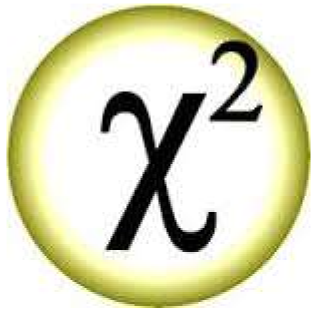
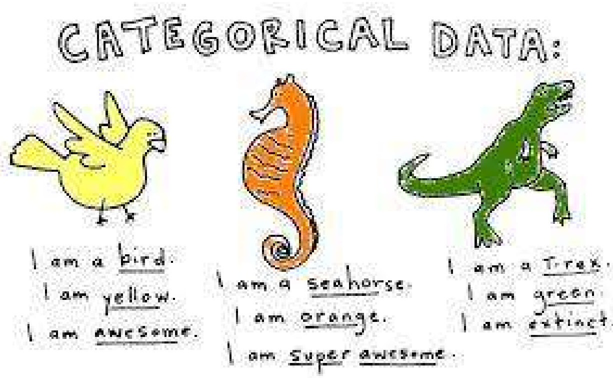
Relationships between Two Categorical Variables: Chi-Square Test



Both Variables Are Nominal

- If both variables are nominal/categorical → Chi-square test of independence
- Independence = no relationship



Cross-tabulation (crosstab) or Contingency Table

	Variable 1 Category 1	Variable 1 Category 2	Variable 1 Category 3
Variable 2 Category 1	n1	n3	n5
Variable 2 Category 2	n2	n4	n6

Example

We want to know whether gun ownership is related to one's marital status. We collect data from a random sample of 150 people; each person reports their marital status and whether they have a gun. Can we conclude that the rates of gun ownership depend on marital status in the population?

Data:

- never married with a gun = 15 people
- previously married with a gun = 15 people
- married with a gun=30 people
- never married no gun = 45 people
- previously married no gun = 30 people
- married no gun = 15 people

Organizing the Data: Step 1

- Decide on the dependent and independent variable: marital status → gun ownership? or gun ownership → marital status?
- “Independent” variable (marital status =columns of the table; “dependent” variable (gun ownership) = rows

Organizing the Data: Step 2

Make a table: C1, C2, C3 are the three values of the “independent” variable (marital status), they form the columns of the table; R1, and R2 are the two values of the “dependent” variable (gun ownership), they form the rows of the table

	C1: Never married	C2: Previously married	C3: Married
R1: Gun	R1C1: 15	R1C2: 15	R1C3: 30
R2: No gun	R2C1: 45	R2C2: 30	R2C3: 15

Organizing the Data: Step 3

- Add row totals(R_t), column totals (C_t), and column percentages
- The N for this table (which we label T, for "total") is 150

	Never married	Previously married	Married	Row Totals
Gun	15 25%	15 33.33%	30 66.67%	60 40%
No gun	45 75%	30 66.67%	15 33.33%	90 60%
Column Totals	60 100%	45 100%	45 100%	150 100%

Describing the Pattern in the Sample

- Column percentages = compare across columns
- In our sample, there are many more gun owners among the married (about 67% of all married people) than among those never married (25% of all never married)
- Previously married – in the middle (33% of previously married have guns)

Note on Percentages

- In this class, we will always use column percentages
- A variable that can be considered independent (cause) should be in columns, dependent (outcome) – in rows
- If no independent or dependent variable, you can just pick which one will be in columns
- How do we recognize column vs row percentages?
  - Columns sum up to 100% → these are column percentages
  - Rows sum up to 100% → these are row percentages

The Idea of Chi-Square Test

- The test is based on calculating expected frequencies
- They show what the table would look like if H0 was true
- Then we compare observed (O) and expected (E) frequencies
- If too far from each other → reject H0

Observed vs Expected

	C1 Never married	C2 Previously married	C3 Married	Row Totals (R_t)
Expected: R1: Gun Observed:	24 40% 15 25%	18 40% 15 33.33%	18 40% 30 66.67%	60 40%
Expected: R2: No gun Observed:	36 60% 45 75%	27 60% 30 66.67%	27 60% 15 33.33%	90 60%
Column Totals (C_t)	60 100%	45 100%	45 100%	150 (T) 100%

Our Example Step-by-Step

1. State hypotheses:

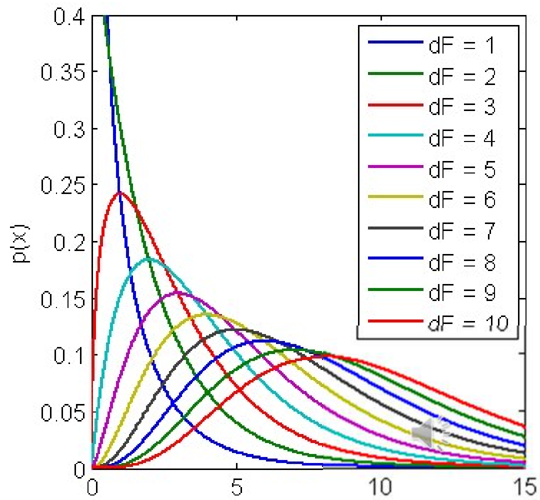
- H0: Marital status and gun ownership are independent (unrelated)
- H1: Marital status is related to gun ownership (the two variables are not independent)
- H0: $O_i = E_i$
- H1: $O_i \neq E_i$ (always non-directional)

