STAT 705: Analysis of Contingency Tables

Adapted from Timothy Hanson

Department of Statistics, University of South Carolina

Stat 705: Analysis of Contingency Tables

Outline of Part I: models and parameters

- Basic ideas: contingency table, cross-sectional or fixed margins (multinomial and product multinomial sampling), independence.
- Various types of studies leading to contingency tables.
- Two groups, 2×2 table: odds ratio, relative risk, and difference in proportions.
- $I \times J$ table with ordinal outcomes: ordinal association.

Contingency tables & their distributions

Let X and Y be categorical variables measured on a subject with I and J levels respectively.

Each subject sampled will have an associated (X, Y); e.g. $(X, Y) =$ (female, Republican). For the gender variable X, $I = 2$, and for the political affiliation Y, we might have $J = 3$.

Say n individuals are sampled and cross-classified according to their outcome (X, Y). A *contingency table* places the raw number of subjects falling into each cross-classification category into the table cells. We call such a table an $I \times J$ table.

If we relabel the category outcomes to be integers $1 \leq X \leq I$ and $1 \leq Y \leq J$ (i.e. turn our experimental outcomes into *random variables*), we can simplify notation: n_{ij} is the number of individuals with $X = i$ and $Y = j$.

Abstract contingency table

In the abstract, a contingency table looks like:

n_{ij}	$Y=1$	$Y=2$	$\cdots$	$Y=J$	Totals
$X=1$	n_{11}	n_{12}	$\cdots$	n_{1J}	n_{1+}
$X=2$	n_{21}	n_{22}	$\cdots$	n_{2J}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$X=I$	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	n_{I+}
Totals	n_{+1}	n_{+2}	$\cdots$	n_{+J}	$n=n_{++}$

If subjects are randomly sampled from the population and cross-classified, both X and Y are random and (X,Y) has a bivariate discrete joint distribution. Let $\pi_{ij}=P(X=i,Y=j)$, the probability of falling into the $(i,j)^{th}$ (row,column) in the table.

Example of 3×3 table

From Chapter 2 in Christensen (1997) we have a sample of $n = 52$ males aged 11 to 30 years with knee operations via arthroscopic surgery. They are cross-classified according to $X = 1, 2, 3$ for injury type (twisted knee, direct blow, or both) and $Y = 1, 2, 3$ for surgical result (excellent, good, or fair-to-poor).

n_{ij}	Excellent	Good	Fair to poor	Totals
Twisted knee	21	11	4	36
Direct blow	3	2	2	7
Both types	7	1	1	9
Totals	31	14	7	$n = 52$

with theoretical probabilities:

π_{ij}	Excellent	Good	Fair to poor	Totals
Twisted knee	π_{11}	π_{12}	π_{13}	π_{1+}
Direct blow	π_{21}	π_{22}	π_{23}	π_{2+}
Both types	π_{31}	π_{32}	π_{33}	π_{3+}
Totals	π_{+1}	π_{+2}	π_{+3}	$\pi_{++} = 1$

Marginal probabilities

The marginal probabilities that $X = i$ or $Y = j$ are

$$P(X = i) = \sum_{j=1}^{J} P(X = i, Y = j) = \sum_{j=1}^{J} \pi_{ij} = \pi_{i+}.$$

$$P(Y = j) = \sum_{i=1}^{I} P(X = i, Y = j) = \sum_{i=1}^{I} \pi_{ij} = \pi_{+j}.$$

A "+" in place of a subscript denotes a *sum* of all elements over that subscript. We must have

$$\pi_{++} = \sum_{i=1}^{I} \sum_{j=1}^{J} \pi_{ij} = 1.$$

The counts have a multinomial distribution $\mathbf{n} \sim \text{mult}(n_{++}, \boldsymbol{\pi})$ where $\mathbf{n} = [n_{ij}]_{I \times J}$ and $\boldsymbol{\pi} = [\pi_{ij}]_{I \times J}$.

$$P(\mathbf{n}) = \binom{n_{++}}{\mathbf{n}} \prod_{(i,j)} \pi_{ij}^{n_{ij}}$$

Product multinomial table

Often the marginal counts for X or Y are fixed by design. For example in a case-control study, a fixed number of cases (e.g. people w/ lung cancer) and a fixed number of controls (no lung cancer) are sampled. Then a risk factor or exposure Y is compared among cases and controls within the table. This results in a separate multinomial distribution for each level of X.

