DATA ANALYSIS IN SPSS POINT AND CLICK

www.ats.ucla.edu/stat/seminars/special/SPSS_analysis.pdf

By Andy Lin
(IDRE Stat Consulting)

TODAY'S SEMINAR

- Continuous, Categorical, and Ordinal variables
- Descriptive Statistics – Summaries of single variables
- Hypothesis Testing –Making inferences with statistics
- T-tests – comparing continuous means between 2 groups
- ANOVA and linear regression – comparing/predicting means among/using several variables
- Chi-square – comparing proportions among categories
- Logistic regression – predicting a binary (yes/no) outcome

CONTINUOUS, CATEGORICAL AND ORDINAL VARIABLES

- Continuous
  - numerical values meaningful
  - usually a measurement
  - e.g. age, height, blood pressure
  - SPSS calls these “scale” variables

CONTINUOUS, CATEGORICAL AND ORDINAL VARIABLES

- Categorical
  - numerical values, typically arbitrary labels, representing membership to a category
  - e.g. gender, occupation, hospital
  - SPSS calls these “nominal” variables
  - Value labels very useful

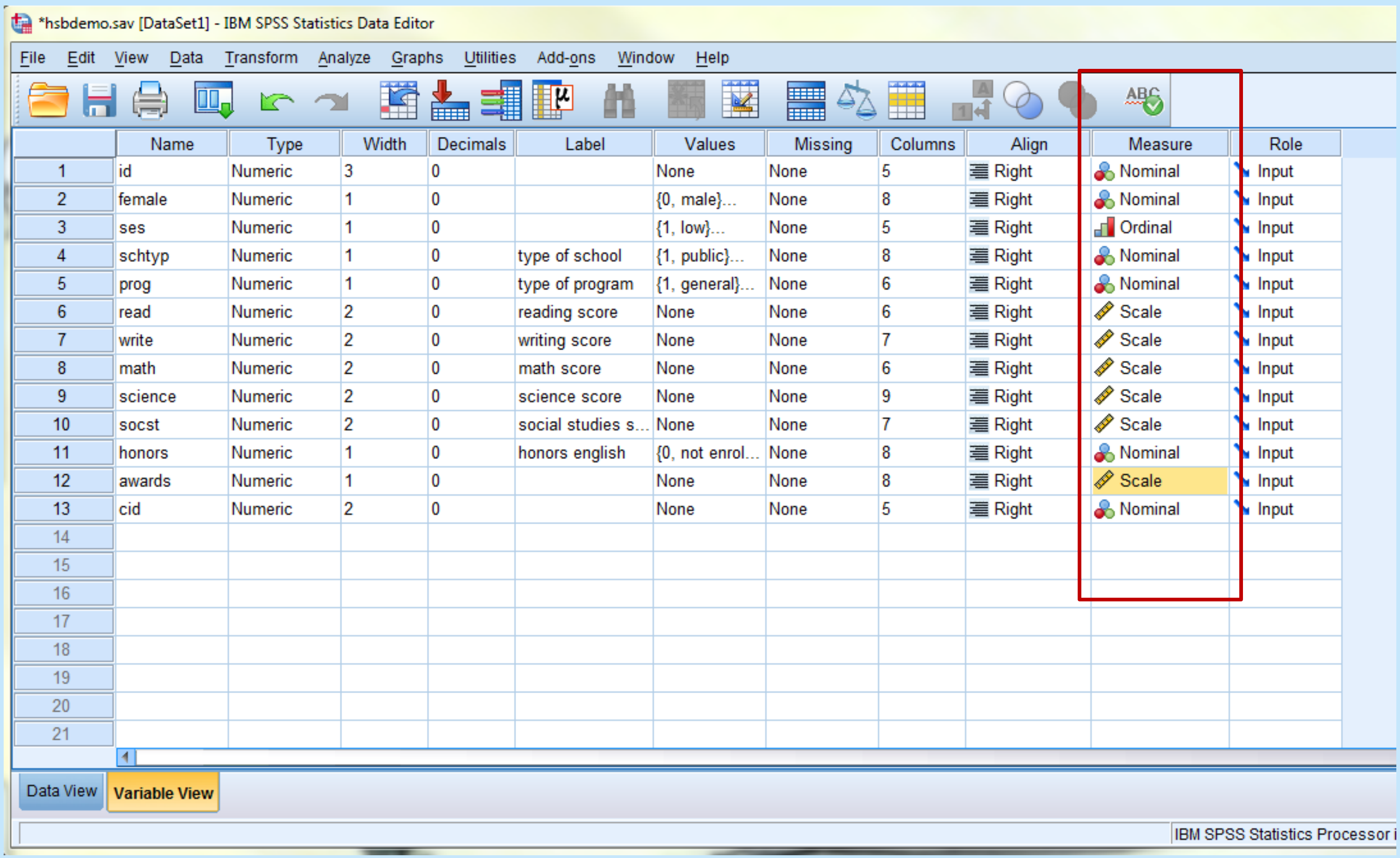
CONTINUOUS, CATEGORICAL AND ORDINAL VARIABLES

