Chapter 2 Contingency Tables

Yibi Huang
Department of Statistics
University of Chicago

Outline

2.2 Difference in Proportions
2.3 Relative Risk and Odds Ratio
2.1 Probability Structure For Contingency Tables

2.2 Difference in Proportions

Two Sample Problems for Proportions

Choose an SRS of size n_{1+} from a large population having proportion π_1 of successes and an independent SRS of size n_{2+} from another population having proportion π_2 of successes.

Population	Population proportion	Sample size	Count of successes	Estimate of π_i
1	π_1	n_1	X_1	$p_1 = X_1/n_1$
2	π_2	n_2	X_2	$p_2 = X_2/n_2$

Example: Physician's Health Study (p.27)

Myocardial Infarction (MI) = heart attack. 2×2 table.

	MI	
Group	Yes	No
Placebo	189	10845
Aspirin	104	10933

Still 2×2:

	MI		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037
Total	293	21778	22071

Information About Physicians' Health Study (p. 27)

Physicians' Health Study was a 5-year randomized study published testing whether regular intake of aspirin reduces mortality from cardiovascular disease[1].

- Participants were male physicians 40-84 years old in 1982 with no prior history of heart attack, stroke, and cancer, no current liver or renal disease, no contraindication of aspirin, no current use of aspirin
- Every other day, the male physicians participating in the study took either one aspirin tablet or a placebo.
- Response: whether the participant had a heart attack (including fatal or non-fatal) during the 5 year period.

[1] *Source*: Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study. *New Engl. J. Med.*, **318**: 262-64,1988.

Wald CI for Diff. of Proportions

Wald CI for $\pi_1 - \pi_2$ is

$$p_1 - p_2 \pm z^* \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Example: Physicians' Health Study (p. 27)

	MI			
Group	Yes	No	Total	
Placebo	189	10845	11034	$\Rightarrow p_1 = 189/11034 \approx 0.0171$
Aspirin	104	10933	11037	$\Rightarrow p_2 = 104/11037 \approx 0.0094$

95% CI for $\pi_1 - \pi_2$:

$$0.0171 - 0.0094 \pm 1.96 \sqrt{\frac{0.0171 \times 0.9829}{11034} + \frac{0.009 \times 0.9906}{11037}}$$

$$= 0.0077 \pm 1.96(0.00154) = 0.0077 \pm 0.0030 = (0.0047, 0.0107)$$

Example: Physicians' Health Study

Conclusion:

- As the 95% CI does not contain 0, the incidence rate of heart attack was significantly lower in aspirin group than in the placebo group
- Can we claim that taking aspirin every other day is effective in reducing the chance of heart attack?
  Yes, because it was a randomized, double-blind, placebo-controlled experiment.

Agresti-Caffo Confidence Interval for $\pi_1 - \pi_2$

For small samples, Wald CI for $\pi_1 - \pi_2$ suffers from similar problem with achieving the nomial of confidence level as Wald CI for a single proportion.

Agresti and Caffo (2000) suggested adding *one success* and *one failure* in each of the two samples.

$$\tilde{p}_1 = \frac{X_1 + 1}{n_1 + 2} \qquad \tilde{p}_2 = \frac{X_2 + 1}{n_2 + 2}$$

Agresti-Caffo CI for $\pi_1 - \pi_2$ is given by

$$(\tilde{p}_1 - \tilde{p}_2) \pm z^* \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1 + 2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2 + 2}}$$

Note we still estimate π_1 and π_2 by $p_1 = X_1/n_1$ and $p_2 = X_2/n_2$, not by $\tilde{p}_1$ and $\tilde{p}_2$.

The actual confidence level of Agresti-Caffo CI is closer to the nominal level than Wald CI and hence is recommended.

Testing the Equality of Two Proportions

The z-statistic for testing H_0: $\pi_1 = \pi_2$ is

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad \text{where } p = \frac{X_1 + X_2}{n_1 + n_2}$$

Under H_0, z is approx. $N(0, 1)$.