For the I multinomial distributions, the conditional probabilities of falling into $Y = j$ must sum to one for *each* level of $X = i$:

$$\sum_{j=1}^{J} \pi_{j|i} = 1 \text{ for } i = 1, \ldots, I, \text{ where } \pi_{j|i} = \pi_{j|i}^{Y|X} = P(Y = j|X = i).$$

How is $(n_{1+}, \ldots, n_{I+})$ distributed?

Clinical trial example

The following 2×3 contingency table is from a report by the Physicians' Health Study Research Group on $n = 22,071$ physicians that took either a placebo or aspirin every other day.

	Fatal attack	Nonfatal attack	No attack
Placebo	18	171	10,845
Aspirin	5	99	10,933

Here we have placed the probabilities of each classification into each cell:

	Fatal attack	Nonfatal attack	No attack
Placebo	$\pi_{1\|1}$	$\pi_{2\|1}$	$\pi_{3\|1}$
Aspirin	$\pi_{1\|2}$	$\pi_{2\|2}$	$\pi_{3\|2}$

The row totals $n_{1+} = 11,034$ and $n_{2+} = 11,037$ are fixed and thus $\pi_{1|1} + \pi_{2|1} + \pi_{3|1} = 1$ and $\pi_{1|2} + \pi_{2|2} + \pi_{3|2} = 1$.

We want to compare probabilities in each column.

Independence

When (X, Y) are jointly distributed, X and Y are independent if

$$P(X = i, Y = j) = P(X = i)P(Y = j) \text{ or } \pi_{ij} = \pi_{i+}\pi_{+j}.$$

Let

$$\pi_{i|j} = \pi_{i|j}^{X|Y} = P(X = i|Y = j) = \pi_{ij}/\pi_{+j}$$

and

$$\pi_{j|i} = \pi_{j|i}^{X|Y} = P(Y = j|X = i) = \pi_{ij}/\pi_{i+}.$$

Then independence of X and Y implies

$$P(X = i|Y = j) = P(X = i) \text{ and } P(Y = j|X = i) = P(Y = j).$$

The probability of any given column response is the same for each row. The probability for any given row response is the same for each column.

Case-control studies

	Case	Control
Smoker	688	650
Non-smoker	21	59
Total	709	709

In a case/control study, fixed numbers of cases n_1 and controls n_2 are (randomly) selected and exposure variables of interest recorded. In the above study we can compare the relative proportions of smokers within those patients admitted with lung cancer (cases) and within those matched patients not admitted with lung cancer (controls). We can measure association between smoking and lung cancer, but cannot infer causation. These data were collected "after the fact." Such data are usually cheap and easy to get. Above: some very old lung cancer data (Agresti, 2013).

These designs (and clinical trials) always yield product multinomial sampling.

Other types of studies

- Prospective studies start with a sample of subjects and observes them through time.
  - Clinical trial randomly allocates "smoking" and "non-smoking" treatments to experimental units and then sees who ends up with lung cancer or not. Problem with ethics here.
  - A cohort study simply follows subjects after letting them assign their own treatments (i.e. smoking or non-smoking) and records outcomes.
- A cross-sectional design samples n subjects from a population and cross-classifies them.
- Are each of these multinomial or product multinomial?

Comparing two proportions

Let X and Y be dichotomous. Let $\pi_1 = P(Y = 1|X = 1)$ and let $\pi_2 = P(Y = 1|X = 2)$.

The *difference* in the probability of $Y = 1$ when $X = 1$ versus $X = 2$ is $\pi_1 - \pi_2$.

The *relative risk* π_1/π_2 is more informative for rare outcomes. However it may also *exaggerate* the effect of $X = 1$ versus $X = 2$ as well and cloud issues.

Odds ratios

The *odds* of success (say $Y = 1$) versus failure ($Y = 2$) are $\Omega = \pi/(1-\pi)$ where $\pi = \pi_{+1} = P(Y = 1)$. When someone says "3 to 1 odds the Gamecocks will win", they mean $\Omega = 3$ which implies the probability the Gamecocks will win is 0.75, from $\pi = \Omega/(\Omega + 1)$. Odds measure the relative rates of success and failure.

