Two-Way Tables: Chi-Square Tests

Edps/Psych/Soc 589

Carolyn J. Anderson

Department of Educational Psychology

ILLINOIS

Outline

- Overview and Definitions
- Chi-squared distribution
- Pearson's X^2 statistic
- Likelihood ratio test statistic
- Examples of
  - Independence
  - Homogeneous distributions
  - Unrelated classifications
  - Other
- Residuals
- (Partitioning Chi-square)
- Comments
- Practice

Definitions

- For a 2–way table, a null hypothesis H_o specifics a set of probabilities

$$H_O : \{\pi_{ij}\} \quad \text{for } i = 1, \ldots, I \quad \text{and } j = 1, \ldots, J$$

- "*Expected Frequencies*" are the values expected if the null hypothesis is true,

$$\mu_{ij} = n\pi_{ij}$$

- To test a null hypothesis, we compare the observed frequencies n_{ij} and the expected frequencies μ_{ij}:

$$\{n_{ij} - \mu_{ij}\}$$

- The test statistics are functions of observed and expected frequencies.
- If the null hypothesis is true, then the test statistics are distributed as chi-squared random variables so they are referred to as

"*Chi-Squared Tests*".

Null Hypotheses

The two most common tests/null hypotheses are

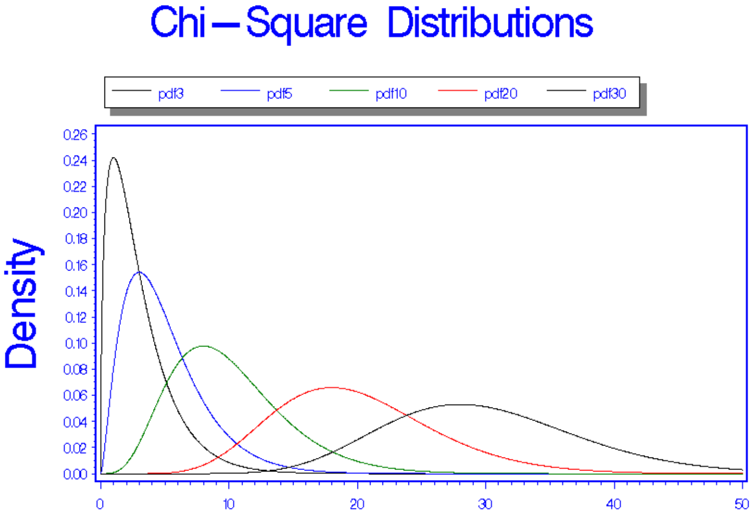
- Chi-squared test of *Independence*.
- Chi-squared test of *Homogeneous Distributions*.

The Chi–Squared Distribution

The “*Degrees of Freedom*”, df, completely specifies a chi-squared distribution.

- $0 \leq$ chi-squared random variable.
- The mean of a chi-squared distribution $= df$.
- The variance of a chi-squared distribution $= 2df$ and the **standard deviation** $= \sqrt{2df}$.
- The shape is skewed to the right.
- As df increase, the mean gets larger and the distribution more spread out.
- As df increase, the distribution becomes more “bell-shaped” (i.e., $df \rightarrow \infty$, $\chi^2_{df} \rightarrow \mathcal{N}$).