2. Select alpha: 0.05

3. Test statistic: Chi-square (always one-tailed)

Chi-Square Distribution

- Discovered by a German statistician Friedrich Robert Helmert in 1875
- Rediscovered and popularized by Karl Pearson



Our Example Step-by-Step

4. Formula: $\chi^2 = \Sigma((O-E)^2/E)$

O = “observed”; the actual number of cases in a cell

E = the number of cases "expected" in the cell if we assume H0 (no relationship, or independence);

$E = C_t \times R_t / T$

R_t = the total for the R^{th} row

C_t = the total for the C^{th} column

T = total number of cases in the table

Calculations

Cell	E= $C_t * R_t/T$	O	O-E	$(O-E)^2$	$(O-E)^2/E$
C_1R_1	60x(60/150) = 60 x .4 = 24	15	15 - 24 = -9	81	81/24 = 3.375
C_2R_1	45x(60/150) = 45 x .4 = 18	15	15 - 18 = -3	9	9/18 = 0.500
C_3R_1	45x(60/150) = 45 x .4 = 18	30	30 - 18 = 12	144	144/18 = 8.000
C_1R_2	60x(90/150) = 60 x .6 = 36	45	45 - 36 = 9	81	81/36 = 2.250
C_2R_2	45x(90/150) = 45 x .6= 27	30	30 - 27 = 3	9	9/27 = 0.333
C_3R_2	45x(90/150) = 45 x .6 = 27	15	15 - 27 = -12	144	144/27 = 5.333
Σ	150		0		χ^2 = 19.791

$$\chi^2 = \Sigma((O-E)^2/E) = 19.791$$

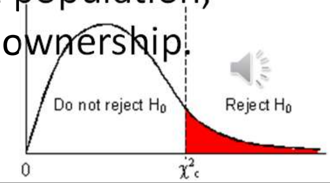
Our Example Step-by-Step

5. Use table B5 to find critical value: df = (R-1)x(C-1) = (2-1)x(3-1) = 2 [R = number of rows, C = number of columns] and alpha=.05 → 5.99

6. Compare computed value and critical value: 19.791 > 5.99

7. State your decision about H0: Reject H0

8. Conclusion: Based on our sample data, we are 95% certain that in the U.S. population, marital status is related to gun ownership.



In Other Words

- The departures from independence (O – E) are so large that chance would produce a chi-square value this large less than 5% of the time when randomly sampling from a population in which the two are independent

Post-Hoc Assessment: Residuals

Cell	$E = C_t * R_t / T$	O	O-E	$(O-E)^2$	$(O-E)^2/E$
C_1R_1	60x(60/150) = 60 x .4 = 24	15	15 - 24 = -9	81	81/24 = 3.375
C_2R_1	45x(60/150) = 45 x .4 = 18	15	15 - 18 = -3	9	9/18 = 0.500
C_3R_1	45x(60/150) = 45 x .4 = 18	30	30 - 18 = 12	144	144/18 = 8.000
C_1R_2	60x(90/150) = 60 x .6 = 36	45	45 - 36 = 9	81	81/36 = 2.250
C_2R_2	45x(90/150) = 45 x .6= 27	30	30 - 27 = 3	9	9/27 = 0.333
C_3R_2	45x(90/150) = 45 x .6 = 27	15	15 - 27 = -12	144	144/27 = 5.333

Focus on residuals larger than critical value of z

- for 90% confidence, 1.645
- for 95% confidence, 1.96
- for 99% confidence, 2.576

Post-Hoc Assessment: Residuals

- C1R1, C3R1, C1R2, C3R2 – residuals>1.96
- Married people are particularly likely to own guns (67% of them do) as compared to the never married people (only 25% of them do)
- Previously married are not significantly different from the other two groups

Important Things to Remember

- Chi-square test helps us determine whether there is a relationship between two variables overall; can't say which categories specifically are different (overall test, like ANOVA!)
- Need to have enough data per cell for chi-square test: fewer than 20% of cells should have EXPECTED counts of <5

Chi-Square in Stata: Problem

- Question: Are the opinions about legalizing marijuana (grass) linked to people's level of education (degree)?
- H0: Opinions about legalizing marijuana and people's level of education are unrelated.
- H1: Opinions about legalizing marijuana are related to people's level of education.
- H0: $O_i = E_i$
- H1: $O_i \neq E_i$

Variables and Percentages in Stata

- Command: `tab grass degree, col chi`
- grass = row variable, degree = column variable
- In Stata command:
  - first variable = row variable (use your "dependent" variable)
  - second variable = column variable (use your "independent" variable)
  - ask for column percentages → option col
  - If you wanted row percentages → option row