An *odds ratio* compares relatives rates of success (or disease or whatever) across two exposures $X = 1$ and $X = 2$:

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}.$$

Odds ratios are always positive and a ratio > 1 indicates the relative rate of success for $X = 1$ is greater than for $X = 2$. However, the odds ratio gives *no information* on the probabilities $\pi_1 = P(Y = 1|X = 1)$ and $\pi_2 = P(Y = 1|X = 2)$.

Odds ratio, cont.

Different values for these parameters can lead to the same odds ratio.

Example: $\pi_1 = 0.833$ & $\pi_2 = 0.5$ yield $\theta = 5.0$. So does $\pi_1 = 0.0005$ & $\pi_2 = 0.0001$.

· One set of values might imply a different decision than the other, but $\theta = 5.0$ in both cases.

· Here, the relative risk is about 1.7 and 5 respectively.

· Note that when dealing with a rare outcome, where $\pi_i \approx 0$, the relative risk is approximately equal to the odds ratio.

Odds ratio, cont.

When $\theta = 1$ we must have $\Omega_1 = \Omega_2$ which further implies that $\pi_1 = \pi_2$ and hence Y does not depend on the value of X. If (X, Y) are both random then X and Y are stochastically independent.

An important property of odds ratio is the following:

$$\begin{aligned} \theta &= \frac{P(Y=1|X=1)/P(Y=2|X=1)}{P(Y=1|X=2)/P(Y=2|X=2)} \\ &= \frac{P(X=1|Y=1)/P(X=2|Y=1)}{P(X=1|Y=2)/P(X=2|Y=2)} \end{aligned}$$

Let's verify this formally.

This implies that for the purposes of estimating an odds ratio, it *does not matter* if data are sampled prospectively, retrospectively, or cross-sectionally. The common odds ratio is estimated $\hat{\theta} = n_{11}n_{22}/[n_{12}n_{21}]$.

Case/control and the odds ratio

	Case	Control
Smoker	688	650
Non-smoker	21	59
Total	709	709

Recall there are $n_1 = n_2 = 709$ lung cancer cases and (non-lung cancer) controls. The margins are fixed and we have product multinomial sampling.

We can estimate $\pi_{1|1} = P(X = 1|Y = 1) = n_{11}/n_{+1}$ and $\pi_{1|2} = P(X = 1|Y = 2) = n_{12}/n_{+2}$ but not $P(Y = 1|X = 1)$ or $P(Y = 1|X = 2)$–we would need $P(X = 1), P(X = 2)$.

However, for the purposes of estimating θ it does not matter!

For the lung cancer case/control data,
$\hat{\theta} = 688 \times 59/[21 \times 650] = 3.0$ to one decimal place.

Odds of lung cancer, cont.

- The odds of being a smoker is 3 times greater for those that develop lung cancer than for those that do not.
- The odds of developing lung cancer is 3 times greater for smokers than for non-smokers.

The second interpretation is more relevant when deciding whether or not you should take up recreational smoking.

Note that we *cannot* estimate the relative risk of developing lung cancer for smokers $P(Y = 1|X = 1)/P(Y = 1|X = 2)$.

Formally comparing groups

You should convince yourself that the following statements are equivalent:

- $\pi_1 - \pi_2 = 0$, the difference in proportions is zero.
- $\pi_1/\pi_2 = 1$, the relative risk is one.
- $\theta = [\pi_1/(1-\pi_1)]/[\pi_2/(1-\pi_2)] = 1$, the odds ratio is one.

All of these imply that there is no difference between groups for the outcome being measured, i.e. Y is independent of X, written $Y \perp X$.

Estimation of $\pi_1 - \pi_2$, π_1/π_2, and θ are coming up...

A measure of ordinal trend: concordant and discordant pairs

Another single statistic that summarizes association for ordinal (X, Y) uses the idea of concordant and discordant pairs. Consider data from the 2014 General Social Survey:

	Education			
Family Income	8th Grade or less	High School	College	Graduate School
$<$ \$10K	2011	4381	1789	220
\$10K – \$15K	706	3238	1474	203
\$15K – \$20K	366	2621	1423	230
\$20K – \$25K	696	2519	1591	305
$>$ \$25K	104	10175	12981	4229

Family income tends to increase with education. How to summarize this association?