Picture of Chi–Squared Distributions



Pearson's Chi-Squared Statistic

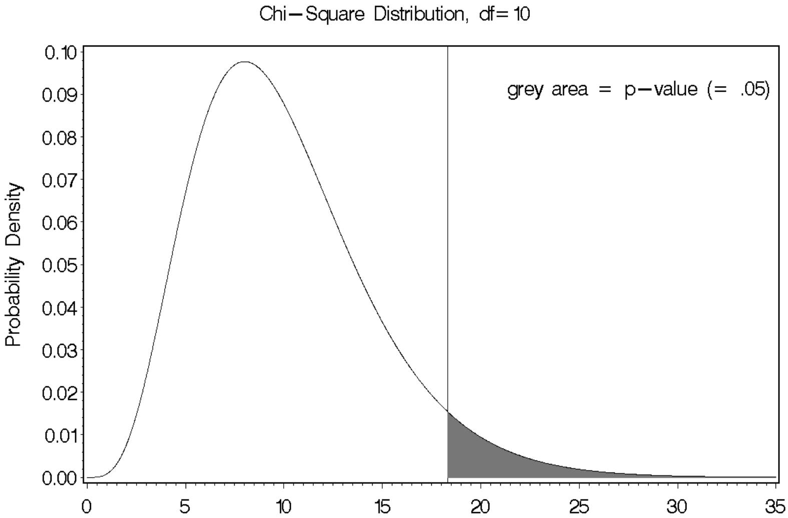
$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-\mu_{ij})^2}{\mu_{ij}}$$

- $0 \leq X^2$
- When $n_{ij} = \mu_{ij}$ for all (i, j), then $X^2 = 0$
- For "large" samples, X^2 has an approximate chi-squared distribution.

  A good rule: "Large" means $\mu_{ij} \geq 5$ for all (i, j).

- The p-value for a test is the right tail probability of X^2.

Chi–Squared Distribution and p-value



Likelihood Ratio Statistic

- Need the maximum likelihood estimates of parameters assuming
  - Null hypothesis is true (simpler, restrictions on parameters).
  - Alternative hypothesis is true (more general, no (or fewer) restrictions on parameters).
- The test statistic is based on

$$\Lambda = \frac{\text{maximum of the likelihood when parameters satisfy } H_O}{\text{maximum of likelihood when parameters are not restricted}}$$

- The numerator $\leq$ denominator $(\max L(H_O) \leq \max L(H_A))$.
- $0 \leq \Lambda \leq 1$.
- If $\max L(H_O) = \max L(H_A)$, then there is no evidence against H_O. (i.e., $\Lambda = 1$)
- The smaller the likelihood under H_O, the more evidence against H_O (i.e., the smaller Λ).

Likelihood Ratio Statistic for 2-way Table

The test statistic is $-2\log(\Lambda)$, which for contingency tables

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log(n_{ij}/\mu_{ij})$$

This is the "*likelihood ratio chi-squared statistic*".

Chi-Squared Test Hypotheses

1. Independence
2. Homogeneous Distributions
3. Unrelated Classifications
4. Other

- 1, 2 , & 3 are all tests of "no association" or "no relationship".
- 1 & 2 are the most common.
- 1, 2, & 3 all use the same formula to compute expected frequencies, but arrive at it from different starting points.
- 4 depends on the (substantive) hypothesis you are testing.
- These four test differ in terms of
  - Experimental procedure (i.e., sampling design)
  - The null and alternative hypothesis
  - Logic used to obtain estimates of expected frequencies assuming H_O is true.

Independence

Situation: Two response variables (either Poisson sampling or multinomial sampling)

Null Hypothesis: Two variables are statistically independent

Alternative Hypothesis: Two variables are dependent.

Definition of statistical independence,

$$H_O : \pi_{ij} \equiv \pi_{i+}\pi_{+j}$$

for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

Statistical dependence is not statistically independent

$$H_A : \pi_{ij} \neq \pi_{i+}\pi_{+j}$$

for at least one $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

To test this hypothesis, we assume H_O is true.

Expected Frequencies Under Independence

Given data, the observed marginal proportions p_{i+} and p_{+j} are the maximum likelihood estimates of π_{i+} and π_{+j}, respectively; that is,

$$\begin{aligned} \hat{\pi}_{i+} &= p_{i+} \\ \hat{\pi}_{+j} &= p_{+j} \end{aligned}$$

"Estimated Expected Frequencies" are

$$\begin{aligned} \hat{\mu}_{ij} &= n\hat{\pi}_{i+}\hat{\pi}_{+j} \\ &= n(n_{i+}/n)(n_{+j}/n) \\ &= \frac{n_{i+}n_{+j}}{n} \end{aligned}$$

Testing Independence

For "large" samples, to test the hypothesis that two variables are statistically independent, use either

$$G^2 = 2\sum_i\sum_j n_{ij}\log(n_{ij}/\hat{\mu}_{ij})$$

or

$$X^2 = \sum_i\sum_j \frac{(n_{ij}-\hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

and compare value to the appropriate chi-squared distribution.