Example: Physicians' Health Study (p. 27)

	MI		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037

$\Rightarrow p_1 = 189/11034 \approx 0.0171$

$\Rightarrow p_2 = 104/11037 \approx 0.0094$

For testing H_0: $\pi_1 = \pi_2$, $\quad p = \frac{189+104}{11034+11037} \approx 0.0132$

$$z = \frac{0.0171 - 0.0094}{\sqrt{0.0132(1-0.0132)\left(\frac{1}{11034} + \frac{1}{11037}\right)}} \approx \frac{0.0077}{0.00154} \approx 5.001$$

2-sided p-value $= 0.00000057$, strong evidence against H_0.

Small Sample Test for 2×2 Tables

Note the test on the previous slide works for large sample only. Use only when the numbers of successes and failures are both at least 5 in both samples (i.e., all n_{ij}'s are ≥ 5.)

		Success	Failure
Population	1	n_{11}	n_{12}
	2	n_{21}	n_{22}

A small sample test for H_0: $\pi_1 = \pi_2$ (Fisher's exact test) will be introduced in Section 2.6.

Relative Risk

Relative Risk (RR)

When π_1 and π_2 are both small, it sometimes makes more sense to look at their ratio π_1/π_2 than their difference $\pi_1 - \pi_2$.

E.g., consider the probability of disease for smokers (π_1) and for nonsmokers (π_2):

- Case 1: $\pi_1 = 0.51$ and $\pi_2 = 0.50$
- Case 2: $\pi_1 = 0.011$ and $\pi_2 = 0.001$.

In both cases $\pi_1 - \pi_2 = 0.01$.

But in Case 1, an increase of 0.01 due to smoking is small relative to the already sizable risk of disease in the nonsmoking population.

Case 2 has smokers with 11 times the chance of disease than nonsmokers.

Need to convey the relative magnitudes of these changes better than differences allow.

Relative Risk (RR)

relative risk (RR) $= \dfrac{\pi_1}{\pi_2}$, estimated by $= \dfrac{p_1}{p_2}$.

Example (Physicians Health Study)

Sample relative risk in the Physicians Health Study is

$$\frac{p_1}{p_2} = \frac{0.0171}{0.0094} = 1.82$$

Sample proportion of heart attacks was 82% higher for placebo group.

- Independence $\iff \dfrac{\pi_1}{\pi_2} = 1$

Inference of Relative Risk (RR) π_1/π_2 (1)

- Sampling distribution for sample RR (p_1/p_2) is highly skewed. The large sample normal approximation is NOT good.
- Sampling distribution of $\log(p_1/p_2)$ is closer to normal.
- It can be shown (by delta method in Stat 244) that

$$\mathrm{Var}(\log(p_1/p_2)) \approx \frac{1-\pi_1}{n_1\pi_1} + \frac{1-\pi_2}{n_2\pi_2}.$$

So the SE of $\log(p_1/p_2)$ is

$$SE = \sqrt{\widehat{\mathrm{Var}}(\log(p_1/p_2))} = \sqrt{\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}} = \sqrt{\frac{1}{X_1} - \frac{1}{n_1} + \frac{1}{X_2} - \frac{1}{n_2}}$$

Confidence Interval for Relative Risk (RR)

CI for log(RR):

$$\log(p_1/p_2) \pm z^* SE = \log(p_1/p_2) \pm z^* \sqrt{\frac{1}{X_1} - \frac{1}{n_1} + \frac{1}{X_2} - \frac{1}{n_2}} = (L, U)$$

CI for RR:

$$(e^L, e^U)$$

Example: Physicians' Health Study (p. 27)

Group	MI: Yes	MI: No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037

$\Rightarrow p_1 = 189/11034 \approx 0.0171$

$\Rightarrow p_2 = 104/11037 \approx 0.0094$

SE for log(RR) is

$$\sqrt{\frac{1}{X_1} - \frac{1}{n_1} + \frac{1}{X_2} - \frac{1}{n_2}} = \sqrt{\frac{1}{189} - \frac{1}{11034} + \frac{1}{104} - \frac{1}{11037}} \approx 0.1213$$

95% CI for log(RR) is

$$\log(p_1/p_2) \pm z^* SE = \log\left(\frac{0.0171}{0.0094}\right) \pm 1.96(0.1213) = 0.5984 \pm 0.2378$$
$$\approx (0.3606, 0.8362).$$

95% CI for RR is $(e^{0.3606}, e^{0.8362}) = (1.4342, 2.3076)$.

