

## Bivariate Analysis - Crosstabulation

- One of most basic research tools – shows how x varies with respect to y
- Interpretation of table depends upon “direction of percentaging”
- example

---

---

---

---

---

---

---

---

---

---

---

---

## Row vs. Column Percents

tab PRAYER DEGREE, row col

Key
Frequency
row percentage
column percentage

BIBLE PRAYER IN PUBLIC SCHOOLS	RS HIGHEST DEGREE					Total
	LT HIGH S	HIGH SCHD	JUNI OR CO	BACHELOR	GRADUATE	
APPROVE	1,492 15.13 27.29	4,797 48.63 36.96	569 5.77 41.99	1,942 19.69 56.18	1,064 10.79 64.62	9,864 100.00 39.60
DI SAPPROVE	3,976 26.43 72.71	8,183 84.39 63.04	786 5.22 58.01	1,615 10.07 49.82	585 3.89 35.48	15,045 100.00 60.40
Total	5,468 21.95 100.00	12,980 52.11 100.00	1,355 5.44 100.00	3,457 13.88 100.00	1,649 6.62 100.00	24,909 100.00 100.00

119. A. The United States Supreme Court has ruled that no state or local government may require the reading of the Lord's Prayer or Bible verses in public schools. What are your views on this—do you approve or disapprove of the court ruling?

---

---

---

---

---

---

---

---

---

---

---

---

## Row vs. Column Percents

by RACE, sort : tabulate PRAYER DEGREE, column noFreq row

--> RACE = W

Key
row percentage
column percentage

BIBLE PRAYER IN PUBLIC SCHOOLS	RS HIGHEST DEGREE					Total
	LT HIGH S	HIGH SCHD	JUNI OR CO	BACHELOR	GRADUATE	
APPROVE	13.71 28.39	48.49 35.30	5.75 44.43	20.79 88.47	11.26 66.91	100.00 61.99
DI SAPPROVE	34.62 71.81	55.60 61.70	5.12 85.87	10.81 41.83	4.15 34.09	100.00 86.41
Total	20.08 100.00	52.64 100.00	5.38 100.00	14.78 100.00	7.50 100.00	100.00 100.00

--> RACE = B

Key
row percentage
column percentage

BIBLE PRAYER IN PUBLIC SCHOOLS	RS HIGHEST DEGREE					Total
	LT HIGH S	HIGH SCHD	JUNI OR CO	BACHELOR	GRADUATE	
APPROVE	31.87 21.44	51.99 25.71	5.74 27.07	9.28 31.35	5.15 41.12	100.00 28.29
DI SAPPROVE	34.54 78.56	50.88 75.29	5.23 72.93	4.86 66.65	2.80 86.88	100.00 71.71
Total	32.87 100.00	51.14 100.00	5.34 100.00	7.44 100.00	3.17 100.00	100.00 100.00

---

---

---

---

---

---

---

---

---

---

---

---

## Reading Tables - Basics

- What's the difference between 20% and 30%?
- What do row, column, and total percents tell you?
- Why are percent distributions more informative than distributions of raw numbers?
- Always best to either show or mention your numbers – two forms of info clearer than one.

---

---

---

---

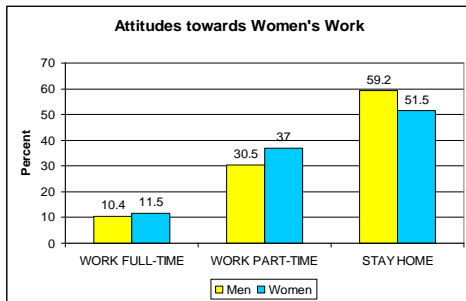
---

---

---

---

## Example




---

---

---

---

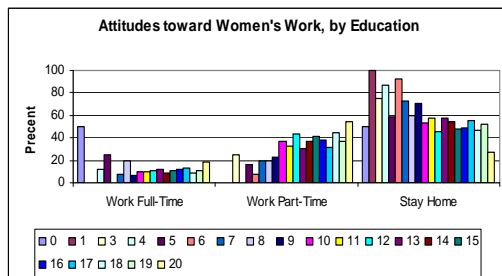
---

---

---

---

## Example before recoding




---

---

---

---

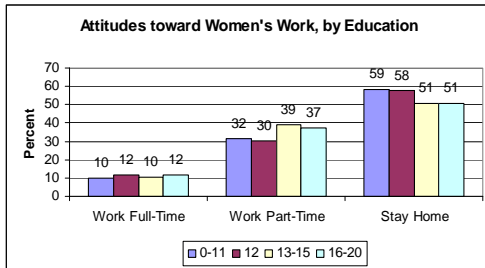
---

---

---

---

## Example after recoding



---

---

---

---

---

---

---

---

## Extensions

- Three-way crosstabulations
  - Web: use the “control” option
  - Stata: use “by” or “table”
- Crosstabulations for subsample
  - Web: use the “filter” option
  - Stata: use the “if” option
- Example: Educational differences in attitudes toward women’s roles – could do either way (wrkby x educ x sex or wrkby x educ if sex=male).

---

---

---

---

---

---

---

---

## Chi-square statistic – Introduction

- Chi-square statistic is useful for testing statistical significance of relationships in crosstabulations, or contingency tables.
- Contingency tables: tables cross-classifying categorical variables.
- 2-way and 3-way (involving two or three variables).