General Rule for computing Degrees of Freedom:

> *The number of parameters specified under the alternative hypothesis minus the number of parameters specified under the null hypothesis.*

Computing Degrees of Freedom

$$df = (\#\text{ parameters in } H_A) - (\#\text{ parameters in } H_O)$$

- Null hypothesis has
  - $(I-1)$ unique parameters for the row margin, $\hat{\pi}_{i+}$.
  - $(J-1)$ unique parameters for the column margin, $\hat{\pi}_{+j}$.
- Alternative hypothesis has
  $(IJ-1)$ unique parameters. The only restriction on the parameters in the H_A is that the probabilities sum to 1.
- Degrees of Freedom so

$$df = (IJ-1) - [(I-1)+(J-1)] = (I-1)(J-1).$$

$df =$ the same number was came up with when we considered how many numbers we need to completely describe the association in an $I \times J$ table.

Example: Two Items from the 1994 GSS

- Item 1: A working mother can establish just as warm and secure a relationship with her children as a mother who does not work.
- Item 2: Working women should have paid maternity leave.

Observed Frequencies: n_{ij}

	Item2					
Item 1	Strongly Agree	Agree	Neither	Disagree	Strongly Disagree	
Strongly Agree	97	96	22	17	2	234
Agree	102	199	48	38	5	392
Disagree	42	102	25	36	7	212
Strongly Disagree	9	18	7	10	2	46
	250	415	102	101	16	884

Example: Estimated Expected Values

- Item 1: A working mother can establish just as warm and secure a relationship with her children as a mother who does not work.
- Item 2: Working women should have paid maternity leave.

Estimated Expected Frequencies: $\hat{\mu}_{ij} = n_{i+}n_{+j}/n$

	Item2					
	Strongly				Strongly	
Item 1	Agree	Agree	Neither	Disagree	Disagree	
Strongly Agree	66.18	109.85	27.00	26.74	4.24	234
Agree	110.86	184.03	45.23	44.79	7.10	392
Disagree	59.96	99.53	24.46	24.22	3.84	212
Strongly Disagree	13.01	21.60	5.31	5.26	0.83	46
	250	415	102	101	16	884

Example: Test Statistics

Statistic		*df*	Value	*p*-value
Pearson Chi-square	X^2	12	47.576	$< .001$
Likelihood Ratio Chi-square	G^2	12	44.961	$< .001$

What's the nature of the dependency? Residuals. . .

Residuals

- Raw Residuals: $n_{ij} - \hat{\mu}_{ij}$

  Problem: These tend to be large when $\hat{\mu}_{ij}$ is large.

  For Poisson random variables, mean = variance.
- Pearson Residuals or often called "standardized residuals"

$$\frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$$

	Strongly Agree	Agree	Neither	Disagree	Strongly Disagree
Strongly Agree	3.79	-1.32	$-.96$	-1.88	-1.09
Agree	$-.84$	1.10	.41	-1.01	$-.79$
Disagree	-2.32	.25	.11	2.39	1.61
Strongly Disagree	-1.11	$-.77$	.73	2.07	1.28

If the null hypothesis is true, then these should be approximately normally distributed with mean = 0, but . . .

Adjusted Residuals

- Problem with Pearson Residuals: The variance (standard deviation) of Pearson residuals is a bit too small.
- Adjusted Residuals or "Haberman residuals" (Haberman, 1973).

$$\frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - p_{i+})(1 - p_{+j})}}$$

If the null hypothesis is true, then these residuals have an asymptotic standard normal distribution.