One measure of positive association is the probability of concordance.

Concordance

Consider two independent, randomly drawn individuals (X_1, Y_1) and (X_2, Y_2). This pair is concordant if either $X_1 < X_2$ and $Y_1 < Y_2$ simultaneously, or $X_1 > X_2$ and $Y_1 > Y_2$ simultaneously. An example would be (\$15K-\$20K, High School) and ($>$ \$25K, College). This pair indicates some measure of increased family income with education.

The probability of concordance Π_c is (note all the ties):

$$\begin{aligned} P(X_2 > X_1, Y_2 > Y_1 \text{ or } X_2 < X_1, Y_2 < Y_1) &= P(X_2 > X_1, Y_2 > Y_1) \\ &\quad + P(X_2 < X_1, Y_2 < Y_1) \\ &= 2P(X_2 > X_1, Y_2 > Y_1) \end{aligned}$$

Using iterated expectation we can show

$$P(X_2 > X_1, Y_2 > Y_1) = \sum_{i=1}^{I} \sum_{j=1}^{J} \pi_{ij} \left(\sum_{h=i+1}^{I} \sum_{k=j+1}^{J} \pi_{hk} \right).$$

γ statistic

Similarly, the probability of discordance is Π_d given by $2P(X_1 > X_2, Y_1 < Y_2)$. For pairs that are untied on both variables (i.e. they do not share the same income or education categories), the probability of concordance is $\Pi_c/(\Pi_c + \Pi_d)$ and the probability of discordance is $\Pi_d/(\Pi_c + \Pi_d)$. The difference in these is the *gamma* statistic

$$\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d}.$$

We have $-1 \leq \gamma \leq 1$. $\gamma = 1$ only if $\Pi_c = 1$, all pairs are perfectly concordant. Let C be the number of concordant pairs and D be the number of discordant pairs. An estimator is $\hat{\gamma} = \frac{C-D}{C+D}$. For the income data, $\hat{\gamma} = (C - D)/(C + D) = 0.4337$, a strong positive association between family income and education. Among untied pairs, the proportion in concordance is 43% greater than discordance.

Outline of Part II: inference and tests

- 2×2 tables: odds ratio, relative risk, difference in proportions, SAS examples.
- $I \times J$ tables: testing independence, SAS examples.
- $I \times J$ tables: following up rejection of $H_0 : X \perp Y$, SAS examples.

Inference for odds ratios

The sample odds ratio $\hat{\theta} = n_{11}n_{22}/n_{12}n_{21}$ can be zero, undefined, or ∞ if one or more of $\{n_{11}, n_{22}, n_{12}, n_{21}\}$ are zero.

An alternative is to add $1/2$ observation to each cell $\tilde{\theta} = (n_{11} + 0.5)(n_{22} + 0.5)/(n_{12} + 0.5)(n_{21} + 0.5)$. This also corresponds to a particular Bayesian estimate.

Both $\hat{\theta}$ and $\tilde{\theta}$ have skewed sampling distributions for small $n = n_{++}$. The sampling distribution of $\log \hat{\theta}$ is relatively symmetric and therefore more amenable to a Gaussian approximation. An approximate $(1 - \alpha) \times 100\%$ CI for $\log \theta$ is given by

$$\log \hat{\theta} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

A CI for θ is obtained by exponentiating the interval endpoints.

Inference for odds ratios: alternative CI

- When $\hat{\theta} = 0$ this doesn't work ($\log 0$ "$=$" $-\infty$).
- Can use $n_{ij} + 0.5$ in place of n_{ij} in MLE estimate and standard error yielding

$$\log \tilde{\theta} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n_{11}+0.5} + \frac{1}{n_{12}+0.5} + \frac{1}{n_{21}+0.5} + \frac{1}{n_{22}+0.5}}.$$

- The exact approach involves testing $H_0 : \theta = t$ for various values of t subject to rows or columns or both fixed and computing or simulating a p-value. Those values of t that give p-values greater than 0.05 define the 95% CI. This is related to Fisher's exact test.