Interpretation. With 95% confidence, after 5 years, the risk of MI for male physicians taking placebo is between 1.43 and 2.30 times the risk for male physicians taking aspirin.

$\Rightarrow$ Risk of MI is at least 43% higher for the placebo group.

Odds Ratio

Odds

Consider a variable with binary outcome {Success, Failure}={S, F} (or {Yes, No})

	Outcome: Success	Outcome: Failure
probability	π	$1-\pi$

The odds of outcome S (instead of F) is

$$odds(S) = \frac{\mathrm{P}(S)}{\mathrm{P}(F)} = \frac{\pi}{1-\pi}.$$

- if odds $= 3$, then S is three times as likely as F;
- if odds $= 1/3$, then F is three times as likely as S.

$$\mathrm{P}(S) = \pi = \frac{odds(S)}{1 + odds(S)}$$

$$odds(S) = 3 \Longrightarrow \mathrm{P}(S) = \frac{3}{1+3} = \frac{3}{4}, \quad \mathrm{P}(F) = \frac{1}{4}$$
$$odds(S) = \frac{1}{3} \Longrightarrow \mathrm{P}(S) = \frac{1/3}{1+1/3} = \frac{1}{4}, \quad \mathrm{P}(F) = \frac{3}{4}$$

Odds Ratio

Population	Population proportion	Sample size	Count of successes	Estimate of π_i
1	π_1	n_1	X_1	$p_1 = X_1/n_1$
2	π_2	n_2	X_2	$p_2 = X_2/n_2$

$$odds(\text{Success}) = \begin{cases} \dfrac{\pi_1}{1-\pi_1} & \text{in population 1} \\ \dfrac{\pi_2}{1-\pi_2} & \text{in population 2} \end{cases}$$

Definition (Odds Ratio)

$$\textit{Odds Ratio}: \quad \theta = \frac{odds_1}{odds_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

Relative Risk v.s. Odds Ratio

$$\text{Odds Ratio} = \text{Relative Risk} \times \frac{1-\pi_2}{1-\pi_1}$$

When $\pi_1 \approx 0$ and $\pi_2 \approx 0$,

$$\text{Odds Ratio} \approx \text{Relative Risk}$$

Odds ratio is more further away from 1 than relative risk (RR)

- If $\pi_1 > \pi_2$, then Odds Ratio $>$ RR > 1.
- If $\pi_1 < \pi_2$, then Odds Ratio $<$ RR < 1.

Estimate of Odds Ratio

		Success	Failure	Total
Population	1	n_{11}	n_{12}	n_{1+}
	2	n_{21}	n_{22}	n_{2+}
	Total	n_{+1}	n_{+2}	n_{++}

$$\widehat{\theta} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{(n_{11}/n_{1+})/(n_{12}/n_{1+})}{(n_{21}/n_{2+})/(n_{22}/n_{2+})} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Odds ratio is thus called the "**cross-product ratio.**"

Properties of the Odds Ratio

- odds > 0, $\theta > 0$
- $\theta = 1$ when $\pi_1 = \pi_2$; i.e., when X, Y are independent.
- The further θ is from 1, the stronger the association.
  (For $Y =$ lung cancer, some studies have $\theta \approx 10$ for $X =$ smoking,
  $\theta \approx 2$ for $X =$ passive smoking.)
- If rows are interchanged (or if columns are interchanged),

  $$\theta \longrightarrow 1/\theta.$$

  e.g., a value of $\theta = 1/5$ indicates the same strength of association as $\theta = 5$, but in the opposite direction.

Log Odds Ratio

- Sampling distribution of $\widehat{\theta}$ is skewed to the right.
  Normal approximation for $\widehat{\theta}$ is NOT good.
- Sampling distribution of $\log \widehat{\theta}$ is closer to normal.
- $\theta = 1 \iff \log\theta = 0$, when X, Y are independent
- If rows (or columns) are interchanged,

  $$\log\theta \longrightarrow \log(1/\theta) = -\log\theta.$$

  The log odds ratio ($\log\theta$) is symmetric about 0, e.g.,

  $$\theta = 2 \iff \log\theta = 0.7$$
  $$\theta = 1/2 \iff \log\theta = -0.7$$

  The absolute value of $\log\theta$ indicates the strength of association

A Confidence Interval for the Odds Ratio

Large-sample (asymptotic) SE of $\log \widehat{\theta}$ is

$$\mathrm{SE}(\log \widehat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

CI for $\log \theta$:

$$(L, U) = \log \widehat{\theta} \pm z^* \times \mathrm{SE}(\log \widehat{\theta})$$