	Strongly Agree	Agree	Neither	Disagree	Strongly Disagree
Strongly Agree	5.22	−2.12	−1.19	−2.33	−1.28
Agree	−1.33	2.03	.59	−1.44	−1.06
Disagree	−3.14	.39	2.92	2.92	1.82
Strongly Disagree	−1.35	−1.09	.80	2.25	1.33

Residuals and SAS

-
  ```
  DATA GSS94;
     INPUT item1 item2 count;
     DATALINES;
   1   1   97
   1   2   96
   ⋮   ⋮    ⋮
   4   5    2
  ```
- PROC FREQ gives raw residuals (DEVIATION option) and "cell contribution" to Pearson chi-squared statistic, which are Squared Pearson residuals (CELLCH2 option).
  ```
  PROC FREQ;
  TABLES item1*item2 / CELLCH2;
  ```
- PROC GENMOD gives Adjusted residuals and lots more.
  ```
  PROC GENMOD;
  CLASS item1 item2;
  MODEL count = item1 item2 / link=log dist=P obstats;
  ```
  "AdjChiRes" are the adjusted chi-square (Haberman) residuals.

Another Example of Independence

“Specifically, there were about $26,000$ applications to the Urbana campus this year. About $18,000$ applicants were admitted using the 69% admissions rate cited in the article. The 160 “I list” applicants had a 77% admissions rate, according to the Tribune. This translates into the admission of 13 more applicants on the Category I list admissions rate versus the standard rate.”

Ignoring the ethical question, is 13 more applicants admitted statistically significant? In other words, is 77% statistically different from 69%?

Let’s look at the statistical question using all methods that we’ve discussed so far.

Admission Scandal Results

Binomial test of whether admission rate from I list is same as general admission rate. The results are significant whether use asymptotic test or binomial exact tests.

I-list: H_o: Probability of Admission of I list $= .69$
(i.e., the proportion general admission)

admit	Frequency	Percent	Cumulative Frequency	Cumulative Percent
yes	123	76.88	123	76.88
no	37	23.13	160	100.00

	Large Sample	Exact Binomial
Proportion	0.7688	
ASE	0.0333	
95% Lower Conf Limit	0.7034	0.6956
95% Upper Conf Limit	0.8341	0.8317

Results Continued

Asymptotic (large sample) Test of H0: Proportion $= 0.69$

ASE under H0	0.0366
Z	2.1538
One-sided Pr $> Z$	0.0156
Two-sided Pr $> \|Z\|$	0.0313

Sample Size $= 160$

		95% Confidence	
Statistic	Value	Interval	
Difference of Proportions	.076	0.009	0.144
Odds ratio	1.478	1.022	2.136
Relative Risk	1.110	1.020	1.209
Correlation	0.013		

Test of Independence

	Admission		
	yes	no	Total
I list	123	37	160
general	18000	8000	26000
Total	18123	8037	26160

Statistics for Table of List by Admission

Statistic	DF	Value	Prob
Chi-Square	1	4.3659	0.0367
Likelihood Ratio Chi-Square	1	4.6036	0.0319
Continuity Adj. Chi-Square	1	4.0141	0.0451
Mantel-Haenszel Chi-Square	1	4.3657	0.0367
Phi Coefficient		-0.0129	

Homogeneous Distributions

Situation: Sample from different populations and observe classification on a response variable. The explanatory variable defines the populations and the number from each population is determined by the researcher.

i.e., independent Binomial/Multinomial sampling.

Null Hypothesis: The distributions of responses from the different populations are the same.

Alternative Hypothesis: The distributions of responses from the different populations are different.

Example of Homogeneous Distributions

Effectiveness of Vitamin C for prevention of common cold.