- Ordinal
  - numerical values denote relative rank, but lack clear absolute meaning
  - e.g. Pain scale (1-10), Likert scales, income bracket
  - SPSS calls these “ordinal” variables

SPSS WORKS A BIT SMOOTHER IF YOU SPECIFY YOUR VARIABLE TYPES

- In Variable View, can specify variable types in “Measure” column
  - SPSS will guess types when you load data
  - Some commands expect certain variables types to be used
    - helpful to have types specified beforehand

SPSS WORKS A BIT SMOOTHER IF YOU SPECIFY YOUR VARIABLE TYPES



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	id	Numeric	3	0		None	None	5	Right	Nominal	Input
2	female	Numeric	1	0		{0, male}...	None	8	Right	Nominal	Input
3	ses	Numeric	1	0		{1, low}...	None	5	Right	Ordinal	Input
4	schtyp	Numeric	1	0	type of school	{1, public}...	None	8	Right	Nominal	Input
5	prog	Numeric	1	0	type of program	{1, general}...	None	6	Right	Nominal	Input
6	read	Numeric	2	0	reading score	None	None	6	Right	Scale	Input
7	write	Numeric	2	0	writing score	None	None	7	Right	Scale	Input
8	math	Numeric	2	0	math score	None	None	6	Right	Scale	Input
9	science	Numeric	2	0	science score	None	None	9	Right	Scale	Input
10	socst	Numeric	2	0	social studies s...	None	None	7	Right	Scale	Input
11	honors	Numeric	1	0	honors english	{0, not enrol...	None	8	Right	Nominal	Input
12	awards	Numeric	1	0		None	None	8	Right	Scale	Input
13	cid	Numeric	2	0		None	None	5	Right	Nominal	Input
14											
15											
16											
17											
18											
19											
20											
21											

WE ANALYZE CONTINUOUS AND CATEGORICAL VARIABLES DIFFERENTLY

- Summarize them differently
  - Continuous = means, standard deviations, quantiles (e.g. quartiles)
  - Categorical = frequencies
- When analyzed as an outcome, the variable's type determines the appropriate analysis
  - Continuous = t-test, ANOVA, regression
  - Categorical = chi-square test of independence, logistic regression
- Ordinal variables can be analyzed both ways
  - treating them as continuous is controversial
  - Safer to analyze as a categorical variable, usually
    - Fewer assumptions
  - Also have their own set of analyses (ordinal logistic regression)

DESCRIPTIVE STATISTICS – CHARACTERIZING THE SAMPLE

- Descriptive statistics provide summaries of the characteristics of the sample
  - NOT used to infer relationships between variables, so no p-values

DESCRIPTIVE STATISTICS – CONTINUOUS VARIABLES

- Continuous variables
  - **Use Analyze -> Descriptive Statistics -> Descriptives**
    - Calculates mean, standard deviation, and minimum and maximum