Chi-Square in Stata

`tab grass degree, col chi`

```
+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

    SHOULD |
 MARIJUANA |
   BE MADE |                  RS HIGHEST DEGREE
     LEGAL | LT HIGH S  HIGH SCHO  JUNIOR CO   bachelor   graduate |     Total
-----------+-------------------------------------------------------+----------
     legal |        74        277         51        110         74 |       586
           |     38.14      47.43      53.13      48.46      55.64 |     47.49
-----------+-------------------------------------------------------+----------
 NOT LEGAL |       120        307         45        117         59 |       648
           |     61.86      52.57      46.88      51.54      44.36 |     52.51
-----------+-------------------------------------------------------+----------
     Total |       194        584         96        227        133 |     1,234
           |    100.00     100.00     100.00     100.00     100.00 |    100.00

          Pearson chi2(4) =  11.6452   Pr = 0.020
```

- Chi-square = 11.645, p<.05
- We reject the null hypothesis of no relationship → 95% confident that opinions about legalizing marijuana are tied to level of education

Post-Hoc Assessment: Where Are the Differences?

- Analysis of residuals
- Need to install a user-written program in Stata (do once):

`net install tab_chi, from(http://fmwww.bc.edu/RePEc/bocode/t)`

- Which cells have the largest differences in Observed – Expected?
- Focus on residuals larger than critical value of z
  - for 90% confidence, 1.645
  - for 95% confidence, 1.96
  - for 99% confidence, 2.576

Post-Hoc Assessment: Residuals

tabchi grass degree, adj

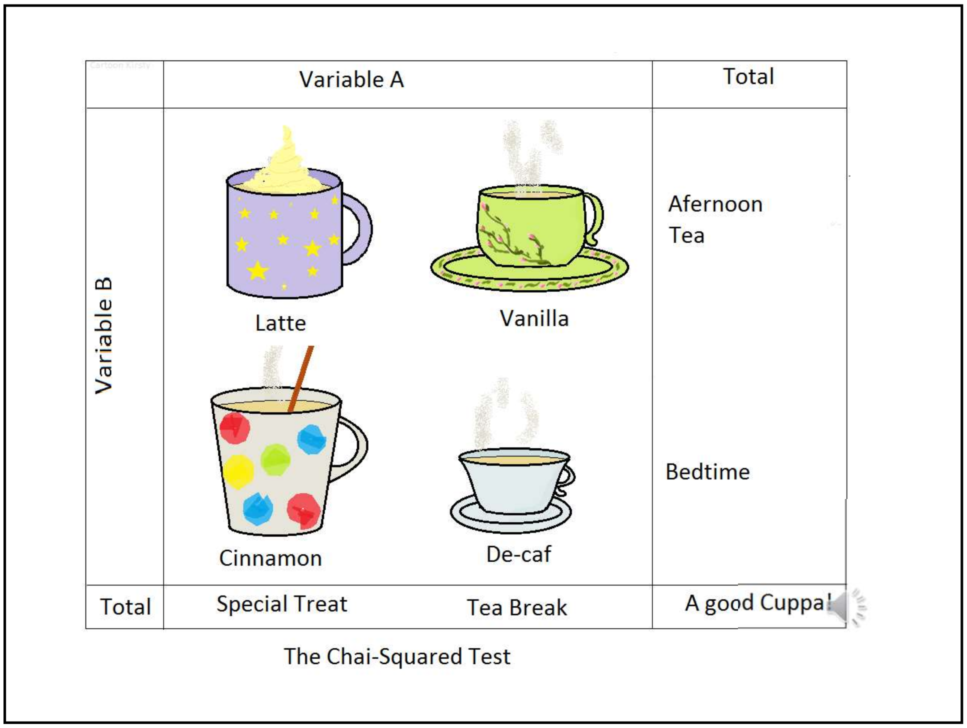
```
          observed frequency
          expected frequency
          adjusted residual

------------------------------------------------------------------------------------------
SHOULD     |
MARIJUANA  |
BE MADE    |                              RS HIGHEST DEGREE
LEGAL      | LT HIGH SCHOOL     HIGH SCHOOL  JUNIOR COLLEGE        bachelor         graduate
-----------+------------------------------------------------------------------------------
    legal  |             74             277              51             110               74
           |         92.126         277.329          45.588         107.797           63.159
           |         -2.839          -0.038           1.152           0.324            1.993
           |
NOT LEGAL  |            120             307              45             117               59
           |        101.874         306.671          50.412         119.203           69.841
           |          2.839           0.038          -1.152          -0.324           -1.993
------------------------------------------------------------------------------------------

          Pearson chi2(4) =  11.6452   Pr = 0.020
 likelihood-ratio chi2(4) =  11.7237   Pr = 0.020
```

Conclusion

- Those with less than high school degree are particularly likely to oppose legalizing marijuana in the population, while those with graduate degrees are particularly likely to favor the legalization


Variable A
Total
Variable B
Latte
Vanilla
Afernoon Tea
Cinnamon
De-caf
Bedtime
Total
Special Treat
Tea Break
A good Cuppa!
The Chai-Squared Test