	Outcome		
	Cold	No Cold	
vitamin C	$17/139 = .12$	$122/139 = .88$	$.12 + .88 = 1.00$
placebo	$31/140 = .22$	$109/140 = .78$	$.22 + .78 = 1.00$
	$48/279 = .17$	$231/279 = .83$	$.17 + .83 = 1.00$

Chi-Square Test for Homogeneous Distributions

The null and alternative hypotheses are:

$$H_O : \pi_1 = \pi_2 \qquad \text{versus} \qquad H_A : \pi_1 \neq \pi_2$$

and more generally,

$$H_O : \pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} = \pi_{+j} \qquad \text{versus} \qquad H_A : \pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} \neq \pi_{+j}$$

for all $i, \ldots, I$ and $j = 1, \ldots, J$.

Assuming H_O is true, the conditional distributions of the response variable given the explanatory variable should all be equal and they should equal the marginal distribution of the response variable; that is,

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} = \pi_{+j}$$

Estimated Expected Frequencies

- Expected frequencies equal
$$\mu_{ij} = n_{i+}\pi_{+j}$$
where n_{i+} is given (fixed by design).
- Given data, our (maximum likelihood) estimates of the marginal probabilities of responses are
$$\hat{\pi}_{j|i} = \hat{\pi}_{+j} = p_{+j} = n_{+j}/n$$
- **Estimated Expected Frequencies** are
$$\begin{aligned}\hat{\mu}_{ij} &= n_{i+}\hat{\pi}_{+j} \\ &= n_{i+}(n_{+j}/n) \\ &= \frac{n_{i+}n_{+j}}{n}\end{aligned}$$

Degrees of Freedom

for test of homogeneous distributions

Null Hypothesis has

$(J-1)$ unique parameters — the $\hat{\pi}_{+j}$, which sum to 1.

Alternative Hypothesis has

$I(J-1)$ unique parameters — for I values of $\hat{\pi}_{j|i}$, which must sum to 1.

Degrees of Freedom equal

$$df = I(J-1) - (J-1) = (I-1)(J-1)$$

Same as for testing independence.

Example: Effectiveness of Vitamin C

Observed Frequencies

	Outcome		
	Cold	No Cold	
vitamin C	17	122	139
placebo	31	109	140
	48	231	279

Expected Values

	Outcome		
	Cold	No Cold	
vitamin C	23.91	115.09	139
placebo	24.09	115.91	140
	48	231	279

Effectiveness of Vitamin C (continued)

Test Statistic		df	Value	p–value
Pearson Chi-Square	X^2	1	4.811	.03
Likelihood Ratio Chi-Square	G^2	1	4.872	.03

Adjusted Residuals

	Outcome	
	Cold	No Cold
vitamin C	-2.31	2.17
placebo	2.10	-2.22

Summary regarding Effectiveness of Vitamin C

$$
\begin{aligned}
\text{Difference of Proportions} &= -.10 & 95\%\ \text{CI}\ (-.19, -.01) \\
\text{Relative Risk} &= .552 & 95\%\ \text{CI}\ (.32, .93) \\
\text{Odds ratio} &= .490 & 95\%\ \text{CI}\ (.26, .93) \\
\text{Correlation} &= -.131 &
\end{aligned}
$$

Test Statistic		df	Value	p–value
Pearson Chi-Square	X^2	1	4.811	.03
Likelihood Ratio Chi-Square	G^2	1	4.872	.03

Adjusted Residuals

	Outcome	
	Cold	No Cold
vitamin C	−2.31	2.17
placebo	2.10	−2.22

Unrelated Classification

Situation: Both margins are fixed by design. The sample can be considered the population.

Example: 1970 draft lottery of 19–26 year olds (Fienberg, 1971). Each day of the year (including Feb 29) was typed on a slip of paper and inserted into a capsule. The capsules were mixed and were assigned a "drawing number" according to their position in the sequence of capsules picked from a bowl. The cross-classification of months by drawing number where drawing numbers are grouped into thirds.

		Drawing Numbers			
		1–122	123–244	245–366	Totals
	Jan	9	12	10	31
	Feb	7	12	10	29
	March	5	10	16	31
	April	8	8	14	30
	May	9	7	15	31
Month	June	11	7	12	30
	July	12	7	12	31
	Aug	13	7	11	31
	Sept	10	15	5	30
	Oct	9	15	7	31
	Nov	12	12	6	30
	Dec	17	10	4	31

Hypothesis of Unrelated Classification

Null Hypothesis: The row and column classifications are unrelated.

