

2 Describing contingency tables

2.1 Probability structure for contingency tables

2.1.1 Contingency tables and their distributions

- ◆ Let X and Y denote two categorical response variables, X with I categories and Y with J categories.
- ◆ A $I \times J$ (I -by- J) “contingency table” or “cross-classification” table consists of I rows and J columns, each “cell” containing frequency count of for a sample. The “cell frequencies” are denoted $\{n_{ij}\}$. The total sample size is denoted $n = \sum_{ij} n_{ij}$.
- ◆ Distinguish three sampling schemes. (SDK) For 2×2 table case, they correspond to
 - A simple random sample from one group that yields a single multinomial distribution for the cross-classification of two binary responses (\Rightarrow independence question)
 - Simple random samples from two groups that yield two independent binomial distributions for a binary response i.e. “stratified random sampling” (\Rightarrow homogeneity question)
 - Randomized assignment of subjects to two equivalent treatments, resulting in the hypergeometric distribution
- ◆ The “joint distribution” $\{\pi_{ij}\}$ where $\pi_{ij} = P(X=i, Y=j)$
- ◆ The “marginal distributions” for the row variable $\{\pi_{i+}\}$ for $\pi_{i+} = \sum_j \pi_{ij}$ and for the column variable $\{\pi_{+j}\}$.
- ◆ Usually, X =explanatory variable and Y =response variable. In cases when X is fixed (not random), consider conditional distribution $\pi_{j|i}$. The question becomes “how the conditional distribution changes as the category of X changes”
 - For diagnostic tests for a disease,
 - “Sensitivity”= $P(Y=\text{diagnosed} + | \text{actually} +)$: we want it to be **{high,low}**
 - “Specificity”= $P(Y=\text{diagnosed} - | \text{actually} -)$: we want it to be **{high,low}**
 - For the following table, **what are the values** of sensitivity and specificity?
 - Relationship between $\pi_{j|i}$ and $\pi_{i,j}, \pi_{i+}$?
- ◆ Two categorical **response variables** are “independent” if **??**
- ◆ When Y is a response and X is an explanatory variable, such independence is translated in conditional distribution as **??** and referred to as “homogeneity”
- ◆ {joint, marginal, conditional} “Sample distributions” are defined similarly and denoted p or $\hat{\pi}$ in place of π .
- ◆ Note that $p_{ij} = n_{ij}/n$
- ◆ Poisson, binomial and multinomial sampling (Agresti)

- Poisson sampling model : treats cell counts $\{Y_{ij}\}$ as independent poisson random variable
- Multinomial sampling model: the total sample size n is fixed but not the row and column totals
- Independent multinomial sampling or product multinomial sampling: observations on Y at each setting of an explanatory variable X are independent, and row totals are considered fixed.
- Hypergeometric sampling distribution: both row and column margins are naturally fixed
- ◆ Seat belt example
- ◆ Types of studies
 - “cases” and “controls”
 - “Case-control studies” use a “retrospective” design. E.g. $P(\text{smoking behavior}|\text{lung cancer})$ rather than $P(\text{lung cancer}|\text{smoking behavior})$ c.f. Bayes theorem
 - Two types of studies using “prospective” sampling design
 - “Clinical trials” randomly allocate subjects to the groups
 - In “cohort studie”, subjects make their own choice
 - “Cross-sectional design” samples subjects and classifies them simultaneously on both variables
 - Observational vs Experimental study
 - case-control, cohort, cross-sectional studies : observational studies (more common but more potential for biases)
 - a clinical trial : an experimental study (can use the power of randomization)
- ◆ Example : smoking and lung cancer

2.2 Comparing two proportions

- ◆ If rows are groups and the columns are the binary categories of the response Y
- ◆ The “difference of proportions” of successes $= \pi_{1|1} - \pi_{1|2}$. (let category 1 to be “success”)
 - Difference of proportions $= 0 \Leftrightarrow$ homogeneity or independence
 - If both variables are responses, one can reverse the role of X and Y , which leads to a different result.
- ◆ The “relative risk” $= \pi_{1|1} / \pi_{1|2}$. Motivation?
 - Relative risk $= 1 \Leftrightarrow$ homogeneity or independence
- ◆ The “odds ratio”
 - For a probability π of success, the “odds” $\Omega = \pi / (1 - \pi)$ (the success is Ω times as likely as a failure)

- Within row i , the odds are $\Omega_i = \pi_i / (1 - \pi_i)$
- The odds ratio is defined as $\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1 / (1 - \pi_1)}{\pi_2 / (1 - \pi_2)} = \frac{\pi_{11} / \pi_{12}}{\pi_{21} / \pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$ (WHY?)
 - $\theta \geq 0$
 - $\theta = 1$ iff X and Y are independent
 - If a cell has zero probability, $\theta = 0$ or infinity.
 - If $\theta > 1$, subjects in row 1 are more likely to have success than are subjects in row 2 ($\pi_1 > \pi_2$) If $\theta < 1$, $\pi_1 < \pi_2$.
 - θ is farther from 1.0 in either direction if the association is stronger.
 - Two values represent the same association if one is inverse of the other
 - Thus, the “log odds ratio” $\log \theta$ is symmetric about 0 and has a couple of desirable properties (like?)
 - It is unnecessary to identify one variable as the response to use θ .
 - Equally valid for prospective, retrospective, or cross-sectional sampling designs.
 - The “sample odds ratio” $\hat{\theta} = n_{11}n_{22} / n_{12}n_{21}$ estimates the same parameter
 - The sample odds ratio is “invariant” to multiplication of counts within rows by a constant as well as multiplication within columns by a constant.
- For 2x2 version of aspirin data, what are the values?
- Case-control studies and the odds ratio
 - Since only $P(X|Y)$ is given, can't compute the difference of proportions nor relative risk for the outcome of interest
 - But the odds ratio can be computed!
 - Odds ratio = relative risk $(1 - \pi_2) / (1 - \pi_1)$. The two are similar when the probability π_i of the outcome of interest is close to zero for both groups!

Cross-classification of aspirin use and myocardial infraction (5 year, blind, randomized study, 1988)

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

The 2x2 version of cross-classification of aspirin use and myocardial infraction (5 year, blind, randomized study, 1988)

	Heart attack	No attack
--	--------------	-----------

Placebo	189	10,845
Aspirin	104	10,933

Estimated conditional distributions for breast cancer diagnoses

	Diagnosis of test	
Breast cancer	Positive	Negative
Yes	0.82	0.18
No	0.01	0.99

Cross-classification of smoking by lung cancer

	Lung cancer	
Smker	Cases	Controls
Yes	688	650
No	21	59