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
reading score	200	28	76	52.23	10.253
writing score	200	31	67	52.78	9.479
math score	200	33	75	52.65	9.368
Valid N (listwise)	200				

DESCRIPTIVE STATISTICS – CATEGORICAL AND ORDINAL VARIABLES

- Continuous variables
  - **Use Analyze -> Descriptive Statistics -> Frequencies**
    - Proportion of sample within each category

Frequency Table

female

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	male	91	45.5	45.5	45.5
	female	109	54.5	54.5	100.0
	Total	200	100.0	100.0	

ses

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	low	47	23.5	23.5	23.5
	middle	95	47.5	47.5	71.0
	high	58	29.0	29.0	100.0
	Total	200	100.0	100.0	

HYPOTHESIS TESTING

- We assess the truthfulness of our research hypothesis using statistics
- Typically we are interested in showing:
  - Groups are different
  - One variable predicts another variable

HYPOTHESIS TESTING

- Formally:
  - We propose an uninteresting (null) hypothesis
    - Groups are not different
    - This variable does not predict that variable
  - We calculate the probability of the outcome if we assume the null hypothesis is true
    - This probability = p-value
  - Assess the null hypothesis in light of this probability
    - A small probability implies the outcome is VERY UNLIKELY if the null hypothesis is true, so we reject the null hypothesis
    - We thus conclude that the interesting, alternative hypothesis is supported

HYPOTHESIS TEST EXAMPLE

- Suppose a man on a street challenges you to a game of dice
  - Highest sum of 2 dice wins
  - First roll -- You: 6, He: 12
  - Second roll – You: 9, He: 12
- You immediately hypothesize he's cheating with a loaded dice
  - Let's test it

HYPOTHESIS TEST EXAMPLE

- 1. Propose the uninteresting null hypothesis:
  - His pair of dice are fair.
- 2. Calculate probability of outcome if we assume null hypothesis is true:
  - If his dice are fair, the probability of rolling 12 and 12 is 1/36*1/36 = 0.00077
    - We don't know the probability of rolling 12 and 12 with a loaded pair of dice, but we do with a fair pair of dice
      - This is why we test the null, not the alternative
- 3. Assess null hypothesis in light of this probability (p-value)
  - Since this outcome is so rare with a fair pair of dice, we reject the null hypothesis that he is using a fair pair of dice
    - Typically, a threshold p-value of 0.05 is used for rejection of the null
  - We thus conclude he is cheating
    - The alternative hypothesis

HYPOTHESIS TEST RESEARCH EXAMPLE

- 1. Propose the uninteresting null hypothesis:
  - This drug has no effect on weight.
  - H_0: mean_weight_control - mean_weight_cases = 0
- 2. Calculate probability of outcome if we assume null hypothesis is true:
  - What is probability of seeing a difference this size assuming the drug has no effect
    - If the drug has no effect, then the difference in mean weight is due to sampling variability (chance)
    - We can calculate the probability of observing the difference in means due to chance alone by calculating how much people randomly vary in our sample
      - If people do not randomly vary much, then a large difference between means is very unlikely if the drug has no effect
      - If people vary wildly by chance, then a large difference in means is possible by chance even if the drug has no effect
        - But becomes less probable as sample size increases
    - Assumptions allow us to calculate these probabilities
      - For instance, we might assume that weight is normally distributed
      - Before, assuming that the dice were fair allowed us to calculate the probability of the outcome

HYPOTHESIS TEST RESEARCH EXAMPLE

- 3. Assess null hypothesis in light of this probability
  - If the size of the difference is very unlikely if the drug has no effect, we reject the hypothesis that the drug has no effect
  - If the size of the difference is NOT unlikely (p-value > 0.05) if the drug has no effect, we fail to reject the null hypothesis that the drug has no effect

WHAT TYPE OF TEST DO I USE?