H_O: Drawing was random; that is, there is no relationship between drawing number and month of birth

Alternative Hypothesis: The row and column classifications are related.

H_A: Drawing was not random; there is a relationship between drawing number and month of birth.

Expected Values

The logic to find the expected values follows that of homogeneous distributions.

- n_{i+} fixed for rows
- n_{+j} fixed for columns
- $n_{+j}/n =$ proportion in column j.

If the null hypothesis is true, then expected frequencies μ_{ij} are

$$\begin{aligned} \mu_{ij} &= (\#\text{ in row } i)(\text{proportion in column } j) \\ &= n_{i+}(n_{+j}/n) \\ &= \frac{n_{i+}n_{+j}}{n} \end{aligned}$$

Degrees of Freedom $= (I-1)(J-1)$.

Example: 1970 Draft

Statistic		df	Value	p–value
Pearson chi-square	X^2	22	37.540	.02
Likelihood ratio chi-square	G^2	22	38.669	.02

What's the nature of the association?

Adjusted Residuals:

		Drawing Number		
		1-122	123–244	245–366
	Jan	$-.52$	.64	$-.12$
	Feb	-1.08	.93	.15
	March	-2.11	$-.15$	2.27
	April	$-.80$	$-.83$	1.63
	May	$-.52$	-1.35	1.87
Month	June	.42	-1.23	.82
	July	.68	-1.35	.68
	Aug	1.07	-1.35	.28
	Sept	.01	2.00	-2.01
	Oct	$-.52$	1.83	-1.32
	Nov	.68	1.04	-1.72
	Dec	2.67	$-.15$	$-.251$

Explanation. . .

Other Hypotheses

These can either be

- Simpler than independence. (Example on following slides)
- More complex. (e.g., symmetry and others . . . later in the semester).

Example of Other Hypothesis

H_o specifies the distribution of one or more of the margins.

Example: (from Wickens, 1989). Suppose there are 2 approaches to solving a problem & the answer is either correct or incorrect.

		Answer		
		Correct	Incorrect	
Method	A			$n/2 = .5$
	B			$n/2 = .5$
				n

- H_O: Independence and equal number of students should choose each method.
- H_A: Method and Answer are dependent and/or unequal number of students choose each method.

The expected frequencies $= n_{i+}n_{+j}/n = n_{+j}/2$.

Another Other Example

Testing Mendal's Theories of natural inheritance

Review:

$$\begin{aligned} Y &= \text{yellow} \longrightarrow \text{dominant trait} \\ g &= \text{green} \longrightarrow \text{recessive trait} \end{aligned}$$

- 1st generation: All plants have genotype Yg and phenotype is yellow.
- 2nd generation: Possible genotypes and phenotypes are

Genotype	Phenotype	Assuming random
YY	yellow	25%
Yg	yellow	25%
gY	yellow	25%
gg	green	25%

Theory predicts that 75% will be yellow and 25% will be green.

Partitioning Chi-Square

Another way to investigate the nature of association

The sum of independent chi-squared statistics are themselves chi-squared statistics with degrees of freedom equal to the sum of the degrees of freedom for the individual statistics.

For example, if

$$\begin{array}{rl}
Z_1^2 & \text{is chi-squared with } df_1 = 1 \\
\text{and } Z_2^2 & \text{is chi-squared with } df_2 = 1 \\
\hline \\
\text{then } (Z_1^2 + Z_2^2) & \text{is chi-squared with } df = df_1 + df_2 = 2
\end{array}$$

. . . and (of course) Z_1^2 and Z_2^2 are independent.

"**Partitioning chi-squared**" uses this fact, but in reverse:

We start with a chi-squared statistic with $df > 1$ and break it into component parts, each with $df = 1$

Partitioning Chi-Square by Example

Why partition? Partitioning chi–squared statistics helps to show that an association which was significant for the overall table primarily reflects differences between some categories and/or some groups of categories.