Example: Aspirin and heart attacks

The following 2×2 contingency table is from a report by the Physicians' Health Study Research Group on $n = 22,071$ physicians that took either a placebo or aspirin every other day.

	Fatal attack	Nonfatal or no attack
Placebo	18	11,016
Aspirin	5	11,032

Here $\hat{\theta} = \frac{18 \times 11032}{5 \times 11016} = 3.605$ and $\log \hat{\theta} = \log 3.605 = 1.282$, and $\text{se}\{\log(\hat{\theta})\} = \sqrt{\frac{1}{18} + \frac{1}{11016} + \frac{1}{5} + \frac{1}{11032}} = 0.506$.

A 95% CI for θ is then $\exp\{1.282 \pm 1.96(0.506)\} = (e^{1.282-1.96(0.506)}, e^{1.282+1.96(0.506)}) = (1.34, 9.72)$.

Inference for difference in proportions & relative risk

Assume (1) multinomial sampling or (2) product binomial sampling. The row totals n_{i+} are fixed (e.g. prospective study or clinical trial) Let $\pi_1 = P(Y = 1|X = 1)$ and $\pi_2 = P(Y = 1|X = 2)$.

The sample proportion for each level of X is the MLE $\hat{\pi}_1 = n_{11}/n_{1+}$, $\hat{\pi}_2 = n_{21}/n_{2+}$. Using either large sample results or the CLT we have

$$\hat{\pi}_1 \dot{\sim} N\left(\pi_1, \frac{\pi_1(1-\pi_1)}{n_{1+}}\right) \perp \hat{\pi}_2 \dot{\sim} N\left(\pi_2, \frac{\pi_2(1-\pi_2)}{n_{2+}}\right).$$

Since the difference of two independent normals is also normal, we have

$$\hat{\pi}_1 - \hat{\pi}_2 \dot{\sim} N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_{1+}} + \frac{\pi_2(1-\pi_2)}{n_{2+}}\right).$$

Inference for difference in proportions

Plugging in MLEs for unknowns, we estimate the standard deviation of the difference in sample proportions by the standard error

$$se(\hat{\pi}_1 - \hat{\pi}_2) = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_{1+}} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_{2+}}}.$$

A Wald CI for the unknown difference has endpoints

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\frac{\alpha}{2}} se(\hat{\pi}_1 - \hat{\pi}_2).$$

For the aspirin and heart attack data,
$\hat{\pi}_1 = 18/(18 + 11016) = 0.00163$ and
$\hat{\pi}_2 = 5/(5 + 11032) = 00045.$

The estimated difference is $\hat{\pi}_1 - \hat{\pi}_2 = 0.00163 - 00045 = 0.0012$ and $se(\hat{\pi}_1 - \hat{\pi}_2) = 0.00043$ so a 95% CI for $\pi_1 - \pi_2$ is $0.0012 \pm 1.96(0.00043) = (0.0003, 0.0020)$.

Inference for relative risk

Like the odds ratio, the relative risk $\pi_1/\pi_2 > 0$ and the sample relative risk $r = \hat{\pi}_1/\hat{\pi}_2$ tends to have a skewed sampling distribution in small samples. Large sample normality implies

$$\log r = \log \hat{\pi}_1/\hat{\pi}_2 \dot{\sim} N(\log \pi_1/\pi_2, \sigma^2(\log r)).$$

where

$$\sigma(\log r) = \sqrt{\frac{1-\pi_1}{\pi_1 n_{1+}} + \frac{1-\pi_2}{\pi_2 n_{2+}}}.$$

Plugging in $\hat{\pi}_i$ for π_i gives the standard error and CIs are obtained as usual for $\log \pi_1/\pi_2$, then exponentiated to get the CI for π_1/π_2.

For the aspirin and heart attack data, the estimated relative risk is $\hat{\pi}_1/\hat{\pi}_2 = 0.00163/0.00045 = 3.60$ and $se\{\log(\hat{\pi}_1/\hat{\pi}_2)\} = 0.505$, so a 95% CI for π_1/π_2 is $\exp\{\log 3.60 \pm 1.96(0.505)\} = (e^{\log 3.60 - 1.96(0.505)}, e^{\log 3.60 + 1.96(0.505)}) = (1.34, 9.70)$.

