Analyses of Qualitative Variables

There are several kinds of analyses involving qualitative variables that I want to review today to help get ready for the various regression models we'll cover the next few weeks.

Univariate Analyses of Binary & Multiple Category Variables

The most common starting place with quantitative variables is the mean, std and Skewness -- are these useful for qualitative variables?

Statistics

	GROUP	GENDER	MARITAL
N	288	288	288
	0	0	0
Mean	1.507	.781	1.573
Std. Deviation	.501	.414	.771
Skewness	-.028	-1.368	1.547
Std. Error of Skewness	.144	.144	.144

SPSS willingly provides these statistics for any variables you ask -- but are they useful summary values?

For **binary** variables:

- the **decimal portion of the mean** tells the **proportion of the sample that is in the higher coded group**
- **the standard deviation** is sqrt((m*(1-m)) where m = the decimal part of the mean. Std is at its largest with a 50% split and smaller with disproportionate samples
- the direction of **skewness** tells the less frequent group

GROUP

		Frequency	Percent	Cumulative Percent
Valid	traditional	142	49.3	49.3
	nontraditional	146	50.7	100.0
	Total	288	100.0	

For Group

- the mean of 1.507 tells us that 50.7% of the sample is coded 2 (non-traditional students) -- matching the % given in the frequency table
- notice the "symmetry" of a 50-50 split

GENDER

		Frequency	Percent	Cumulative Percent
Valid	male	63	21.9	21.9
	female	225	78.1	100.0
	Total	288	100.0	

For Gender

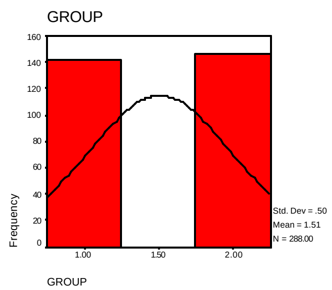
- the mean of .781 tell us that 78.1% of the sample is coded 1 (female) -- again matching the % given in the frequencies
- notice the "asymmetry", with the negative skewness indicating the smaller value has the lower frequency

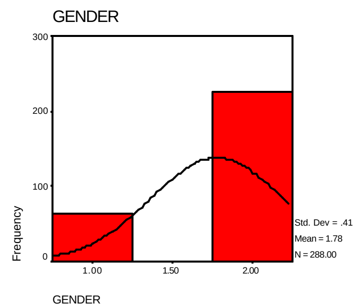
MARITAL

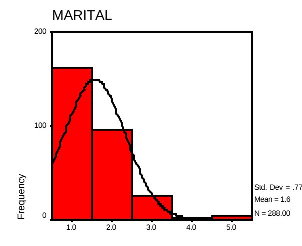
		Frequency	Percent	Cumulative Percent
Valid	single	162	56.3	56.3
	married	95	33.0	89.2
	divorced	26	9.0	98.3
	separated	2	.7	99.0
	widowed	3	1.0	100.0
	Total	288	100.0	

For multiple category variables these parametric summary statistics have no meaning!

- There are multiple frequency patterns of these 5 categories that will produce this mean
- Std and skewness assume the values have a meaningful order and valuing, while these "values" represent kinds, not amounts.







Bivariate Analyses with One Quantitative and One Binary Variable

Because the means and standard deviations of binary variables are meaningful, there are several statistically equivalent analyses available.

- t-test and ANOVA can be used to test whether the two groups have different means on the quantitative variable (ANOVA can be applied with multiple-category variables)
- correlation can also be used to examine the same question
- the effect size of the t-test and ANOVA will match and both will equal the absolute value of the correlation

t-test assessing relationship between gender and loneliness (Rural and Urban Loneliness Scale)

Group Statistics

	GENDER	N	Mean	Std. Deviation
loneliness	male	63	31.60	8.526
	female	225	37.00	11.509

