Sections 2.1, 2.2

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 770: Categorical Data Analysis

Contingency tables & their distributions

Let X and Y be categorical variables measured on an 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 who are $X = i$ 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}	$\pi_{++} = 1$

2.1.2 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 $\boldsymbol{\pi} = [\pi_{ij}]_{I \times J}$. What is $(n_{1+}, \ldots, n_{I+})$ distributed?

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; more on this on slide 14. Another example is a clinical trial, where the number receiving treatment A and the number receiving treatment B are both fixed.

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

$$\sum_{j=1}^{J} \pi_{j|i} = 1 \text{ for } i = 1, \ldots, I.$$

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 $\pi_{1|2} + \pi_{2|2} + \pi_{3|2} = 1$.

Want to compare probabilities in each column.

2.1.3 Sensitivity and specificity

Diagnostic tests indicate the presence or absence of a disease or infection. Tests are typically imperfect, i.e. there is positive probability of incorrectly diagnosing a subject as not infected when they are in fact infected and vice-versa.

Let $D+$ or $D-$ be the true infection/disease status and $T+$ or $T-$ be the result of a diagnostic test.

$$\text{sensitivity} = P(T+|D+).$$

$$\text{specificity} = P(T-|D-).$$

2×2 table

Let $\pi_{11} = P(T+, D+)$, $\pi_{12} = P(T+, D-)$, $\pi_{21} = P(T-, D+)$, $\pi_{22} = P(T-, D-)$. Let subjects be randomly sampled from the population so that $\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22} = 1$. Then sensitivity is given by

$$Se = P(T+|D+) = P(T+, D+)/P(D+) = \pi_{11}/\pi_{+1}$$

and specificity is

$$Sp = P(T-|D-) = P(T-, D-)/P(D-) = \pi_{22}/\pi_{+2}.$$

To get MLEs for sensitivity and specificity, simply replace each π_{ij} by its MLE $\hat{\pi}_{ij} = n_{ij}/n$ where n_{ij} is the number falling into category (i, j) and $n = n_{++}$.

If $n_1 = n_{+1}$ and $n_0 = n_{+2}$ are fixed ahead of time, we have product multinomial sampling. The MLEs are *exactly the same*.

Example: Rapid strep test

Sheeler et al. (2002) describe a modest prospective trial of $n = 232$ individuals complaining of sore throat who were given the rapid strep (*streptococcal pharyngitis*) test T. The true status of each individual D was determined by throat culture.

A 2×2 contingency table looks like

	$D+$	$D-$	Total
$T+$	44	4	48
$T-$	19	165	184
Total	63	169	232

Estimating sensitivity, specificity, and prevalence

	$D+$	$D-$	Total
$T+$	44	4	48
$T-$	19	165	184
Total	63	169	232

- An estimate of Se is $\widehat{Se} = \widehat{P}(T+|D+) = \frac{44}{63} = 0.70$.
- An estimate of Sp is $\widehat{Sp} = \widehat{P}(T-|D-) = \frac{165}{169} = 0.98$.
- The estimated prevalence of strep among those complaining of sore throat $P(D+)$ is $\widehat{P}(D+) = \frac{63}{232} = 0.27$.

2.1.4 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

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

and

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

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.

2.1.5 Poisson, binomial, multinomial sampling

Let n_{ij} be the cell count in the $(i,j)^{th}$ classification.
Poisson sampling assumes $n_{ij} \overset{ind.}{\sim} \text{Poisson}(\mu_{ij})$. Then

$$p(\mathbf{n}|\boldsymbol{\mu}) = \mathcal{L}(\boldsymbol{\mu}) = \prod_{i=1}^{I}\prod_{j=1}^{J} e^{-\mu_{ij}}\mu_{ij}^{n_{ij}}/n_{ij}!.$$

The sample size $n = n_{++}$ is random.
When $n = n_{++}$ is fixed but the row n_{i+} and column n_{+j} totals are not we have *multinomial* sampling and

$$p(\mathbf{n}|\boldsymbol{\pi}) = \mathcal{L}(\boldsymbol{\pi}) = n!\prod_{i=1}^{I}\prod_{j=1}^{J} \pi_{ij}^{n_{ij}}/n_{ij}!.$$

Product multinomial sampling

Finally, sometimes row (or column) totals are fixed ahead of time (e.g. sampling n_{1+} = women and n_{2+} men and asking them if they smoke). Then we have *product multinomial* sampling. Agresti prefers using $n_i = n_{i+}$ for simplicity.

