Notes on Contingency Tables

Daniel J. Eck

Introduction

In these notes we will introduce and cover the basics of contingency tables, motivate applications for which contingency tables are useful, and study generalized linear models for analyzing contingency tables. Contingency tables record observations on two (or more) categorical variables and can be useful in measuring associates between these variables. Consider the joint observations of two categorical variables: X with I categories and Y with J categories. We can summarize this data in an $I \times J$ contingency table:

		Y		
		1	$\cdots$	J
X	1			
	$\vdots$			
	I			

where each cell in the above contains a count. If both X and Y are random, then let

$$\pi_{ij} = \mathbb{P}\left(X \text{ in row } i, Y \text{ in column } j\right)$$

define the joint distribution of X and Y. The marginal distribution of X and Y are defined by

$$\pi_{i+} = \mathbb{P}(X \text{ in row } i) \qquad \pi_{+j} = \mathbb{P}(Y \text{ in column } j).$$

The conditional distribution of Y given that X belongs to row i is defined by

$$\pi_{j|i} = \mathbb{P}\left(Y \text{ in column } j | X \text{ in row } i\right) = \frac{\pi_{ij}}{\pi_{i+}}.$$

Verify that the joint, marginal, and conditional distributions above are proper probability distributions. Argue that the above conditional distribution is meaningful even if counts X are not random.

Contingency tables are popular in medicine and biostatistics, consider the example that:

$$X = \text{true disease status (yes, no)}$$
$$Y = \text{test result (positive, negative)}$$

	positive	negative
yes		false negative
no	false positive	

Then, in this context, two conditional distributional quantities are defined as

$$\pi_{1|1} = \mathbb{P}\left(\text{positive}|\text{yes}\right) = \textbf{sensitivity}$$
$$\pi_{2|2} = \mathbb{P}\left(\text{negative}|\text{no}\right) = \textbf{specificity}$$

Descriptive Statistics

We will let n_{ij} be the count in row i and column j, and let the total sample size be $n = \sum_i \sum_j n_{ij}$. The margins of the contingency table are

$$n_{i+} = \sum_j n_{ij} \qquad n_{+j} = \sum_i n_{ij}$$

With this construction, a natural estimate of π_{ij} is $p_{ij} = n_{ij}/n$. We have similar estimates for the marginals $p_{i+} = \sum_j p_{ij}$, $p_{+j} = \sum_i p_{ij}$, and the conditionals $p_{j|i} = n_{ij}/n_{i+}$.

Independence

If both X and Y are random, they are independent if

$$\pi_{ij} = \pi_{i+}\pi_{+j} \qquad \text{for all } i \text{ and } j$$

and this implies that $\pi_{j|i} = \pi_{+j}$ for all i and j. **Come up with an intuitive explanation for why $\pi_{j|i} = \pi_{+j}$ for all i and j when X and Y are independent. Your answer should involve no mathematical expressions.** Even if X is not random, the condition that $\pi_{j|i} = \pi_{+j}$ for all i and j is called homogeneity. This might still be relevant in a situation where X is deliberately chosen and Y is observed as a response.

Sampling Models

We will list some possible joint distributions for counts in an $I \times J$ contingency table. These models illustrate different modeling techniques under different sampling schemes:

1. Poisson (random total)
$$Y_{ij} = \text{ count in cell } (i,j) \sim \text{Poisson}(\mu_{ij})$$
and the Y_{ij}s are independent.

2. Multinomial (fixed total n)
$$\begin{aligned} N_{ij} &= \text{ count in cell } (i,j) \\ \{N_{ij}\} &\sim \text{ multinomial}\,(n, \{\pi_{ij}\}) \end{aligned}$$

3. Independent Multinomial: Assume row totals $n_i = n_{i+}$ are fixed.
$$\left.\begin{aligned} \{N_{1j}\} &\sim \text{multinomial}(n_1, \{\pi_{j|1}\}) \\ &\vdots \\ \{N_{Ij}\} &\sim \text{multinomial}(n_I, \{\pi_{j|I}\}) \end{aligned}\right\} \text{ independent}$$
When $J = 2$, this is independent binomial sampling, for which we may just write π_i in place of $\pi_{1|i}$ and $\pi_{2|i} = 1 - \pi_i$.

4. Hypergeometric: Assume that the row and column totals are fixed.

Note that each successive model is obtained from the previous model by assuming more restrictive structure on the data. Here is a simple conceptual example:

$$\begin{aligned} X &= \text{ seat belt use (yes, no)} \\ Y &= \text{ accident result (fatal, non-fatal)} \end{aligned}$$

Possible sampling schemes:

1. use all accidents last year $\implies$ Poisson sampling
2. sample a fixed number n of accidents $\implies$ multinomial
3. sample a fixed number n_1 where seat belt was used, and a fixed number n_2 where it was not $\implies$ independent binomial.

Other important situations

1. **Case-control study**: categories of the Y variable have fixed counts. These are often retrospective where X is determined from past information.
2. **Prospective study**: follow a group of subjects over time. In a **clinical trial**, X is randomly assigned. In a **cohort study**, X is chosen by subject. One can use the independent multinomial in either case.
3. **Cross-sectional study**: sample subjects from populations, then record (X, Y) for each. In this case, use multinomial when n is fixed.

Clinical trials are **experimental**: X assigned by the investigator. Experimental designs allow for inferences about causation.

Case-control, cohort, and cross-sectional studies are **observational studies**: X is not controlled by the investigator. Observational studies do not necessarily yield causal conclusions using traditional techniques. The rapidly growing field of causal inference (which combines researchers in experimental design, classical statistics, computer science, social science, political science, epidemiology, and others) seeks to extract causal conclusions from observational data. We will only scrape the surface of causal inference in this course.