Demonstrate the method by example by partitioning G^2 for a 3×3 table into $(3-1)(3-1) = 4$ parts.

Example: A sample of psychiatrists were classified with respect to their school of psychiatric thought and their beliefs about the origin of schizophrenia. (Agresti, 1990; Gallagher, et al, 1987).

School of	Origin of Schizophrenia		
Psychiatric Thought	Biogenic	Environmental	Combination
Eclectic	90	12	78
Medical	13	1	6
Psychoanalysis	19	13	50

Check for Relationship & Then Partition

First we check if these two variables are independent or not.

Statistic	df	Value	p–value
X^2	4	22.378	$< .001$
G^2	4	23.036	$< .001$

School of Psychiatric Thought	Origin of Schizophrenia		
	Biogenic	Environmental	Combination
Eclectic	90	12	78
Medical	13	1	6
Psychoanalysis	19	13	50

Sub-table 1:

	Bio	Env
Eclectic	90	12
Medical	13	1

$\longrightarrow df = 1$
$G^2 = .294$
p-value $= .59$

Sub-table 2:

	Env	Com
Eclectic	12	78
Medical	1	6

$\longrightarrow df = 1$
$G^2 = .005$
p-value $= .94$

Sub-table 3:

	Bio	Env
Medical	13	1
Psychan	19	13

$\longrightarrow df = 1$
$G^2 = 6.100$
p-value $= .01$

Sub-table 4:

	Env	Com
Medical	1	6
Psychoan	13	50

$\longrightarrow df = 1$
$G^2 = .171$
p-value $= .68$

But. . . $.294 + .005 + 6.100 + .171 = 6.570 \neq 23.036$

Independent Component Tables

A general method proposed by Lancaster (1949).

$\sum_{a<i}\sum_{b<j} n_{ab}$	$\sum_{a<i} n_{aj}$
$\sum_{b<j} n_{ib}$	n_{ij}

Using this with our example:

School of Psychiatirc Thought	Origin of Schizophrenia Biogenic	 Environmental	 Combination
Eclectic	90	12	78
Medical	13	1	6
Psychoanalysis	19	13	50

Sub-Table 1:

	Bio	Env
Eclectic	90	12
Medical	13	1

$$\begin{aligned} \longrightarrow df &= 1 \\ G^2 &= .294 \\ X^2 &= .264 \\ \hat{\theta} &= .577 \end{aligned}$$

Sub-Table 2:

	Bio +Env	Com
Eclectic	102	78
Medical	14	6

$$\begin{aligned} \longrightarrow df &= 1 \\ G^2 &= 1.359 \\ X^2 &= 1.314 \\ \hat{\theta} &= .560 \end{aligned}$$

Sub-Table 3:

	Bio	Env
Ecl+Med	103	13
Psychoan	19	13

$$\begin{aligned} \longrightarrow df &= 1 \\ G^2 &= 12.953 \\ X^2 &= 14.989 \\ \hat{\theta} &= 5.421 \end{aligned}$$

Sub-Table 4:

	Bio +Env	Com
Ecl+Med	116	84
Psychoan	32	50

$$\begin{aligned} \longrightarrow df &= 1 \\ G^2 &= 8.430 \\ X^2 &= 8.397 \\ \hat{\theta} &= 2.158 \end{aligned}$$

Description of Association

from Agresti (1990):

> *"The psychoanalytic school seems more likely than other schools to ascribe the origins of schizophrenia as being a combination. Of those who chose either the biogenic or environmental origin, members of the psychoanalytic school were somewhat more likely than the other schools to chose the environmental origin."*

With this partitioning, likelihood ratio chi-squared statistics add up to G^2 for full table

$$.294 + 1.359 + 12.953 + 8.430 = 23.036$$

Pearson X^2's don't add up to value in full table:

$$.264 + 1.314 + 14.989 + 8.397 = 24.964 \neq 22.378$$

. . . but this is OK because they are not suppose to add up exactly.