Example: Seat-belts and traffic deaths

Car accident fatality records for children < 18, Florida 2008.

	Injury outcome		
Seat belt use	Fatal	Non-fatal	Total
No	54	10,325	10,379
Yes	25	51,790	51,815

- $\hat{\theta} = 54(51790)/[10325(25)] = 10.83$.
- $se(\log \hat{\theta}) = 0.242$.
- 95% CI for $\hat{\theta}$ is $(\exp\{\log(10.83) - 1.96(0.242)\}, \exp\{\log(10.83) + 1.96(0.242)\}) = (6.74, 17.42)$.
- We reject that $H_0 : \theta = 1$ (at level $\alpha = 0.05$). We reject that seatbelt use is not related to mortality.

SAS code

- `norow`, `nocol` and `nopercent` remove row, column and cell percentages from the table (not shown); these are conditional probabilities.
- `measures` gives estimates and CIs for odds ratio and relative risk.
- `riskdiff` gives estimate and CI for $\pi_1 - \pi_2$.
- `exact` plus `or` or `riskdiff` gives exact p-values for hypothesis tests of no difference and/or CIs.

```
data table;
input use$ outcome$ count @@;
datalines;
no  fatal 54 no  nonfatal 10325
yes fatal 25 yes nonfatal 51790
;
proc freq data=table order=data; weight count;
 tables use*outcome / measures riskdiff nopercent nocol;
* exact or riskdiff; * exact test for H0: pi1=pi2 takes forever...;
run;
```

SAS output: inference for $\pi_1 - \pi_2$, π_1/π_2, and θ

Statistics for Table of use by outcome

Column 1 Risk Estimates

	Risk	ASE	(Asymptotic) 95% Confidence Limits		(Exact) 95% Confidence Limits	
Row 1	0.0052	0.0007	0.0038	0.0066	0.0039	0.0068
Row 2	0.0005	0.0001	0.0003	0.0007	0.0003	0.0007
Total	0.0013	0.0001	0.0010	0.0016	0.0010	0.0016
Difference	0.0047	0.0007	0.0033	0.0061		

Difference is (Row 1 - Row 2)

Column 2 Risk Estimates

	Risk	ASE	(Asymptotic) 95% Confidence Limits		(Exact) 95% Confidence Limits	
Row 1	0.9948	0.0007	0.9934	0.9962	0.9932	0.9961
Row 2	0.9995	0.0001	0.9993	0.9997	0.9993	0.9997
Total	0.9987	0.0001	0.9984	0.9990	0.9984	0.9990
Difference	-0.0047	0.0007	-0.0061	-0.0033		

Difference is (Row 1 - Row 2)

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	10.8345	6.7405	17.4150
Cohort (Col1 Risk)	10.7834	6.7150	17.3165
Cohort (Col2 Risk)	0.9953	0.9939	0.9967

Three CIs give three equivalent tests...

Note that $(54/10379)/(25/51815) = 10.78$ and $(10325/10379)/(51790/51815) = 0.995$.

`Col1 risk` is relative risk of *dying* and `Col2 risk` is relative risk of *living*.

We can test all of $H_0 : \theta = 1$, $H_0 : \pi_1/\pi_2 = 1$, and $H_0 : \pi_1 - \pi_2 = 0$. *All* of these null hypotheses are equivalent to $H_0 : \pi_1 = \pi_2$, i.e. living is independent of wearing a seat belt.

Inference for testing independence in $I \times J$ tables

Assume one mult$(n, \boldsymbol{\pi})$ distribution for the whole table. Let $\pi_{ij} = P(X = i, Y = j)$; we must have $\pi_{++} = 1$.

If the table is 2×2, we can just look at $H_0 : \theta = 1$.

In general, independence holds if $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$, or equivalently, $\mu_{ij} = n\pi_{i+}\pi_{+j}$.

That is, independence implies a constraint; the parameters $\pi_{1+}, \ldots, \pi_{I+}$ and $\pi_{+1}, \ldots, \pi_{+J}$ define all probabilities in the $I \times J$ table under the constraint.

