea
  - WSS just compares the group variances
  - BSS compares each group to the "grand mean"
- ANOVA's F ratio compares WSS to BSS
  - Want "within group variance" to be small
    - want to say that strata are unique
    - want to say that the groups are clearly distinguishabbe
  - To support Ha (that there is a difference), want more variance between groups and less within each

F test looks at their ratio

- Between-group Sum of Squares (BSS; Systematic variance)
  - Hope to show that groups are <u>different</u> from each other
  - Therefore want differences *between* groups to be <u>large</u>
  - Want the group means to be very different, as before, but also…
- Within-group Sum of Squares (WSS; Random variance)
  - Want to conceive of each group as <u>consistently</u> unique
  - Therefore want differences *within* groups to be <u>small</u>
  - Observations should be similar to each other, i.e. close to the mean

ANOVA Example: Hypotheses

- Research Q: **Does family size vary w/ ethnicity?**
- Research hypothesis (H_A): **The number of children ever born to a woman, *on average*, differs between Chinese, Hispanic, & Hutterite groups.**

$$H_A : \mu_1 \neq \mu_2 \neq \mu_3$$

- Null hypothesis (H_0): **The number of children ever born to a woman is, *on average*, the same for Chinese, Hispanic, & Hutterite groups.**

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

ANOVA "Source table" in SPSS

- Look for this:

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	265.765	2	132.882	3.310	.066
Within Groups	562.000	14	40.143		
Total	827.765	16			

- Again, SPSS provides an "exact" p value (the estimated area under the curve to the right of the calculated F) so you don't need Appendix D (or anything else) to get a "tabular" F
- Also, SPSS gives the "Total Sum of Squares", (TSS = BSS+WSS), representing all of the variation from the means

Interval Covariation

- Remember definition of "relationship":
  - Values of the DV change across values of the IV
- If both variables are interval, 3 options:
  - **ANOVA**
    - compares IV categories with 2 kinds of DV variance
    - Uses the F test
  - **Correlation** ("r" value, and associated sig. level)
    - summary measure closely related to regression analysis
    - usually relating only two variables
  - **Regression** (model, equation, and "r-squared")
    - omnibus test of entire model, including multiple variables
    - summary equation: coefficients (beta weights), alpha, eta
    - Summary statistic: r^2, the Holy Grail of statistics

Scatterplots in SPSS

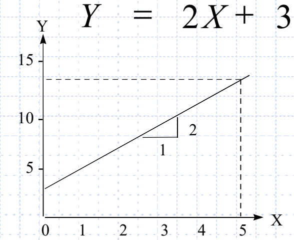
- Not the same as MeansPlot (as in ANOVA)
- Graphs – Scatter – click "define" – Y is dependent variable, X is independent

Regression is like Algebra…

- The formula for a *slope* was "Y = m*x + b"

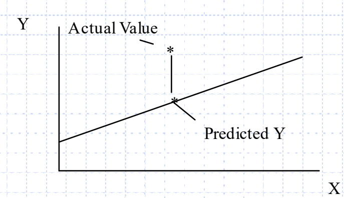
$$Y = 2X + 3$$

X	Y
0	3
1	5
2	7
3	9
4	11
5	13



…with errors

Not all of the points are on the line. The vertical distance between the line (which represents predicted values) and each actual observed data point is the "residual", or error



Core Calculations (a & b)

- For the regression equation Y = α + Bx,
  - B ("beta") is the population parameter slope (expected increase in Y for an increase of 1 in X) which we estimate w/ "b", where:

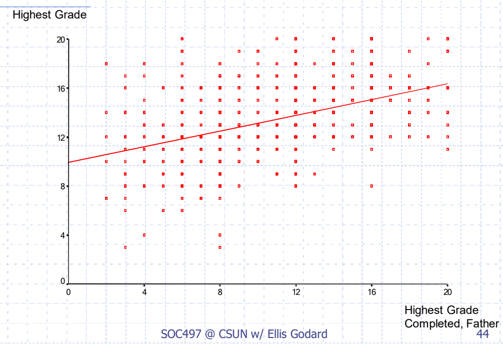
$$b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

  - α ("alpha") is the population parameter Y-intercept (predicted Y when X=0) which we estimate w/ "a":

$$a = \bar{Y} - b * \bar{X}$$

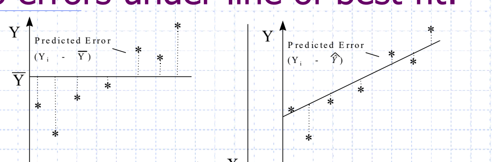
SOC497 @ CSUN w/ Ellis Godard 43

The "Line of Best Fit"



SOC497 @ CSUN w/ Ellis Godard 44

R^2 compares errors under null to errors under line of best fit:



Total Sum of Squares

$TSS = \sum (Y_i - \bar{Y})^2$

Sum of Squared Errors

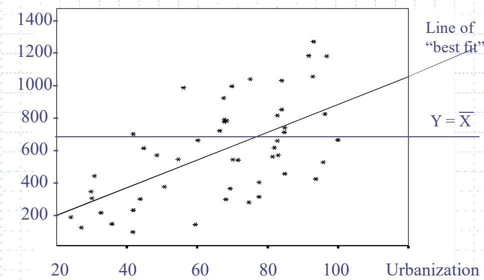
$SSE = \sum (Y_i - \hat{Y})^2$

$$r^2 = \frac{TSS - SSE}{TSS}$$

SOC497 @ CSUN w/ Ellis Godard 45

What does regression *test* ask?

Which line best summarizes the data points? Do they differ enough from Y=$\bar{X}$ to think the line of best fit is a better inference?



SOC497 @ CSUN w/ Ellis Godard 46

Benefits of Assuming Linearity

- **Describes the relationship between 2 variables**
- **Inferential values: Allows us to predict…**
  - the *average* value of the DV, for each value of the IV
  - the avg *increase* in DV *for each increase of 1* in the IV

Coefficients

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	223.190	1.393		160.193	.000
	Index of online media use	-5.102	.366	-.201	-13.932	.000

a. Dependent Variable: weight

  - …how much of the variance in the DV is explained by variance in the IV

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.201[a]	.041	.040	79.830

a. Predictors: (Constant), Index of online media use

  - **Can write equation:**
    - **Weight = 223.190 – (5.102*OLUSE)**
    - **Note p values**

SOC497 @ CSUN w/ Ellis Godard 47

Summary statistics in SPSS

VARIABLE	B	SE B	Beta	T	Sig T
IV name	"b"	se of "b"		test-stat for b	p-value
(constant)	"a"	se of "a"		test-stat for a	p-value

These are the hypotest values (t and p).

These coefficients are standardized in units of standard deviation
The standardized Beta = (σ_x / σ_y) * b
These are *not* the coefficients you want this week.

These are the coefficients you want for the equation Y = a + bx (estimating Y = α + Bx).
b = change, on average, in Y for each increase of 1 in X.
a = predicted value of Y when X=0 (but *don't infer beyond observed range of X*)

SOC497 @ CSUN w/ Ellis Godard 48

Q4. Which stat tests have p values?

1. All statistical tests — 85%
2. All interval ones — 3%
3. Only non-interval ones — 5%
4. Only bivariate ones — 3%
5. Only nominal ones — 5%

0

SOC497 @ CSUN w/ Ellis Godard 49

Substantive vs. Statistical Significance

- It's possible to obtain a *statistically* significant result for substantively *insignificant* differences
  - Especially true with relatively large samples.
  - It's possible to have 99.9% confidence that some effect (such as gender differences in income) is statistically inferable but extremely small
- What difference is *substantively* important is not the same as what difference is large enough to make an inference from a sample
  - Standard for *statistical* sig is $p<.05$ (tho that varies)
  - But there's no universal standard for *substantive* sig
    - Among other issues, would vary by variable (Units vs $$'s)

SOC497 @ CSUN w/ Ellis Godard 50

Demo: 3 Tests of Interval Data

- Dataset:
  - TestsDemo.sav
- Variables:
  - Dependent ("test variable") is INCOME
  - For t-tests, IVs include SEX & RACE
  - For ANOVA, IVs include MARITAL & RESIDE
  - For regression, IVs include AGE, EDUCATE, & LABOR
- Activities:
  - Start w/ Missings and Frequencies
  - Walk through the 3 tests
  - Compare and interpret results

SOC497 @ CSUN w/ Ellis Godard 51

Lab: 3 Tests of Means

- Your dependent variable is MINUTES, the length of time it took to complete the survey for MEANS.SAV
- Conduct 3 tests to examine variation in MINUTES
  - T-test w/ GENDER as the independent variable
  - ANOVA using PARTYID as the I.V.
  - Regression using YEARBORN as the I.V.
- Be sure to look at frequency distributions first
  - You should *always* look at central tendency, dispersion and shape first, but needn't submit them today
  - Be sure to check for & handle missing values
- Summarize the extent to which each of these potential independent variables accounts for variation in MINUTES

SOC497 @ CSUN w/ Ellis Godard 52

Q5. Today's backgrounds all had…

1. Skies — 0%
2. Stairs — 98%
3. Stores — 3%
4. Styrofoam — 0%
5. Sunsets — 0%

0

SOC497 @ CSUN w/ Ellis Godard 55

Team Scores

Points	Team	Points	Team
5	Shicker & Friends		
4.56	Thomas & Friends		
4	Facebook friends		
4	Phil Lesh & Friends		

SOC497 @ CSUN w/ Ellis Godard 56