Comparing two proportions

Suppose that $I = J = 2$:

		Y		
X	N_{11}	N_{12}	n_1
	N_{21}	N_{22}	n_2

Assume the independent binomial (multinomial with two categories) model:

$$N_{11} \sim \text{binomial}(n_1, \pi_1)$$
$$N_{21} \sim \text{binomial}(n_2, \pi_2)$$

Note that π_1 and π_2 are conditional probabilities that Y is in the first column. Note that homogeneity is the condition that $\pi_1 = \pi_2$. There are three widely used measurements to compare two proportions

1. difference of proportions: $\pi_1 - \pi_2$
2. relative risk: $RR = \pi_1/\pi_2$
3. odds ratio:
$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

What are each of these quantities under homogeneity? The odds ratio is also used in the multinomial model:

$$\theta = \frac{\pi_{1|1}/\pi_{2|1}}{\pi_{1|2}/\pi_{2|2}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

It can also be used in the Poisson model:

$$\theta = \frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}.$$

Useful properties of odds ratio:

- Interchanging rows (or columns) changes θ to $1/\theta$
- Interchanging X and Y doesn't change θ
- Multiplying a row (or a column) by a factor c doesn't change θ. This is scale invariance.
- Relationship to relative risk:

$$\theta = RR \cdot \frac{1-\pi_2}{1-\pi_1}$$

Interpretation of the odds-ratio in this setting: An odds ratio is a measure of association between an exposure and an outcome. The odds ratio represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure. Odds ratios are most commonly used in case-control studies, however they can also be used in cross-sectional and cohort study designs as well See here for more details. Odds ratios are used to compare the relative odds of the occurrence of the outcome of interest (e.g. disease or disorder), given exposure to the variable of interest (e.g. health characteristic, aspect of medical history). The odds ratio can also be used to determine whether a particular exposure is a risk factor for a particular outcome, and to compare the magnitude of various risk factors for that outcome.

- $\theta = 1$ Exposure does not affect odds of outcome
- $\theta > 1$ Exposure associated with higher odds of outcome
- $\theta < 1$ Exposure associated with lower odds of outcome

Example: Myocardial Infarction (MI)

Suppose that an experimenter is interested in investigating the potential benefits of aspirin for preventing MI. The experimenter sets up a clinical trial in which individuals are randomly assigned to the placebo or aspirin regimens. The occurrence of myocardial infarction is recorded over a 5 year period.

	Yes	No
Placebo	189	10845
Aspirin	104	10933

Let π_1 be the probability of MI under placebo and let π_2 be the probability of MI under aspirin. **Estimate $\hat{\pi}_1$ and $\hat{\pi}_2$ and compute the risk difference, the risk ratio, and the odds ratio. Interpret these results.** Read Section 2.2 in Agresti [2013] for more details.

Remark: Prospective studies (like this example) allow the risk difference, the risk ratio, and the odds ratio to be estimated. Retrospective studies allow only the odds ratio to be estimated, the others cannot be estimated. This is because in a retrospective study the experimenter usually does not have the ability to estimate $\mathbb{P}(Y \text{ in column } j \mid X \text{ in row } i)$, and the risk difference and the risk ratio cannot be estimated as a result. However, the odds ratio can be written in terms of $\mathbb{P}(X \text{ in row } i \mid Y \text{ in column } j)$ (quantities that can be estimated in a retrospective study):

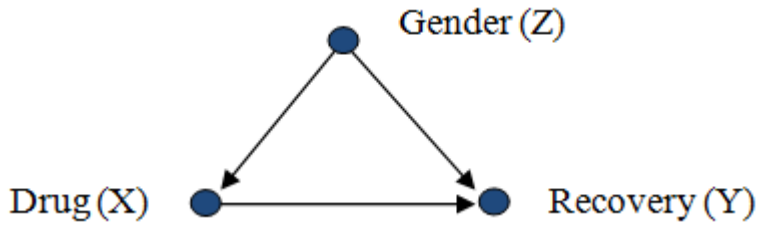
$$\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{\mathbb{P}(Y=1 \mid X=1)/\mathbb{P}(Y=2 \mid X=1)}{\mathbb{P}(Y=1 \mid X=2)/\mathbb{P}(Y=2 \mid X=2)} = \frac{\mathbb{P}(X=1 \mid Y=1)/\mathbb{P}(X=2 \mid Y=1)}{\mathbb{P}(X=1 \mid Y=2)/\mathbb{P}(X=2 \mid Y=2)}.$$

Remark: Studies such as this one were used to support the recommendation of an aspirin regimen for prevention of myocardial infraction. However, these recommendations are changing in light of new evidence. The new recommendations come in part after studies have showed an association between aspirin and major bleeding events in adults with no history of cardiovascular disease, such as myocardial infarction. What is interesting to note here is that associations between bleeding and low-dose aspirin usages are not necessarily new. So this begs the question: did studies like the one we just analyzed get it wrong? The short answer is no, the benefit-risk balance has likely shifted in the presence of new studies, the development of safer risk reduction strategies including statins and recognition of the role of blood pressure. This last part is

particularly interesting. Basic education on the role of blood pressure in causing cardiovascular disease, such as myocardial infarction, has played a part in lowering the overall risk of negative outcomes in the general population. The consequence of this is a reversal in a general medical recommendation. You can see the current aspirin recommendations from the US Preventive Services Task Force here.

Conditional Association

We will now consider a third categorical variable Z. The main question of interest: Does the X-Y relationship change across different levels of Z? This is an important question and the conceptual reasoning involved underlies causal inference for the effect that a treatment has on a response variable in the presence of measurable confounding variables.



Example: Is a drug more effective at curing a disease among females than males? Let X be a drug or a placebo, Y be a disease indicator (1 = cured, 0 = not cured), and Z be the gender. Z may be called a **stratification variable**. Stratification can be used to control for potential confounding variables that are measured.