- Distribution of the outcome determines the type of statistical test to use
  - Continuous normally-distributed outcomes:
    - Do group means differ?
      - 2 groups – t-test
      - 3 or more groups – ANOVA and linear regression
    - Does this measurement predict the outcome?
      - Regression
    - Some departure from normality acceptable, especially with large samples
      - Large departures may require non-parametric tests
  - Categorical outcomes:
    - Do group proportions on outcomes differ?
      - Chi-square test of independence, logistic regression (binary outcome)
    - Does this measurement predict the outcome?
      - Logistic regression (binary outcome)

WHAT TYPE OF TEST DO I USE?

- A huge variety of methods to analyze data
- Outcomes with the following properties may require more advanced or special methods
  - Correlated outcome
    - Repeated measurements
    - Clustering
  - Count outcomes
  - Time to event outcomes
- Data with significant amount of missing or censoring may also require special methods
- Link to table of outcomes and appropriate statistical test:
  - http://www.ats.ucla.edu/stat/spss/whatstat/

T-TEST: DO 2 GROUPS MEANS DIFFER?

- We use t-tests to assess whether 2 group means differ
  - Example: Do patients who take a proposed diet drug have a lower mean weight than those who take placebo?
  - Assumes outcome is normally distributed
- We will test in SPSS whether males and females differ in their math achievement test scores
  - **Analyze -> Compare Means -> Independent Samples t-test**
    - **Test Variable** = outcome
    - **Grouping variable** = group
      - Must tell SPSS numerical values of groups to be compared
    - Use Paired Samples t-test if outcomes may be correlated

T-TEST: DO 2 GROUPS MEANS DIFFER?

- Males and females do not significantly differ in math scores

Group Statistics

	female	N	Mean	Std. Deviation	Std. Error Mean
math score	male	91	52.95	9.665	1.013
	female	109	52.39	9.151	.877

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
math score	Equal variances assumed	.619	.432	.413	198	.680	.551	1.333	-2.078	3.179
	Equal variances not assumed			.411	187.575	.682	.551	1.340	-2.092	3.193

T-TEST: DO 2 GROUPS MEANS DIFFER?

- Typically 3 items are reported
  - t – a measure of the difference between means relative to their variability
  - df – the degrees of freedom, a measure of our effective sample size
  - p-value – probability of observing a t of this size, given the degrees of freedom (sample size)
    - SPSS output column **Sig. (2-tailed)**
- t(198) = 0.413, p = 0.68

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
math score	Equal variances assumed	.619	.432	.413	198	.680	.551	1.333	-2.078	3.179
	Equal variances not assumed			.411	187.575	.682	.551	1.340	-2.092	3.193

T-TEST: DO 2 GROUPS MEANS DIFFER?

- T-test assumes groups have equal variances
  - Levene's Test assesses this assumption
  - If "Sig." for Levene's Test is < 0.05, consider using bottom row of results
    - Corrects for unequal variances

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
math score	Equal variances assumed	.619	.432	.413	198	.680	.551	1.333	-2.078	3.179
	Equal variances not assumed			.411	187.575	.682	.551	1.340	-2.092	3.193

ANOVA – ARE THERE DIFFERENCES AMONG MANY GROUP MEANS?

- ANOVA compares means of several groups
  - Example: Do patients taking diet drugs A, B, and C have different mean body weight from each other and from controls?
  - T-test is special case of ANOVA
  - Assumes outcome normally distributed
    - And homogeneity of variance
- We will test in SPSS whether three different programs differ on their writing test scores
  - Analyze -> General Linear Model -> Univariate
    - Dependent variable = outcome
    - Fixed Factor(s) = grouping variables
    - If you have more than one factor, use the Model window to specify which effects you want
      - Main effects vs interactions

ANOVA OUTPUT

- The three SES classes differ in their writing score means
  - Known as an omnibus test
    - Tested before pairwise tests
- Sometimes the entire ANOVA table reported
- In text, usually report:
  - F – how much groups means differ relative to their variability
  - df – measure of effective sample size
    - Need df for ses (2) and df for Error (197)
  - p-value (**Sig.**) – probability of observing F this size under null hypothesis
  - F(2, 197) = 4.97, p = 0.008