CI for θ:

$$(e^L, e^U).$$

Example (Physicians Health Study)

	MI	
Group	Yes	No
Placebo	189	10845
Aspirin	104	10933

$$\widehat{\theta} = \frac{189 \times 10933}{104 \times 10845} = 1.83$$

$$\log \widehat{\theta} = \log(1.83) = 0.605$$

$$\mathrm{SE}(\log \widehat{\theta}) = \sqrt{\frac{1}{189} + \frac{1}{10845} + \frac{1}{104} + \frac{1}{10933}} = 0.123$$

95% CI for $\log \theta : 0.605 \pm 1.96(0.123) = (0.365, 0.846)$

95% CI for $\theta : (e^{0.365}, e^{0.846}) = (1.44, 2.33)$

Remarks

- Apparently $\theta > 1$.
- $\widehat{\theta}$ not midpoint of CI because of skewness
- Better estimate if we use $\{n_{ij} + 0.5\}$. Especially if any $n_{ij} = 0$.

2.1 Probability Structure for Contingency Tables

Two-Way Contingency Tables

	Y categories				
X categories	$Y = 1$	$Y = 2$	$\cdots$	$Y = J$	X margin
$X = 1$	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+}
Y margin	n_{+1}	n_{+2}	$\cdots$	n_{+J}	$n = n_{++}$

n_{ij} = count of obs. such that $X = i$ and $Y = j$.

- The subscript $+$ denotes the sum over the index it replaces. E.g., when $I = J = 2$,

$$n_{i+} = n_{i1} + n_{i2}, \qquad n_{+j} = n_{1j} + n_{2j},$$
$$n_{++} = n_{+1} + n_{+2} = n_{11} + n_{12} + n_{21} + n_{22}$$

- Note $n_{i+} = \#$ of obs. such that $X = i$, and hence $(n_{1+}, n_{2+}, \ldots, n_{I+})$ are called the *marginal counts of X*.

Joint Distributions of Random Variables (Review)

Suppose units in a population of interest (e.g., all traffic accidents) can be classified on X (e.g., driver wearing seat belt or not) and Y (result of crash in the accident).

Let $\pi_{ij} = \mathrm{P}(X = i, Y = j)$. The probabilities $\{\pi_{ij}\}$ form the *joint distribution* of X and Y.

Example. (Hypothetical)

seat-belt use (X)	result of crash (Y) $Y=1$ (fatal)	$Y=2$ (nonfatal)	$Y=3$ (no injury)
$X=1$ (yes)	$\pi_{11}=0.01$	$\pi_{12}=0.50$	$\pi_{13}=0.20$
$X=2$ (no)	$\pi_{21}=0.03$	$\pi_{22}=0.25$	$\pi_{23}=0.01$

e.g., $\pi_{13} = P(X = 1, Y = 3) = 0.20$ means that in 20% of the traffic accidents, the driver wears seat-belt and are not injured in the accident.

Marginal Distributions of Random Variables (Review)

Example. (Hypothetical)

Seat-Belt Use (X)	result of crash (Y) $Y=1$ fatal	$Y=2$ nonfatal	$Y=3$ no injury	X margin
$X=1$ (Yes)	$\pi_{11}=0.01$	$\pi_{12}=0.50$	$\pi_{13}=0.20$	$\pi_{1+}=0.71$
$X=2$ (No)	$\pi_{21}=0.03$	$\pi_{22}=0.25$	$\pi_{23}=0.01$	$\pi_{2+}=0.29$
Y margin	$\pi_{+1}=0.04$	$\pi_{+2}=0.75$	$\pi_{+3}=0.21$	$\pi_{++}=1$

- In what percentages of traffic accidents the driver wears a seat belt? $P(X = 1) = \pi_{1+} = \pi_{11} + \pi_{12} + \pi_{13} = 0.71$
- The row sums $\{\pi_{i+}\}$ form the *marginal distribution of X* since $\mathrm{P}(X = i) = \sum_j \mathrm{P}(X = i, Y = j) = \sum_j \pi_{ij} = \pi_{i+}$.
- Likewise, the column sums $\{\pi_{+j}\}$ form the *marginal distribution of Y*.