We are interested in the distribution of (X, Y) **conditional** on Z. In observational studies, Z may be a confounding variable. Each Z category defines a **partial table** for X and Y. For example, when X, Y, Z are all binary ($2 \times 2 \times 2$ table):

$Z = 1$:

		Y	
X		n_{111}	n_{121}
		n_{211}	n_{221}

$Z = 2$:

		Y	
X		n_{112}	n_{122}
		n_{212}	n_{222}

These represent conditional associations. The marginal table sums the partial tables across the levels Z and these represent marginal associations (ignoring Z):

		Y	
X		n_{11+}	n_{12+}
		n_{21+}	n_{22+}

In general, let μ_{ijk} equal the expected count in row i, column j, and table k. The conditional odds ratios

$$\theta_{XY(k)} = \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}}$$

are estimated by

$$\hat{\theta}_{XY(k)} = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}$$

The marginal odds ratio

$$\theta_{XY} = \frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}}$$

is estimated from the marginal tables, as usual.

Here are some counterintuitive but possible situations:

- There are conditional associations ($\theta_{XY(k)} \neq 1$) but no marginal association ($\theta_{XY} = 1$).
- There are no conditional associations ($\theta_{XY(k)} = 1$) but there is a marginal association ($\theta_{XY} \neq 1$).
- **Simpson's paradox**: The conditional associations are in the opposite direction from the marginal,

$$\theta_{XY(k)} > 1 \qquad \theta_{XY} < 1$$

Conceptualize and describe an example for which Simpson's paradox can occur.

The X and Y variables are conditionally independent given $Z = k$ if $\theta_{XY(k)} = 1$. If this is true for all k, X and Y are conditionally independent given Z. This is not the same as marginal independence between X and Y ($\theta_{XY} = 1$). **Prove that conditional independence is not the same as marginal independence.**

For multinomial sampling, we can show that conditional independence occurs when

$$\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}} \qquad \text{for all } i, j, k.$$

Let Z have K categories. We say that X and Y have homogeneous association over Z if

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}$$

Conditional independence is a special case. Keep in mind that, even under homogeneous association, the conditional and marginal associations may be different.

For a detailed presentation of confounding, confounding biases, and additional biases in the context of causal inference that goes beyond the scope of this course, see this paper. One of the takeaways from this paper is that the strength of the confounder and the stratum specific sample size of the confounding variable play an important role in the practical importance of controlling for a confounding variable.

R Example: Death Penalty Data

```
deathpenalty <- read.table("deathpenalty.txt")
deathpenalty$DeathPenalty <- as.factor(deathpenalty$DeathPenalty)
deathpenalty$Defendant <- as.factor(deathpenalty$Defendant)
deathpenalty$Victim <- as.factor(deathpenalty$Victim)
deathpenalty
```

```
##   DeathPenalty Defendant Victim Freq
## 1          Yes     White  White   53
## 2           No     White  White  414
## 3          Yes     Black  White   11
## 4           No     Black  White   37
## 5          Yes     White  Black    0
## 6           No     White  Black   16
## 7          Yes     Black  Black    4
## 8           No     Black  Black  139
```

We re-level so that results match tables in Agresti.

```
deathpenalty <- transform(deathpenalty,
                          DeathPenalty = relevel(DeathPenalty, "Yes"),
                          Defendant = relevel(Defendant, "White"),
                          Victim = relevel(Victim, "White"))
```

The partial tables:

```
dp <- xtabs(Freq ~ Defendant + DeathPenalty + Victim, data=deathpenalty)
dp
```

```
## , , Victim = White
##
##           DeathPenalty
## Defendant Yes  No
##     White  53 414
##     Black  11  37
##
## , , Victim = Black
##
##           DeathPenalty
## Defendant Yes  No
##     White   0  16
##     Black   4 139
```

```
addmargins(dp)
```

```
## , , Victim = White
##
##           DeathPenalty
## Defendant Yes  No Sum
##     White  53 414 467
##     Black  11  37  48
##     Sum    64 451 515
##
## , , Victim = Black
##
##           DeathPenalty
## Defendant Yes  No Sum
##     White   0  16  16
##     Black   4 139 143
##     Sum     4 155 159
##
## , , Victim = Sum
##
##           DeathPenalty
## Defendant Yes  No Sum
##     White  53 430 483
##     Black  15 176 191
##     Sum    68 606 674
```

Another format for the partial tables:

```
dpflat <- ftable(DeathPenalty ~ Victim + Defendant, data=dp)
dpflat
```

```
##                 DeathPenalty Yes  No
## Victim Defendant
## White  White                  53 414
##        Black                  11  37
## Black  White                   0  16
##        Black                   4 139
```

Estimated proportions:

```
prop.table(dpflat)
```

```
##                 DeathPenalty         Yes          No
## Victim Defendant
## White  White                  0.078635015 0.614243323
##        Black                  0.016320475 0.054896142
## Black  White                  0.000000000 0.023738872
##        Black                  0.005934718 0.206231454
```

Estimated conditional odds ratios:

```
dp[1,1,1] * dp[2,2,1] / (dp[1,2,1] * dp[2,1,1])  # white victim
```

```
## [1] 0.4306105
```

```
dp[1,1,2] * dp[2,2,2] / (dp[1,2,2] * dp[2,1,2])  # black victim
```

```
## [1] 0
```

Marginal table:

```
mdp <- xtabs(Freq ~ Defendant + DeathPenalty, data=deathpenalty)
mdp
```

```
##          DeathPenalty
## Defendant Yes  No
##     White  53 430
##     Black  15 176
```

Estimated marginal odds ratio:

```
mdp[1,1] * mdp[2,2] / (mdp[1,2] * mdp[2,1])
```

```
## [1] 1.446202
```

Simpson's paradox: White defendants are marginally more likely to get the death penalty, but less likely after conditioning on victims' race.