Fox a fixed $X = i$, there are J counts $(n_{i1}, n_{i2}, \ldots, n_{iJ})$ adding to n_{i+} and this vector is multinomial. Since there are I values of covariate X, we have I independent multinomial distributions, or the *product* of I mult$(n_i, \boldsymbol{\pi}_{|i})$ distributions where $\boldsymbol{\pi}_{|i} = (\pi_{1|i}, \ldots, \pi_{J|i})$.

$$p(\mathbf{n}|\boldsymbol{\pi}) = \mathcal{L}(\boldsymbol{\pi}) = \prod_{i=1}^{I} n_i! \prod_{j=1}^{J} \pi_{j|i}^{n_{ij}} / n_{ij}!.$$

Note: under product multinomial sampling, *only* conditional probabilities $\pi_{j|i}$ can be estimated. To estimate π_{ij} requires information on the π_{i+} occurring naturally.

2.1.6 Seat belt example

Mass. Hwy. Dept. to study seat belt use Y (yes, no) and fatality (fatal, not fatal) X of crashes on the Mass. Turnpike.
Could just analyze data as they arise naturally. Then $n_{ij} \overset{ind.}{\sim} \text{Pois}(\mu_{ij})$. *Poisson sampling.*

If $n = 200$ police records sampled from crashes on turnpike. Then $(n_{11}, n_{12}, n_{21}, n_{22})$ is $\text{mult}(200, \boldsymbol{\pi})$. *Multinomial sampling.*

Could sample $n_1 = 100$ fatal crash reports and $n_2 = 100$ nonfatal reports. Then $(n_{11}, n_{12}) \sim \text{mult}(100, (\pi_{1|1}, \pi_{2|1}))$ independent of $(n_{21}, n_{22}) \sim \text{mult}(100, (\pi_{1|2}, \pi_{2|2}))$. *Product multinomial sampling.*
Here, there's no information on the prevalence of fatal versus non-fatal accidents.

Read experimental design approach.

2.1.7 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 those who smoke within those that developed lung cancer (cases) and those that did not (controls). We can measure association between smoking and lung cancer, but cannot infer causation. These data were collected "after the fact." Data usually cheap and easy to get. Above: lung cancer (p. 42).

Always yield product multinomial sampling.

2.1.8 Types of studies

- Prospective studies start with a sample 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.
- Carefully read this section. Classify each study as multinomial or product multinomial.

2.2.1 –2.2.6 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 probability of $Y = 1$ when $X = 1$ versus $X = 2$ is $\pi_1 - \pi_2$.

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

Comparing two proportions, cont.

Example: Let $Y = 1$ indicate presence of a disease and $X = 1$ indicate an exposure.

When $\pi_2 = 0.001$ and $\pi_1 = 0.01$, $\pi_1 - \pi_2 = 0.009$. However, $\pi_1/\pi_2 = 10$. You are 10 times more likely to get the disease when $X = 1$ than $X = 2$. However, in either case the probability of getting the disease ≤ 0.01.

When $\pi_2 = 0.401$ and $\pi_1 = 0.41$, $\pi_1 - \pi_2 = 0.009$. However, $\pi_1/\pi_2 = 1.02$. You are 2% more likely to get the disease when $X = 1$ than $X = 2$. This doesn't seem as drastic as 1000%.

These sorts of comparisons figure into reporting results concerning public health and safety information. e.g. HRT for post-menopausal women, relative safety of SUVs versus sedans, etc.

2.2.3 & 2.2.4 Odds ratios

The *odds* of success (say $Y = 1$) versus failure ($Y = 2$) are $\Omega = \pi/(1-\pi)$ where $\pi = P(Y = 1)$. 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; see Sec. 2.2.7.

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:

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

You should 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 $\hat{\theta} = n_{11}n_{22}/[n_{12}n_{21}]$.

2.2.6 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 $P(Y = 1|X = 2)$.

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 in Section 3.1.