Pearson statistic

Pearson's statistic is

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

where $\hat{\mu}_{ij} = n(n_{i+}/n)(n_{+j}/n)$, the MLE under H_0.

There are $I - 1$ free $\{\pi_{i+}\}$ and $J - 1$ free $\{\pi_{+j}\}$. Then $IJ - 1 - [(I - 1) + (J - 1)] = (I - 1)(J - 1)$.
When H_0 is true, $X^2 \dot{\sim} \chi^2_{(I-1)(J-1)}$.

Likelihood ratio statistic

The LRT statistic boils down to

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

and is also $G^2 \dot{\sim} \chi^2_{(I-1)(J-1)}$ when H_0 is true.

- $X^2 - G^2 \xrightarrow{p} 0$.
- The approximation is better for X^2 than G^2 in smaller samples.
- The approximation can be okay when some $\hat{\mu}_{ij} = n_{i+}n_{+j}/n$ are as small as 1, but most are at least 5.
- When in doubt, use small sample methods.
- Everything holds for product multinomial sampling too (fixed marginals for one variable)!

SAS code: tests for independence, seat-belt data

- `chisq` gives X^2 and G^2 tests for independence (coming up in these slides).
- `expected` gives expected cell counts under independence.
- `exact` plus `chisq` gives exact p-values for testing independence using X^2 and G^2.

```
proc freq data=table order=data; weight count;
 tables use*outcome / chisq norow nocol expected;
 exact chisq;
run;
```

SAS output: table and asymptotic tests for independence

The FREQ Procedure

Table of use by outcome

use	outcome		
Frequency | Expected | Percent	fatal	nonfatal	Total
no	54	10325	10379
	13.184	10366	
	0.09	16.60	16.69
yes	25	51790	51815
	65.816	51749	
	0.04	83.27	83.31
Total	79	62115	62194
	0.13	99.87	100.00

Statistics for Table of use by outcome

Statistic	DF	Value	Prob
Chi-Square	1	151.8729	<.0001
Likelihood Ratio Chi-Square	1	104.0746	<.0001

SAS output: exact tests for independence

Pearson Chi-Square Test	
Chi-Square	151.8729
DF	1
Asymptotic Pr > ChiSq	<.0001
Exact Pr >= ChiSq	2.663E-24

Likelihood Ratio Chi-Square Test	
Chi-Square	104.0746
DF	1
Asymptotic Pr > ChiSq	<.0001
Exact Pr >= ChiSq	2.663E-24

These test the null H_0 that wearing a seat belt is independent of living. What do we conclude?

Example: Belief in God, a 3×6 table

Highest degree	Belief in God: Don't believe	No way to find out	Some higher power	Believe sometimes	Believe but doubts	Know God exists
Less than high school	9	8	27	8	47	236
High school or junior college	23	39	88	49	179	706
Bachelor or graduate	28	48	89	19	104	293

General Social Survey data cross-classifies opinion on whether God exists by highest education degree obtained.

SAS code, belief in God data

```
data table;
input degree$ belief$ count @@;
datalines;
1 1   9 1 2   8 1 3  27 1 4   8 1 5  47 1 6 236
2 1  23 2 2  39 2 3  88 2 4  49 2 5 179 2 6 706
3 1  28 3 2  48 3 3  89 3 4  19 3 5 104 3 6 293
;
proc format; value $dc
 '1' = 'less than high school'
 '2' = 'high school or junior college'
 '3' = 'bachelors or graduate';
value $bc
 '1' = 'dont believe'
 '2' = 'no way to find out'
 '3' = 'some higher power'
 '4' = 'believe sometimes'
 '5' = 'believe but doubts'
 '6' = 'know God exists';
run;
proc freq data=table order=data; weight count;
 format degree $dc. belief $bc.;
 tables degree*belief / chisq expected norow nocol;
run;
```