Two-Way Table Inference

Consider observing a 2×2 table:

n_{11}	n_{12}
n_{21}	n_{22}

We will assume the independent binomial model

Y_1	$n_1 - Y_1$
Y_2	$n_2 - Y_2$

$$Y_i \sim \text{ indep. binomial}(n_i, \pi_i)$$

which regards row totals as fixed. Recall: This model is implied (conditionally) by the Poisson and multinomial models.

Confidence intervals for measures of association

- **Difference in proportions**: $\pi_1 - \pi_2$

  $$\hat{\pi}_1 - \hat{\pi}_2 = \frac{y_1}{n_1} - \frac{y_2}{n_2}$$

  An approximate $(1-\alpha) \times 100\%$ confidence interval:

  $$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}.$$

  This confidence interval is problematic if π_1 and π_2 are near 0 or 1.

- **Relative Risk RR**: π_1/π_2

  $$r = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{y_1/n_1}{y_2/n_2}$$

  An approximate $(1-\alpha) \times 100\%$ confidence interval for $\log(RR)$:

  $$\log(r) \pm z_{\alpha/2} \sqrt{\frac{1-\hat{\pi}_1}{y_1} + \frac{1-\hat{\pi}_2}{y_2}}.$$

  Exponentiation of the endpoints of the above gives a confidence interval for the RR.

- **Odds Ratio**: θ

  $$\hat{\theta} = \frac{n_{11} n_{22}}{n_{12} n_{21}}$$

  An approximate $(1-\alpha) \times 100\%$ confidence interval for $\log(\theta)$:

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

  Exponentiation of the endpoints of the above gives a confidence interval for the odds ratio.

Keep in mind that it would be inappropriate to interpret a CI that spans the null value (risk difference $= 0$, $RR = 1$, or $\theta = 1$) as indicating evidence for lack of association between the exposure and outcome. One should also report whether or not the range of values in the CI contains values which are of practical significance.

R Example: Seat Belt Data

An analysis of injury outcomes in auto accidents corresponding to seat belt usage for children under 18 in Florida in 2008:

```
seatbelt <- data.frame(Use=c("No","No","Yes","Yes"),
                       Injury=c("Fatal","Nonfatal","Fatal","Nonfatal"),
                       Freq=c(54,10325,25,51790))

sb.tab <- xtabs(Freq ~ Use + Injury, data=seatbelt)
sb.tab
```

```
##      Injury
## Use   Fatal Nonfatal
##   No     54    10325
##   Yes    25    51790
```

Difference of Proportions 95% CI

```
n1 <- sb.tab[1,1] + sb.tab[1,2]
n2 <- sb.tab[2,1] + sb.tab[2,2]

pihat1 <- sb.tab[1,1] / n1
pihat2 <- sb.tab[2,1] / n2

pihat1 - pihat2 + c(-1,1) * qnorm(1-0.05/2) *
  sqrt(pihat1*(1-pihat1)/n1 + pihat2*(1-pihat2)/n2)
```

```
## [1] 0.003323406 0.006117250
```

Relative Risk 95% CI

```
( r <- pihat1 / pihat2 )
```

```
## [1] 10.78335
```

```
logr.CI <- log(r) + c(-1,1) * qnorm(1-0.05/2) *
  sqrt((1-pihat1) / sb.tab[1,1] + (1-pihat2) / sb.tab[2,1])
exp(logr.CI)
```

```
## [1]  6.715008 17.316533
```

Odds Ratio 95% CI

```
( OR.est <- sb.tab[1,1] * sb.tab[2,2] / (sb.tab[1,2] * sb.tab[2,1]) )
```

```
## [1] 10.83452
```

```
logOR.CI <- log(OR.est) + c(-1,1) * qnorm(1-0.05/2) * sqrt(sum(1/sb.tab))
exp(logOR.CI)
```

```
## [1]  6.740538 17.415047
```

Note that the odds ratio is unstable even though the sample size is decently large. This is because the fatal counts are relatively rare events.

Delta Method

It is easy to get an approximate confidence interval for a mean based on a sample mean using the central limit theorem (CLT). What about a confidence interval for a transformed mean? We could just transform the confidence interval for the mean, but what if the estimator is more nearly normal on the transformed scale? Suppose a statistic T_n and a parameter θ have the following asymptotic distribution

$$\sqrt{n}\,(T_n - \theta) \xrightarrow{d} N(0, \sigma^2)$$

as $n \to \infty$. We want a CI for $g(\theta)$ for some smooth function g. A simple Taylor expansion at θ gives

$$g(T_n) \approx g(\theta) + g'(\theta)(T_n - \theta)$$

where the above approximation improves as $n \to \infty$. The above implies that

$$\sqrt{n}\,(g(T_n) - g(\theta)) \approx g'(\theta)\sqrt{n}(T_n - \theta) \xrightarrow{d} N\left(0, (g'(\theta))^2\sigma^2\right)$$

as $n \to \infty$. This approximation is only useful if $g'(\theta) \neq 0$. Thus, the asymptotic variance is $(g'(\theta))^2\sigma^2/n$. So, for n large enough, an approximate confidence interval for $g(\theta)$ is

$$g(T_n) \pm z_{\alpha/2}\frac{|g'(\theta)|\sigma}{\sqrt{n}}.$$

Example: Let $Y \sim$ binomial(n, π) and let $T_n = \hat{\pi} = Y/n$. **Use the CLT and the delta method to construct an asymptotic confidence interval for logit**(π). Note that this recipe does not work when $\hat{\pi} = 0$ or 1. Why?