Conditional Distributions

A conditional distribution of Y given X refers to the probability distribution of Y when we restrict attention to a fixed level of X.

$$\mathrm{P}(Y = j \mid X = i) = \frac{\mathrm{P}(X = i, Y = j)}{\mathrm{P}(X = i)} = \frac{\pi_{ij}}{\pi_{i+}}$$

Example. (Hypothetical)

seat-belt use (X)	result of crash (Y) $Y=1$ fatal	$Y=2$ nonfatal	$Y=3$ no injury	X margin
$X=1$ (yes)	$\pi_{11}=0.01$	$\pi_{12}=0.50$	$\pi_{13}=0.20$	$\pi_{1+}=0.71$
$X=2$ (no)	$\pi_{21}=0.03$	$\pi_{22}=0.25$	$\pi_{23}=0.01$	$\pi_{2+}=0.29$
Y margin	$\pi_{+1}=0.04$	$\pi_{+2}=0.75$	$\pi_{+3}=0.21$	$\pi_{++}=1$

- $P(Y = 1|X = 1) = \frac{0.01}{0.71} = 0.014 \Rightarrow$ Among traffic accidents that the driver had worn a seat belt, only 1.4% of the drivers died.
- $P(Y = 1|X = 2) = \frac{0.03}{0.29} = 0.103$ $\Rightarrow$ Among those that the driver hadn't, 10.3% of them died.

Conditional distributions of Y given X:

seat-belt use (X)	result of crash (Y) $Y=1$ fatal	$Y=2$ nonfatal	$Y=3$ no injury	total
$X=1$ (yes)	$\frac{0.01}{0.71}=0.014$	$\frac{0.50}{0.71}=0.704$	$\frac{0.20}{0.71}=0.282$	1
$X=2$ (no)	$\frac{0.03}{0.29}=0.103$	$\frac{0.25}{0.29}=0.862$	$\frac{0.01}{0.29}=0.034$	1

Conditional distributions of X given Y: $P(X = i \mid Y = j) = \pi_{ij}/\pi{+j}$

seat-belt use (X)	result of crash (Y) $Y=1$ fatal	$Y=2$ nonfatal	$Y=3$ no injury
$X=1$ (yes)	$\frac{0.01}{0.04}=0.25$	$\frac{0.50}{0.75}=0.667$	$\frac{0.20}{0.21}=0.282$
$X=2$ (no)	$\frac{0.03}{0.04}=0.75$	$\frac{0.25}{0.75}=0.333$	$\frac{0.01}{0.21}=0.034$
total	1	1	1

Interpret

$P(X = 2|Y = 1) = P(X = \text{no seat-belt} \mid Y = \text{fatal}) = 0.75$.

Among all traffic accidents that the driver died, 75% of them didn't wear the seat-belt.

Independence

X and Y are said to be *independent*

- if the conditional distribution of Y given X doesn't change with the level of X,
- or if the conditional distribution of X given Y doesn't change with the level of Y

The two conditions are equivalent

Example. If the conditional distributions of $Y|X$ are like

seat-belt use (X)	result of crash (Y) $Y = 1$ fatal	$Y = 2$ nonfatal	$Y = 3$ no injury
$X = 1$ (yes)	0.04	0.75	0.21
$X = 2$ (no)	0.04	0.75	0.21

or if the conditional distributions of $X|Y$ are like

seat-belt use (X)	result of crash (Y) $Y = 1$ fatal	$Y = 2$ nonfatal	$Y = 3$ no injury
$X = 1$ (yes)	0.71	0.71	0.71
$X = 2$ (no)	0.29	0.29	0.29

then seat-belt use and the severity of traffic accidents are independent.

Type of Studies

Types of Studies

Many types of studies result in data in the form of a contingence table.

The analysis and the conclusion can be drawn depend on how the study is done.

Example (Prenatal Vitamin and Autism) Researchers wanted to study whether mothers used prenatal vitamins during the three months before pregnancy (periconceptional period) affects whether the children had autism.