---

---

---

---

---

---

---

---

## Chi-square statistic

- Chi-square statistics reported by a cross-tabulation tests the hypothesis that the row variable and the column variable are independent of each other. If they are independent, joint probabilities can be determined by marginal probabilities.
  - $(N_{i+}/N_{++}) \times (N_{+j}/N_{++})$
- For a two-way contingency table of dimension IxJ, the degrees of freedom are (I-1)(J-1).
- A large chi-square rejects the hypothesis. Larger the chi-square => the stronger the evidence against independence => the stronger the relationship.
- Example of religious homogamy.

---

---

---

---

---

---

---

---

---

---

## Chi-square statistic

. tabulate RELIG RELIGSP, chi 2 expected

Key
Frequency
expected Frequency

RS RELIGIOUS PREFERENCE	RS SPOUSE'S RELIGION					Total
	PROTESTANT	CATHOLIC	JEWISH	NONE	OTHER (SP)	
PROTESTANT	659 456.9	60 212.1	3 20.3	41 65.4	9 17.3	772 772.0
CATHOLIC	51 196.5	256 91.2	4 8.8	19 28.1	2 7.4	332 332.0
JEWISH	1 16.6	2 7.7	22 0.7	2 2.4	1 0.6	28 28.0
NONE	19 50.9	19 23.6	4 2.3	41 7.3	3 1.9	86 86.0
OTHER (SPECIFY)	11 20.1	7 9.3	0 0.9	3 2.9	13 0.8	34 34.0
Total	741 741.0	344 344.0	33 33.0	106 106.0	28 28.0	1,252 1,252.0

Pearson chi 2(16) = 1.7e+03 Pr = 0.000

---

---

---

---

---

---

---

---

---

---

## Comparison of means

- Like crosstabulations but better for continuous (i.e., interval) variables.
- Example: Mean age at first marriage by sex and race (aged x sex x race)

---

---

---

---

---

---

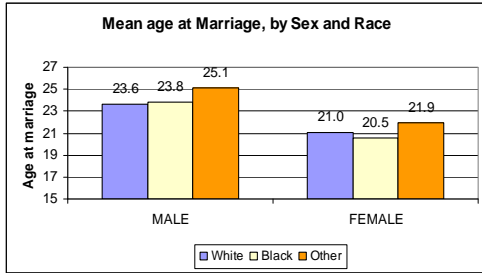
---

---

---

---

## Comparison of means



---

---

---

---

---

---

---

---

## Regression: Concepts

- In essence, regression is conditional mean.
- e.g.,  $X$  = independent variable, and  $Y$  = dependent variable
- Regression says  $Y = F(X) + \epsilon$
- $F(X)$  is called the regression function.
- If  $F(X)$  is linear,  $F(X) = \beta_0 + \beta_1 X$ 
  - Example: income and happiness
- $F(X)$  could be nonlinear.  
e.g.,  $F(X) = \beta_0 + \beta_1 X + \beta_2 X^2$ .

---

---

---

---

---

---

---

---

## Regression: Error Term

- Because a regression model does not predict observed values exactly, there is an error term.
- This is due to our desire for data reduction.
- In bivariate case, a linear model looks like
- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

---

---

---

---

---

---

---

---

## Multivariate Regression

- Regression one of many alternatives in bivariate case, but only real alternative in multivariate case
- $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$ ,
- $\beta$ 's are also called "partial" regression coefficients. Say  $\beta_3 = 2.0$  (with  $x_{i1} = \$$  of income in thousands), e.g., controlling for sex and marital status, every \$1,000 "buys" you two points on the happiness scale.

---

---

---

---

---

---

---

---

## Multivariate Regression cont'd

- $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$
  - If  $x_1$  is defined so 0=not married & 1=married
  - If  $x_2$  is defined so 0=male & 1=female
- $\beta_0$  is predicted happiness for unmarried man w/ no income
- $\beta_0 + \beta_1$  is predicted happiness for married man w/ no income
- $\beta_0 + \beta_1 + \beta_2$  is predicted happiness for married woman w/ no income
- $\beta_0 + \beta_1 + \beta_2 + \beta_3 x_{i3}$  is predicted happiness for married woman w/ income level  $x_i$

---

---

---

---

---

---

---

---

## Regression: Interpretations

- A. Causal Interpretation (Classical)  
Observed = True Model + Error
- B. Descriptive Interpretation (Demographic)  
Observed = Fitted + Residual
- C. Best Predictor Interpretation (Programmatic)  
Observed = Prediction + Deviation

---

---

---

---

---

---

---

---

### Tradeoff between Goodness-of-Fit and Parsimony

- In modeling observed data with a statistical model, you always have the problem of balancing the need for accuracy on the one hand and the need for parsimony on the other hand.
- By accuracy I mean the ability of your statistical model to reproduce the observed data. (Prediction).

---

---

---

---

---

---

---

---

### Consequences of Including Irrelevant Independent Variables

- Should we always include as many independent variables as possible?
- The answer is no. Do not include irrelevant independent variables. Why?
  - 1. Missing theoretically interesting findings
  - 2. Violating the parsimony rule (Occam's Razor)
  - 3. Wasting degrees of freedom
  - 4. Making estimates imprecise.

---

---

---

---

---

---

---

---