We now consider the asymptotic distribution for the log odds ratio (the log of the odds ratio, a quantity that arises naturally from the logistic regression model). Standard errors for the log odds ratio result from a multiparameter version of the delta method. Suppose that $\{N_1, \ldots, N_c\}$ have a multinomial (n, π) distribution where we recall that c is the number of categories and π is the vector of cell probabilities. Recall that $\hat{\pi}_i = n_i/n$ has $\mathrm{E}(\hat{\pi}_i) = \pi_i$ and $\mathrm{Var}(\hat{\pi}_i) = \pi_i(1 - \pi_i)/n$. Also note that

$$\mathrm{cov}(\hat{\pi}_i, \hat{\pi}_j) = -\pi_i\pi_j/n.$$

Asymptotically,

$$\sqrt{n}\,(\hat{\pi} - \pi) \xrightarrow{d} N(0, \Sigma)$$

where the diagonal of Σ is the above univariate variances, and the off-diagonal of Σ contains the pairwise covariances. Now consider a differentiable mapping $g(\pi)$. Let

$$\phi_i = \frac{\partial g(\pi)}{\partial \pi_i}, \quad i = 1, \ldots, c.$$

Then, as $n \to \infty$,

$$\frac{\sqrt{n}\,[g(\hat{\pi}) - g(\pi)]}{\sigma} \xrightarrow{d} N(0, 1),$$

where $\sigma = \sum \pi_i\phi_i^2 - (\Sigma\pi_i\phi_i)^2$. Using plug-in, a $(1 - \alpha) \times 100\%$ asymptotic CI is

$$g(\hat{\pi}) \pm z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}.$$

We now apply the delta method to the log odds ratio, taking

$$g(\pi) = \log(\theta) = \log(\pi_{11}) + \log(\pi_{22}) - \log(\pi_{12}) - \log(\pi_{21}).$$

We have

$$\phi_{11} = \frac{\partial \log(\theta)}{\partial \pi_{11}} = \frac{1}{\pi_{11}} \tag{1}$$

$$\phi_{22} = \frac{1}{\pi_{22}}, \quad \phi_{12} = -\frac{1}{\pi_{12}}, \quad \phi_{21} = -\frac{1}{\pi_{21}}, \tag{2}$$

$\sum_i \sum_j \pi_{ij}\phi_{ij} = 0$ and $\sigma^2 = \sum_i \sum_j \pi_{ij}\phi_{ij}^2 = \sum_i \sum_j 1/\pi_{ij}$. Therefore, the standard error for the log odds ratio is

$$\sigma(\log \hat{\theta}) = \frac{\sigma}{\sqrt{n}} = \sqrt{\sum_i \sum_j \frac{1}{n\pi_{ij}}}$$

Testing Independence

Assume that an $I \times J$ table arises from a multinomial sample of size n. Let

$$\pi_{ij} = \mathbb{P}\left(X \text{ in row } i, Y \text{ in column } j\right).$$

(How many nonredundant parameters are there?) Suppose that we want to test that

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j} \text{ for all } i, j \qquad \text{(i.e. } X, Y \text{ are independent)}$$

(How many nonredundant parameters under H_0?) Let

$$\mu_{ij} = E(N_{ij}) = n\pi_{ij} = n\pi_{i+}\pi_{+j} \text{ under } H_0 \tag{3}$$

Under H_0 and assuming no empty rows or columns, we can show that the maximum likelihood estimates (MLEs) are

$$\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = n\left(\frac{n_{i+}}{n}\frac{n_{+j}}{n}\right) = \frac{n_{i+}n_{+j}}{n}.$$

In the general case, $\hat{\pi}_{ij} = n_{ij}/n$. These are calculated by maximizing the kernel of the multinomial likelihood (or log likelihood) given by

$$\prod_i \prod_j \pi_{ij}^{n_{ij}}, \quad \text{where all } \pi_{ij} \geq 0 \quad \text{and} \quad \sum_i \sum_j \pi_{ij} = 1.$$

Pearson's χ^2 test (score test)

The Pearson χ^2 test statistic is

$$X^2 = \sum_{\text{all cells of table}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \overset{H_0}{\sim} \chi^2_{\text{df}}$$

where df $= (I-1)(J-1)$. Note that $(I-1)(J-1) = (IJ-1) - ((I-1) + (J-1))$. The left side of the subtraction is the total number of parameters in a $I \times J$ contingency table, and the right hand side of the subtraction is the total number of parameters under H_0, constant row and column marginals (i.e. independence). Reject H_0 if

$$X^2 > \chi^2_{(I-1)(J-1)}(\alpha)$$

or use p-value.

Likelihood ratio χ^2 test

Recall that a likelihood ratio test has a test statistic of the form

$$-2\log(\Lambda) \quad \text{where} \quad \Lambda = \frac{\sup_{\theta\in\Theta_o} L(\theta)}{\sup_{\theta\in\Theta} L(\theta)}$$

where Θ_o corresponds to the parameter space of the restricted model in the null hypothesis, and Θ corresponds to the parameter space in the general model. In this multinomial setting, the ratio of likelihoods has the form

$$\Lambda = \frac{\prod_i \prod_j (n_{i+}n_{+j})^{n_{ij}}}{n^n \prod_i \prod_j n_{ij}^{n_{ij}}}.$$

After a bit of algebra we derive the likelihood ratio test statistic, denoted by G^2, as

$$G^2 = 2 \sum_{\text{all cells of table}} \text{observed} \cdot \log\left(\frac{\text{observed}}{\text{expected}}\right) = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log(n_{ij}/\hat{\mu}_{ij}) \overset{H_0}{\sim} \chi^2_{(I-1)(J-1)}.$$

Reject H_0 if

$$G^2 > \chi^2_{(I-1)(J-1)}(\alpha)$$

or use p-value (convention: $0\log(0) = 0$). Note that X^2 and G^2 are asymptotically equivalent under H_0. The χ^2 approximation tends to be better for X^2. For 2×2 tables testing of independence is equivalent to testing homogeneity in the independent binomial model:

$$H_0 : \pi_1 = \pi_2$$

We can show that $X^2 = z^2$ where

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})(1/n_1 + 1/n_2)}} \quad \text{and} \quad \hat{\pi} = \frac{y_1 + y_2}{n_1 + n_2}.$$

Remark: The X^2 and G^2 tests are not necessarily compatible with the Wald CIs. For example, rejecting H_0 is not equivalent to $\theta = 1$ being in a Wald confidence interval. There are score and profile likelihood confidence intervals that are compatible with the tests, see Sections 3.2.5 and 3.2.6 in Agresti [2013] for more details.