Annotated output from proc freq

degree	belief					
Frequency / Expected / Percent	dont believe	no way to find out	some higher power	believe sometimes	believe but doubts	know God exists	Total
less than high school	9 10.05 0.45	8 15.913 0.40	27 34.17 1.35	8 12.73 0.40	47 55.275 2.35	236 206.86 11.80	335 16.75
high school or junior college	23 32.52 1.15	39 51.49 1.95	88 110.57 4.40	49 41.192 2.45	179 178.86 8.95	706 669.37 35.30	1084 54.20
bachelors or graduate	28 17.43 1.40	48 27.598 2.40	89 59.262 4.45	19 22.078 0.95	104 95.865 5.20	293 358.77 14.65	581 29.05
Total	60 3.00	95 4.75	204 10.20	76 3.80	330 16.50	1235 61.75	2000 100.00

Statistics for Table of degree by belief

Statistic	DF	Value	Prob
Chi-Square	10	76.1483	<.0001
Likelihood Ratio Chi-Square	10	73.1879	<.0001

Statistic	Value	ASE
Gamma	-0.2483	0.0334

Following up chi-squared tests for independence

Rejecting $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ does not tell us about the nature of the association.

Pearson and standardized residuals

The Pearson residual is

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

where, as before, $\hat{\mu}_{ij} = n_{i+}n_{+j}/n$ is the estimate under $H_0 : X \perp Y$.

When $H_0 : X \perp Y$ is true, under multinomial sampling $e_{ij} \dot{\sim} N(0, v)$, where $v < 1$, in large samples.

Note that $\sum_{i=1}^{I} \sum_{j=1}^{J} e_{ij}^2 = X^2$.

Standardized Pearson residuals

Standardized Pearson residuals are Pearson residuals divided by their standard error under multinomial sampling.

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

where $p_{ij} = n_{ij}/n$ are MLEs under the full (non-independence) model. Values of $|r_{ij}| > 3$ happen very rarely when $H_0 : X \perp Y$ is true and $|r_{ij}| > 2$ happen only roughly 5% of the time.

Pearson residuals and their standardized version tell us which cell counts are much larger or smaller than what we would expect under $H_0 : X \perp Y$.

Residuals, belief in God data

Annotated output from `proc genmod`:

```
proc genmod order=data; class degree belief;
 model count = degree belief / dist=poi link=log residuals;
run;
```

The GENMOD Procedure

Observation	Raw Residual	Pearson Residual	Deviance Residual	Std Deviance Residual	Std Pearson Residual	Likelihood Residual
1	-1.050027	-0.33122	-0.337255	-0.375301	-0.368586	-0.374018
2	-7.912722	-1.983598	-2.196043	-2.466133	-2.227559	-2.41867
3	-7.17002	-1.226585	-1.273736	-1.473157	-1.418624	-1.459585
4	-4.730002	-1.325706	-1.423967	-1.591184	-1.481383	-1.569931
5	-8.275002	-1.113022	-1.142684	-1.370537	-1.33496	-1.35979
6	29.137492	2.0258686	1.9809013	3.5103847	3.5900719	3.5648903
7	-9.520085	-1.669418	-1.762793	-2.644739	-2.504646	-2.567827
8	-12.49071	-1.740695	-1.819318	-2.754505	-2.635467	-2.688045
9	-22.56805	-2.146245	-2.226274	-3.471424	-3.346635	-3.398513
10	7.8079994	1.2165594	1.1808771	1.7790347	1.8327913	1.8093032
11	0.1400133	0.0104692	0.0104678	0.016927	0.0169292	0.0169284
12	36.630048	1.4158081	1.403181	3.3524702	3.3826387	3.3773731
13	10.56995	2.5317662	2.3247777	2.8023308	3.0518386	2.8824417
14	20.402111	3.883624	3.51114	4.2710987	4.724204	4.4230839
15	29.737956	3.862983	3.5931704	4.5015643	4.8395885	4.6270782
16	-3.078006	-0.655073	-0.671253	-0.812499	-0.792914	-0.806333
17	8.1349809	0.8308573	0.8195034	1.0647099	1.0794611	1.0707466
18	-65.76757	-3.472204	-3.587324	-6.88618	-6.665198	-6.725887

Direction and 'significance' of standardized Pearson residuals r_{ij}

$|r_{ij}| > 3$ indicate severe departures from independence; these are in boxes below.

−	−	−	−	−	+
−	−	−	+	+	+
+	+	+	−	+	−

Which cells are over-represented relative to independence? Which are under-represented? In general, what can one say about belief in God and education? Does this correspond with the γ statistic?