Tests of Between-Subjects Effects

Dependent Variable: writing score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	858.715[a]	2	429.358	4.970	.008
Intercept	511958.919	1	511958.919	5925.673	.000
ses	858.715	2	429.358	4.970	.008
Error	17020.160	197	86.397		
Total	574919.000	200			
Corrected Total	17878.875	199			

a. R Squared = .048 (Adjusted R Squared = .038)

ANOVA: POST-HOC TESTS

- Use Post Hoc window to perform pairwise comparison
- Tukey and Bonferroni are common adjustments
- High different from low and middle

Multiple Comparisons

Dependent Variable: writing score

Tukey HSD

(I) ses	(J) ses	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
low	middle	-1.31	1.658	.710	-5.22	2.61
	high	-5.30*	1.824	.011	-9.60	-.99
middle	low	1.31	1.658	.710	-2.61	5.22
	high	-3.99*	1.549	.029	-7.65	-.33
high	low	5.30*	1.824	.011	.99	9.60
	middle	3.99*	1.549	.029	.33	7.65

Based on observed means.
The error term is Mean Square(Error) = 86.397.

*. The mean difference is significant at the 0.05 level.

LINEAR REGRESSION – DOES THIS MEASUREMENT PREDICT THE OUTCOME?

- Linear regression and ANOVA are equivalent methods
  - Use the same command in SPSS
- Regression typically used to model effects of continuous predictors
  - Example: does resting heart rate predict weight?
- Regression can also model categorical predictors, like ANOVA
- We control for the effects of predictors by adding them to the regression model
  - If gender and age are in regression model, we interpret coefficient of gender as the effect of gender, after controlling for age

LINEAR REGRESSION IN SPSS

- Let us model whether gender and reading test score predict writing test score
  - **Analyze -> General Linear Model -> Univariate**
    - Dependent variable = outcome
    - Fixed Factor(s) = categorical predictors (factors)
    - Covariate(s) = continuous predictors
  - **Options Window -> Check Parameter Estimates**
    - This outputs the regression table
  - **Model Window**
    - By default, SPSS will fully interact all factors (categorical predictors) in the model
      - But, you can specify exactly which main effects and interactions you want
    - Interactions model the effects of one variable changing with levels of the another variable
      - Example – the effect of weight on heart rate may differ between man and women

LINEAR REGRESSION OUTPUT

- Regression coefficients: change in outcome per unit-change in predictor
- For each unit increase in reading score, writing score increases by .566, after controlling for gender
- Males(female=0) on average score 5.487 lower on writing, after controlling for reading score

Parameter Estimates

Dependent Variable: writing score

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	25.715	2.645	9.724	.000	20.500	30.931
read	.566	.049	11.459	.000	.468	.663
[female=0]	-5.487	1.014	-5.410	.000	-7.487	-3.487
[female=1]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

LINEAR REGRESSION REPORTING

- For each effect, report:
- B coefficient – magnitude of effect
- Std. Error or t –provide same info
- p-value (**Sig.**) - probability of observing B coefficient of this size relative to its standard error

Parameter Estimates

Dependent Variable: writing score

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	25.715	2.645	9.724	.000	20.500	30.931
read	.566	.049	11.459	.000	.468	.663
[female=0]	-5.487	1.014	-5.410	.000	-7.487	-3.487
[female=1]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

CHI-SQUARE TEST OF INDEPENDENCE

- Used to see if 2 categorical variables are associated
  - Are the proportions across levels of one variable the same across levels of the other variable?
    - Are the proportions that fall within each BMI category the same for patients with Type I vs Type II diabetes?
- Let's test whether the proportions within each school program type are the same across genders
- Analyze -> Descriptives -> Frequencies
- Statistics Window -> Check Chi-square