Loglinear models for contingency tables

We previously introduced loglinear models as GLMs using the log inverse change of parameters map (log link) with a Poisson response. These models are commonly used to analyze contingency tables. These models estimate the dependency of expected cell counts on categorical levels and are flexible enough to incorporate interactions among categorical variables.

Two-way tables

Consider an $I \times J$ contingency table with multinomial sampling (sampling model 2 above). Recall that expected frequencies are $\mu_{ij} = n\pi_{ij}$. Loglinear models use formulas with respect to expected frequencies μ_{ij} rather than cell probabilities π_{ij}, so they also apply to Poisson sampling for $N = IJ$ independent cell counts Y_{ij} that have $\mu_{ij} = E(Y_{ij})$. In either case, the observed cell counts are denoted by n_{ij}.

Under independence of the row categorical variable X and the column categorical variable Y, the μ_{ij} have the structure

$$\mu_{ij} = \mu\alpha_i\beta_j.$$

For example, under multinomial sampling we have $\mu_{ij} = n\pi_{i+}\pi_{+j}$. Thus we can write

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$$

for a row effect λ_i^X and a column effect λ_j^Y. This is the loglinear model of independence. Identifiability requires constraints such as $\sum_i \lambda_i^X = \sum_j \lambda_j^Y = 0$. The maximum likelihood fitted values are $\hat{\mu}_{ij} = n_{i+}n_{+j}/n$, the estimated expected frequencies for chi-squared tests of independence. The tests using X^2 and G^2 above are goodness-of-fit tests of this loglinear model.

Saturated model for two-way table

Statistically dependent variables satisfy a more complex loglinear model,

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY},$$

where the λ_{ij}^{XY}s are terms denoting interactions between X and Y. If we consider the constraint that $\lambda_I^X = \lambda_J^Y = 0$ then the remaining λ_i^X and λ_j^Y are, equivalently, coefficients of indicator variables for the first $(I-1)$ and $(J-1)$ variables, respectively. Thus λ_{ij}^{XY} is the coefficient of the product of indicator variables. The constraints $\lambda_I^X = \lambda_J^Y = 0$ imply that only $(I-1)(J-1)$ of these terms are nonredundant. Tests of independence analyze whether these λ_{ij}^{XY} parameters equal zero, so they have residual df $= (I-1)(J-1)$.

The number of parameters equals IJ, the number of cells, and the model is therefore saturated. The maximum likelihood fitted values are $\hat{\mu}_{ij} = n_{ij}$. Under this model, direct relationships exist between log odds ratios and λ_{ij}^{XY}. For instance, for 2×2 tables,

$$\begin{aligned}
\log(\theta) = \log\left(\frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}\right) &= \log(\mu_{11}) + \log(\mu_{22}) - \log(\mu_{12}) - \log(\mu_{21}) \\
&= (\lambda + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}) + (\lambda + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY}) \\
&\quad - (\lambda + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}) - (\lambda + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{XY}) \\
&= \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}.
\end{aligned}$$

Unsaturated models are preferable in practice since their fit has simpler interpretations and their fit smooths over the sample data. For tables with at least three variables, unsaturated models can include association terms.

Multinomial models for cell probabilities

Conditional on the sum of n of cell counts, Poisson loglinear models for μ_{ij} become multinomial models for cell probabilities

$$\pi_{ij} = \frac{\mu_{ij}}{\sum_a \sum_b \mu_{ab}}.$$

Thus,

$$\pi_{ij} = \frac{\exp\left(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}\right)}{\sum_a \sum_b \exp\left(\lambda + \lambda_a^X + \lambda_b^Y + \lambda_{ab}^{XY}\right)}.$$

The λ intercept parameter cancels in the multinomial model above. This parameter relates to the total sample size which is random in the Poisson model but fixed in the multinomial model. Therefore, the saturated multinomial model has $IJ - 1$ parameters corresponding to the constraint that $\sum_i \sum_j \pi_{ij} = 1$.

Loglinear models for independence and interaction in three-way tables

Above we demonstrated different modes of independence like conditional independence. Loglinear models for three-way tables describe these kinds of independence. A three-way $I \times J \times K$ contingency table with categorical variables X, Y, Z has several potential types of independence.

Mutual independence: The three variables are mutually independent when

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$$

for all i, j, k. For expected frequencies, mutual independence has loglinear form

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z.$$

Joint independence: Variable Y is jointly independent of X and Z when

$$\pi_{ijk} = \pi_{i+k}\pi_{+j+}$$

for all i, j, k. This is just ordinary two-way independence between Y and a single variable that collapses the X and Z levels. The loglinear model for this example of joint independence is

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}.$$

Conditional independence: Categorical variables X and Y are conditionally independent given Z, when independence holds for each partial table that fixes a level of Z. That is, if $\pi_{ij|k} = \mathbb{P}(X = i, Y = j|Z = k)$, then

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

for all i, j, k. Conditional independence of X and Y given Z has loglinear model form

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.$$

Conditional associations does not imply marginal independence. Conditional independence and marginal independence (the usual notion of independence between two variables) both hold when one of the stronger types of independence (mutual and joint independence) applies. See Section 9.2.1 in Agresti [2013] for further explanations of these notions of independence.