Independent Samples Test

	t-test for Equality of Means		
	t	df	Sig. (2-tailed)
loneliness	-3.466	286	.001

For this analysis $r = \text{sqrt} \dfrac{t^2}{t^2 + df} = \text{sqrt} \dfrac{3.466^2}{3.466^2 + 286} = .20$

ANIOVA assessing relationship between gender and loneliness (Rural and Urban Loneliness Scale)

Descriptives

loneliness

	N	Mean	Std. Deviation
male	63	31.60	8.526
female	225	37.00	11.509
Total	288	35.82	11.140

ANOVA

loneliness

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1435.894	1	1435.894	12.015	.001
Within Groups	34178.075	286	119.504		
Total	35613.969	287			

For this analysis $r = \text{sqrt} \dfrac{F}{F + df} = \text{sqrt} \dfrac{12.015}{12.015 + 286} = .20$

Correlation of assessing relationship between gender and loneliness (Rural and Urban Loneliness Scale)

Since the mean and std of a binary variable "makes sense" and correlation is primarily influenced by scores on the two variables co-vary around their respective means, the correlation will give the same summary as the t-test and ANOVA.

Correlations

		GENDER	loneliness
GENDER	Pearson Correlation	1	.201**
	Sig. (2-tailed)	.	.001
	N	288	288
loneliness	Pearson Correlation	.201**	1
	Sig. (2-tailed)	.001	.
	N	288	288

**. Correlation is significant at the 0.01 level (2-tailed).

Bivariate Analyses with Two Binary Variables

Because the means and standard deviations of binary variables are meaningful, there are several statistically equivalent analyses available.

- X^2 test for independence (also called X^2 for contingency tables)
- t-test and ANOVA can be used to test whether the two groups have different means on the quantitative variable ---- that is different proportions of their respective samples that are in the higher coded group (ANOVA can be applied with multiple-category variables)
- the t-tests and ANOVA can be used with either variable as the IV
- correlation can also be used to examine the same question
- the effect size of the X^2, t-test, ANOVA will all match and all will equal the absolute value of the correlation

X^2 for independence applied to a 2x2 contingency table of gender & group

GENDER * GROUP Crosstabulation

Count

		GROUP		
		traditional	nontraditional	Total
GENDER	male	40	23	63
	female	102	123	225
Total		142	146	288

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6.493[b]	1	.011

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 31.06.

For this analysis $r = \text{sqrt}\ \frac{X^2}{N} = \text{sqrt}\ \frac{6.493}{288} = .15$

t-test with gender as "the IV"

Group Statistics

	GENDER	N	Mean	Std. Deviation
GROUP	male	63	1.37	.485
	female	225	1.55	.499

Independent Samples Test

		t-test for Equality of Means		
		t	df	Sig. (2-tailed)
GROUP	Equal variances assumed	-2.568	286	.011

For this analysis $r = \text{sqrt}\ \frac{t^2}{t^2 + df} = \text{sqrt}\ \frac{2.568^2}{2.568^2 + 286} = .15$

t-test with group as "the IV"

Group Statistics

	GROUP	N	Mean	Std. Deviation	Std. Error Mean
GENDER	traditional	142	.72	.451	.038
	nontraditional	146	.84	.366	.030

Independent Samples Test

		t-test for Equality of Means		
		t	df	Sig. (2-tailed)
GENDER	Equal variances assumed	-2.568	286	.011

For this analysis $r = \text{sqrt}\ \frac{t^2}{t^2 + df} = \text{sqrt}\ \frac{2.568^2}{2.568^2 + 286} = .15$

Correlation of assessing relationship between gender and group

Correlations

		GENDER	GROUP
GENDER	Pearson Correlation	1	.150*
	Sig. (2-tailed)	.	.011
	N	288	288
GROUP	Pearson Correlation	.150*	1
	Sig. (2-tailed)	.011	.
	N	288	288

*. Correlation is significant at the 0.05 level (2-tailed).

As with one binary and one quantitative variable, all the different analyses for two binary variables produce the same result.