CHI-SQUARE TEST OF INDEPENDENCE IN SPSS

- Let’s test whether the proportions within each school program type are the same across genders
- Analyze -> Descriptives -> Crosstabs
  - Rows(s) – 1 categorical variable here
  - Column(s) – 1 categorical variable here
  - Layer 1 of 1 – Any additional categorical variables to test for association
    - Can test 3- or more-way association
- Statistics Window -> Check Chi-square

CHI-SQUARE TEST OF INDEPENDENCE OUTPUT

- Gender and program type not associated
- Typically report:
  - Chi-square – measures how much observed proportions differ from proportions expected if variables are not associated
  - df – (number of rows-1)*(number of columns-1)
  - P-value (**Asymp. Sig.**) – probability of observing this chi-square with these df
  - $\chi^2(2) = .053, p = .974$

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	.053[a]	2	.974
Likelihood Ratio	.053	2	.974
Linear-by-Linear Association	.003	1	.955
N of Valid Cases	200		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 20.48.

LOGISTIC REGRESSION

- Used when analyzing what predicts a binary outcome
  - Binary = 0/1, yes/no
- More specifically, we are modeling what affects the odds of the outcome
  - Let p = probability that outcome = 1
  - Odds(p) = p/(1-p)
    - So, if p = .5, odds = .5/.5 = 1
    - If p = .75, odds = .75/.25 = 3
- Example – are gender and age predictive of the odds of developing Parkinson's disease?

LOGISTIC REGRESSION IN SPSS

- Let's see if gender and math score predict the membership to an honors program (1 = in honors, 0 = not)
- **Analyze -> Generalized Linear Models -> Generalized Linear Models**
  - **Type of Model tab**
    - Choose Binary logistic
  - **Response tab**
    - Dependent variable = move binary outcome here
    - Click Reference Category button
      - Choose **First(lowest value)**
        - Will model odds outcome = 1
  - **Predictors tab**
    - Factors – categorical predictors
    - Covariates – continuous predictors
  - **Model tab**
    - Choose which main effects and interactions you want
  - Above is the minimum specification for logistic regression
  - To get odds ratios reported
    - **Statistics tab**
      - Check **Include exponential parameter estimates**

LOGISTIC REGRESSION OUTPUT

- OR=1 : no diff , OR<1: Lower Odds, OR>1: Higher Odds
- Being male decreases the odds of being in the honors program by 68.4%
  - 1 – Exp(B) for female = 0
- Each point increase in math score increases odds of being in honors by 20%
  - Exp(B) – 1 for math

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test			Exp(B)	95% Wald Confidence Interval for Exp(B)	
			Lower	Upper	Wald Chi-Square	df	Sig.		Lower	Upper
(Intercept)	-10.671	1.5988	-13.805	-7.538	44.552	1	.000	2.320E-005	1.011E-006	.001
[female=0]	-1.121	.4240	-1.952	-.290	6.987	1	.008	.326	.142	.748
[female=1]	0[a]	.	.	.	.	.	.	1	.	.
math	.183	.0284	.127	.238	41.248	1	.000	1.200	1.135	1.269
(Scale)	1[b]									

Dependent Variable: honors english
Model: (Intercept), female, math

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

LOGISTIC REGRESSION OUTPUT

- Typically report
  - Coefficient = **B**
  - Odds ratios – **Exp(B),** the factor by which the odds of the outcome change per unit change in the predictor
  - Std. Error – variability in coefficient
  - Chi-square and df – used to test coefficient
  - p-value – **Sig.**, probability of observing chi-square
  - Often report confidence interval on odds ratio
  - B = -1.21, S.E. = .424, Odds ratio = .326, $\chi^2(1)$=.008

ADDITIONAL RESOURCES

- Data analysis examples
  - How to conduct more advanced analyses
  - http://www.ats.ucla.edu/stat/dae/
- Annotated output
  - How to read output
  - http://www.ats.ucla.edu/stat/AnnotatedOutput/
- Web seminars for SPSS
  - More detailed guides
  - http://www.ats.ucla.edu/stat/seminars/#SPSS