Example: alcohol, cigarette, and marijuana use

This data comes from a survey of 2276 nonurban highschool seniors near Dayton, Ohio. The surveyors asked whether the seniors had ever used alcohol, cigarettes, or marijuana. This survey was taken by Wright State University School of Medicine and the United Health Services in Dayton, Ohio. The data is inputted below

```
mat <- cbind(rep(c(1,0), each = 4),
             rep(c(1,1,0,0), 2),
             rep(c(1,0), 4),
  c(911,538,44,456,3,43,2,279))
colnames(mat) <- c("a", "c", "m", "y")
dat <- as.data.frame(mat)
dat
```

```
##   a c m   y
## 1 1 1 1 911
## 2 1 1 0 538
## 3 1 0 1  44
## 4 1 0 0 456
## 5 0 1 1   3
## 6 0 1 0  43
## 7 0 0 1   2
## 8 0 0 0 279
```

The table below displays fitted values (MLEs of μ_{ijk}) for several candidate loglinear models. The fit for model (ac,am,cm) is close to the observed data, which are the fitted values for the saturated model. The other candidate loglinear models provide poor fit.

```
sat <- glm(y ~ a*c*m, data = dat, family = "poisson")
full_cond <- glm(y ~ a + c + m + a*m + a*c + c*m, data = dat, family = "poisson")
am_cm <- glm(y ~ a + c + m + a*m + c*m, data = dat, family = "poisson")
ac_m <- glm(y ~ a + c + m + a*c, data = dat, family = "poisson")
main <- glm(y ~ a + c + m, data = dat, family = "poisson")

fits <- cbind(mat[, -4], round(predict(main, type = "response"), 1),
      round(predict(ac_m, type = "response"), 1),
      round(predict(am_cm, type = "response"), 1),
      round(predict(full_cond, type = "response"), 1),
      round(predict(sat, type = "response"), 1))
colnames(fits)[4:8] <- c("(a,c,m)", "(ac,m)", "(am,cm)","(ac,am,cm)","saturated")
fits
```

```
##   a c m (a,c,m) (ac,m) (am,cm) (ac,am,cm) saturated
## 1 1 1 1   540.0  611.2   909.2      910.4       911
## 2 1 1 0   740.2  837.8   438.8      538.6       538
## 3 1 0 1   282.1  210.9    45.8       44.6        44
## 4 1 0 0   386.7  289.1   555.2      455.4       456
## 5 0 1 1    90.6   19.4     4.8        3.6         3
## 6 0 1 0   124.2   26.6   142.2       42.4        43
## 7 0 0 1    47.3  118.5     0.2        1.4         2
## 8 0 0 0    64.9  162.5   179.8      279.6       279
```

A good-fitting loglinear model provides a foundation for describing and making inferences about associations among categorical predictors. Standard goodness-of-fit methods and inferential techniques apply to loglinear models. As such, X^2 and G^2 test whether a model provides good fit by comparing fitted cell frequencies to observed counts. For loglinear models, the dfs for the associated tests equal the number of total cell counts minus the number of model parameters. The table below shows goodness-of-fit for several candidate models. Most of these models fit poorly. However, the model consisting of all pairwise associations (ac, am, cm) fits well when compared to the saturated model ($p = 0.53$). It also has lower AIC than the saturated model.

```
tests <- rbind(
c(round(anova(main, sat, test = "LRT")$Dev[2], 1),
  round(anova(main, sat, test = "Rao")$Rao[2], 1), 3),
c(round(anova(ac_m, sat, test = "LRT")$Dev[2], 1),
  round(anova(ac_m, sat, test = "Rao")$Rao[2], 1), 3),
c(round(anova(am_cm, sat, test = "LRT")$Dev[2], 1),
  round(anova(am_cm, sat, test = "Rao")$Rao[2], 1), 2),
c(round(anova(full_cond, sat, test = "LRT")$Dev[2], 1),
  round(anova(full_cond, sat, test = "Rao")$Rao[2], 1), 1)
)
colnames(tests) <- c("G^2", "X^2", "df")
rownames(tests) <- c("(a,c,m)", "(ac,m)", "(am,cm)","(ac,am,cm)")
tests
```

```
##               G^2    X^2 df
## (a,c,m)    1286.0 1411.4  3
## (ac,m)      843.8  704.9  3
## (am,cm)     187.8  177.6  2
## (ac,am,cm)    0.4    0.4  1
```

```
pchisq(tests[4,1], df = tests[4,3], lower = F)
```

```
## [1] 0.5270893
```

```
AIC(sat); AIC(full_cond)
```

```
## [1] 65.04343
```

```
## [1] 63.41741
```

Tests about which conditional associations are relevant involve comparing loglinear models. For example, the test of conditional association between alcohol use and cigarette smoking compares model (ac, cm) and (ac, am, cm). The test statistic is given by the difference in G^2 values corresponding to these model fits (ie LRT)

$$G^2_{(\text{am,cm})} - G^2_{(\text{ac,am,cm})} = 187.8 - 0.4 = 187.4,$$

with df = 2 - 1 = 1. The corresponding χ^2 test suggests strong evidence for the presence of a conditional association between alcohol use and cigarette smoking

```
pchisq(187.4, 1, lower = F)
```

```
## [1] 1.174559e-42
```