Odds & the Odds Ratio

Another useful index of the relationship between two binary variables is the odds ratio.

Back to the 2x2 contingency table for gender * group

GENDER * GROUP Crosstabulation

Count

		GROUP		Total
		traditional	nontraditional	
GENDER	male	40	23	63
	female	102	123	225
Total		142	146	288

For a given gender, the odds of being in a particular group are given by the frequency in that group divided by the frequency in the other group

For males, the odds of being in the traditional group are:
40 / 23 = 1.7391 meaning that if you are male, the odds are 1.7391 to 1 that you are a traditional student

For females, the odds of being in the traditional group are"
102 / 123 = .8293 meaning that if you are female, the odds are .8293 to 1 that you are a traditional student

The Odds Ratio is simply the ratio of the odds of being a traditional student for the two genders.

For this analysis the odds ratio is
1.7391 / .8293 = 2.0972 meaning that males are twice as likely to be traditional students as are females.

The odds ratio is the same if we compute it "the other way"

For traditional students, the odds of being male is 40 / 102 = .3922

For nontraditional students the odds of being male are 23 / 123 = .1970

The odds ratio is .3922 / .1970 = 1.990 -- oops???

Nope -- rounding error!!

For traditional students 40 / 102 = .392156
For nontraditional students 23 / 123 = .186992

Giving the odds ratio 2.0972

For sufficient accuracy, keep 5-6 decimals when calculating these summary statistics !

When there is no relationship between the variables (that is, when the variables are statistically independent) then the odds will be the same for the two categories and the ration will be 1 (or 1:1).

Multivariate Analyses with a Binary Criterion

The OLS analyses available for this situation are linear discriminant analysis and multiple regression, which will produce equivalent results when the criterion in binary.

Multiple Regression

Model Summary[b]

Model	R	R Square	Adjusted R Square
1	.909[a]	.826	.825

a. Predictors: (Constant), total social support, AGE, GENDER

b. Dependent Variable: GROUP

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	59.485	3	19.828	450.46	.000[a]
	Residual	12.501	284	.044		
	Total	71.986	287			

a. Predictors: (Constant), total social support, AGE, GENDER

b. Dependent Variable: GROUP

Coefficients[a]

Model		Unstandardized	Standardized Coefficients	t	Sig.
		B	Beta		
1	(Constant)	.448		4.957	.000
	GENDER	-.008	-.006	-.246	.806
	AGE	.040	.901	35.483	.000
	total social support	-.018	-.040	-1.515	.131

a. Dependent Variable: GROUP

Linear Discriminant Function

Eigenvalues

Function	Eigenvalue	Canonical Correlation
1	4.758[a]	.909

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.174	498.060	3	.000

andardized Canonical Discriminant Function Coefficien

	Function
	1
GENDER	-.017
AGE	.996
total social support	-.102

Notice that the **R** from the regression and the $\mathbf{R_c}$ from the discriminant are the same.

The standardized weights are different by a transformation that reflects the difference between the desired properties of y' and *ldf* values.

One way to demonstrate the equivalence of multiple regression and discriminant function for this model is that the y' and *ldf* values for individuals are equivalent -- that they are perfectly correlated.

With both models, the predicted value is applied to a cutoff to make a classification decision.

Correlations

		Standardized Predicted Value	Discriminant Scores from Function 1 for Analysis 1
Standardized Predicted Value	Pearson Correlation	1	1.000**
	Sig. (2-tailed)	.	.
	N	288	288
Discriminant Scores from Function 1 for Analysis 1	Pearson Correlation	1.000**	1
	Sig. (2-tailed)	.	.
	N	288	288

**. Correlation is significant at the 0.01 level (2-tailed).

One difficulty with both of these models is that the math "breaks down" as variables are skewed (as are r, t, F & X^2). They are particularly sensitive to skewing in the criterion variable -- that is when the groups are substantially disproportionate. This weakness has been well-documented and is the reason for the advent and adoption of the models we will be studying during the remainder of the module.