Necessary Conditions for Partitioning

You are not restricted to use the method proposed by Lancaster; however, for partitioning to lead to a full decomposition of G^2 the following are necessary conditions (Agresti, 1990)

- The degrees of freedom for the sub-tables must sum to the degrees of freedom for the original table.
- Each cell count in the original table must be a cell in one and only one sub-table.
- Each marginal total of the original table must be a marginal total for one and only one sub-table.

A better approach to studying the nature of association — estimating parameters that describe aspects of association and models the represent association.

Summary Comments on Chi-Squared Tests

- Chi–squared tests of no association only indicate evidence there is against H_O.
- Chi–squared tests are limited to "large" samples.
  - As n increases relative to the size of the table, the distribution of X^2 and G^2 are better approximated by the chi–squared distribution.
  - Since the sampling distributions of X^2 and G^2 are only approximated by chi–square distributions, p–values should only be reported to 2 decimal places (3 at most).
  - The distribution of X^2 converges faster to chi–squared than the distribution of G^2. (More about this later in semester).
  - There are small sample methods available — "exact tests"
- The tests that we've discussed have not used additional information that we may have about the variables.
- In the case of ordinal variables, there are better methods.

SAS

```
data gss;
input FECHLD MAPAID count ;
label FECHLD='Mother working doesnt hurt children'
MAPAID='Should working women have paid maternity';
datalines;
1 1 97
1 2 96
...
4 5 2
;
proc freq order=data;
weight count;
tables FECHLD*MAPAID/nopercent norow nocol expected chisq ;
run;
```

R—Many options

File “gss_data.txt” on web-site:

```
fechld mapaid count
1 1 97
1 2 96
...
# Input data
(gss ← read.table("gss_data.txt",header=TRUE) )
(gss.tab ← xtab(count ∼ fechld + mapaid,data=gss) )
addmargins(gss.tab)
# Pearson chi-square statistic, df, and pvalue:
chisq.test(gss.tab,correct=FALSE)
```

R—Many options

```
# set libraries
library(MASS)
# Pearson and Likelihood ratio statistics, df, and pvalues
# data frame (case data or ‘‘data frame’’)
loglm(count ∼ fechld + mapaid, data=gss)
# Using table format
loglim( ∼ fechld + mapaid, data=gss.tab)
```

R—Many options

```
# Or Fit model of independence
fechld.f ← as.factor(gss$fechld)
mapaid.f ← as.factor(gss$mapaid)
model.glm ← glm(count ∼ fechld.f + mapaid.f, data=gss,
family=poisson)
summary(model.glm)
# -- to get p-values
1-pchisq(44.961,12)
# --- Phi coeffcient from psych package
phi(gss, digits=2)
```

Practice: 2018 GSS Items

- "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if the family has a very low income and cannot afford any more children".
- "We hear a lot of talk these days about liberals and conservatives. I'm going to show you a seven-point scale on which the political views that people might hold are arranged from extremely liberal–point 1–to extremely conservative–point 7. Where would you place yourself on this scale?"

Note: I deleted the "moderates" and collapsed liberals and conservatives (later we can look at full scale).

Practice: The Data

Support Abortion?	Political View		
	Liberal	Conservative	Total
yes	337	137	474
no	313	156	469
Total	493	450	943

Practice: Do following

1. What are the null and alternative hypotheses?
2. Use G^2, test (including interpretation).
3. Use X^2, test (including interpretation).
4. Confirm that $r = \sqrt{X^2/n}$

Practice: Getting Started

```
 library(Epi)     # for twoby2 comand
 library(MASS)    # for loglm
 library(vcd)     # for assocstats
 library(psych)   # to get phi coefficient
# --- create data frame --- (n x p) format
var.levels ← expand.grid(abortion=c("yes","no"),
pview=c(‘‘Liberal’’,‘‘Conservative’’))
( gss <- data.frame(var.levels, count=c(337, 313, 137, 156)
) )
```