This analysis indicates statistically significant associations. However, it is often the case that statistical significance will result when sample sizes are large. We now invesitage whether these associations are of practical significance. Confidence intervals are a more useful tool for this exercise. The table below reports $\hat{\lambda}^{\text{ac}}_{11} = 2.054$ with standard error SE = 0.174.For these constraints, this is the estimated conditional log odds ratio. A 95% Wald confidence interval for the true conditional AC odds ratio is

$$\exp(2.054 \pm 1.96 * 0.174) = (5.5, 11).$$

Therefore, strong positive association exists between cigarette use and alcohol use, for both users and nonusers of marijuana.

```
library(car)
```

```
## Loading required package: carData
```

```
summary(full_cond)
```

```
##
## Call:
## glm(formula = y ~ a + c + m + a * m + a * c + c * m, family = "poisson",
##     data = dat)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7         8
##  0.02044  -0.02658  -0.09256   0.02890  -0.33428   0.09452   0.49134  -0.03690
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.63342    0.05970  94.361  < 2e-16 ***
## a            0.48772    0.07577   6.437 1.22e-10 ***
## c           -1.88667    0.16270 -11.596  < 2e-16 ***
## m           -5.30904    0.47520 -11.172  < 2e-16 ***
## a:m          2.98601    0.46468   6.426 1.31e-10 ***
## a:c          2.05453    0.17406  11.803  < 2e-16 ***
## c:m          2.84789    0.16384  17.382  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 2851.46098  on 7  degrees of freedom
## Residual deviance:    0.37399  on 1  degrees of freedom
## AIC: 63.417
##
## Number of Fisher Scoring iterations: 4
```

Likelihood equations and asymptotic distributions

We now discuss likelihood equations, maximum likelihood estimation, and theoretical properties. For three-way tables, the joint Poisson probability that cell counts $\{Y_{ijk} = n_{ijk}\}$ is

$$\prod_{i,j,k} \frac{e^{-\mu_{ijk}} \mu_{ijk}^{n_{ijk}}}{n_{ijk}!},$$

with product taken over all cells of the table. The kernel of the log likelihood is

$$l(\mu) = \sum_{i,j,k} n_{ijk} \log(\mu_{ijk}) - \sum_{i,j,k} \mu_{ijk}.$$

The general loglinear model (saturated model) for a three-way table simplifies to

$$\begin{aligned} l(\mu) = n\lambda &+ \sum_i n_{i++}\lambda_i^X + \sum_j n_{+j+}\lambda_j^Y + \sum_k n_{++k}\lambda_k^Z \\ &+ \sum_{ij} n_{ij+}\lambda_{ij}^{XY} + \sum_{ik} n_{i+k}\lambda_{ik}^{XZ} + \sum_{jk} n_{+jk}\lambda_{jk}^{YZ} \\ &+ \sum_{i,j,k} n_{ijk}\lambda_{ijk}^{XYZ} - \sum_{ijk} \exp\left(\lambda + \cdots + \lambda_{ijk}^{XYZ}\right). \end{aligned}$$

Recall that the Poisson distribution is an exponential family, and coefficients of the parameters are sufficient statistics (you proved this). For this saturated model, the cell counts n_{ijk} are coefficients of λ_{ijk}^{XYZ}, so there is no reduction of the data. For simpler models, some of the parameters are zero and the above simplifies. For example, in the mutual independence model sufficient statistics are the coefficients of λ_i^X, λ_j^Y, and λ_k^Z. These are, respectively, n_{i++}, n_{+j+}, and n_{++k}.

The fitted values corresponding to a loglinear model are solutions to the likelihood equations. Let $\mathbf{n} = (n_1, \ldots, n_N)^T$ and $\mu = (\mu_1, \ldots, \mu_N)^T$ where $n = \sum_i n_i$. Loglinear models for positive Poisson means have the form

$$\log(\mu) = M\beta.$$

For example, consider the independence model for a 2×2 table, $\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$. With constraints $\lambda_2^X = \lambda_2^Y = 0$, we have

$$\begin{pmatrix} \log(\mu_{11}) \\ \log(\mu_{12}) \\ \log(\mu_{21}) \\ \log(\mu_{22}) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ \lambda_1^X \\ \lambda_1^Y \end{pmatrix}.$$

For the model $\log(\mu) = M\beta$, we have $\log(\mu_i) = \sum_j x_{ij}\beta_j$ for all i. For Poisson sampling we have

$$\begin{aligned} l(\mu) &= \sum_i n_i \log(\mu_i) - \sum_i \mu_i \\ &= \sum_i n_i \left(\sum_j x_{ij}\beta_j \right) - \sum_i \exp\left(\sum_j x_{ij}\beta_j \right). \end{aligned}$$

Here, the sufficient statistic for β_j is the coefficient $\sum_i n_i x_{ij}$. Notice that

$$\frac{\partial l(\mu)}{\partial \beta_j} = \sum_i n_i x_{ij} - \sum_i \mu_i x_{ij}, \qquad j = 1, 2, \ldots, p.$$

The likelihood equations equate these derivatives to zero. Thus,

$$M^T \mathbf{n} = M^T \hat{\mu}.$$

This is the observed equals expected property of exponential families that we saw earlier.

We now derive the covariance matrix for MLEs. The Hessian matrix has elements

$$\begin{aligned}
\frac{\partial^2 l(\mu)}{\partial \beta_j \partial \beta_k} &= -\sum_i x_{ij} \frac{\partial \mu_i}{\partial \beta_k} \\
&= -\sum_i x_{ij} \left\{ \frac{\partial}{\partial \beta_k} \left[\exp\left(\sum_h x_{ih} \beta_h \right) \right] \right\} \\
&= -\sum_i x_{ij} x_{ik} \mu_i.
\end{aligned}$$

The Fisher information matrix can therefore be written as

$$M^T \text{diag}(\mu) M.$$

Thus, for Poisson sampling, the asymptotic covariance matrix corresponding to the MLE of β is

$$\text{cov}(\hat{\beta}) = \left[M^T \text{diag}(\mu) M \right]^{-1}.$$

Similar asymptotic results hold with multinomial sampling. When Y_i, $i = 1, \ldots, N$ are indpendent Poisson random variables, the conditional distribution of Y_i given $n = \sum_i Y_i$ is multinomial with parameters $\pi_i = \mu_i / (\sum_a \mu_a)$ (**Prove this result**).

Acknowledgments

Aspects of these notes closely follow Trevor Park's slides. We also borrow materials from Agresti [2013].

References

A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, 2013.