Mother	Child: autism	no autism	Total
took vitamin	n_{11}	n_{12}	n_{1+}
no vitamin	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n_{++}

Example (Prenatal Vitamin and Autism)

Study 1: randomly sample 400 children age aged 24 - 60 month and classify each according to they have autism and whether their mother took prenatal vitamins during the periconceptional period

- n_{++} is fixed at 400
- If sampled properly, the makeup of the sample will be close to the makeup of the population, all joint, marginal, and conditional probabilities are estimable
  - joint: $\widehat{\pi}_{ij} := p_{ij} = n_{ij}/n_{++}$,
  - marginal: $\widehat{\pi}_{i+} := p_{i+} = n_{i+}/n_{++}$, $\widehat{\pi}_{+j} := p_{+j} = n_{+j}/n_{++}$,
  - conditional: $\widehat{\mathrm{P}}(Y = j|X = i) = n_{ij}/n_{i+}$, $\widehat{\mathrm{P}}(X = i|Y = j) = n_{ij}/n_{+j}$
- Drawback: The prevalence of the disease are usually low (e.g., 1% to 2% for autism), the number of diseased subjects (n_{11} and n_{21}) are usually very small. May not be powerful enough to detect the effect of vitamin (or the risk factor).

Example (Prenatal Vitamin and Autism – Cont'd)

Study 2A (Cohort Study): randomly sample 200 mothers who had taken prenatal vitamins during the periconceptional period and 200 mothers who didn't, and see their children have autism at age 5.

Study 2B (Randomized experiment): randomly split 400 women to two groups. Given women in the treatment group prenatal vitamins until they get pregnant and give placebo to those in the control group until they get pregnant, and see if their children have autism at age 5.

In both Study 2A and 2B

- n_{1+}, n_{2+} are fixed at 200
- Only conditional probabilities $\mathrm{P}(X = i|Y = j) = \mathrm{P}(\text{autism or not}|\text{vitamin or not})$ are estimable.
- Drawback: number of cases n_{11} and n_{21} will be very small if the disease is rare. unless the sample sizes n_{1+}, n_{2+} are very big ($>$ 1000 or even $>$ 10000)

Example (Prenatal Vitamin and Autism – Cont'd)

Study 3 (Retrospective Study): randomly sample 200 children age 3-5 with autism and 200 children age 3-5 with typical development, and see if their mother took prenatal vitamins during the periconceptional period.

In Study 3

- n_{+1}, n_{+2} are fixed at 200
- Only conditional probabilities $\mathrm{P}(X = i|Y = j) = \mathrm{P}(\text{vitamin or not}|\text{autism or not})$ are estimable.
- Advantage: number of disease cases n_{11} and n_{21} can be large without making the overall sample size too big.
- Drawback: Only $\mathrm{P}(\text{vitamin or not}|\text{autism or not})$ are estimable, but we are more interested in $\mathrm{P}(\text{autism or not}|\text{vitamin or not})$.

Comparing Proportions in 2×2 Tables

In many studies, only the conditional probabilities $\mathrm{P}(Y = j|X = i)$ are of interest, e.g.,

		heart attack (Y)	
	$\mathrm{P}(Y\|X)$	Yes	No
smoker (X)	Yes	π_1	$1 - \pi_1$
	No	π_2	$1 - \pi_2$

The problem reduces to the comparison of

$$\pi_1 = \mathrm{P}(\text{heart attack} \,|\, \text{smoker}) \text{ and}$$
$$\pi_2 = \mathrm{P}(\text{heart attack} \,|\, \text{nonsmoker}).$$

To estimate π_1 and π_2, the study must be *prospective* —— sampling from the young population and then 10 or 20 years later measures the rates of heart attack for the smokers and nonsmokers

- *randomized experiment*: subjects are randomly assigned to smoke or not to smoke
- *cohort studies*: subjects make their own choice about whether to smoke

A Case Control Study Example (p.32)

- **cases**: 262 young and middle-aged women (age < 69) admitted to 30 coronary care units in northern Italy with acute heart attack during a 5-year period
- **controls**: each of the 262 cases above was matched with two control patients admitted to the same hospitals with other acute disorders[2].

	Heart Attack (Y)	
Ever Smoker (X)	Cases	Controls
Yes	172	173
No	90	346
Total	262	519

- This is a *retrospective* ("look into the past") study

[2] Source: A. Gramenzi et al., J. Epidemiol. Community Health, 43:214-217, 1989.

In the case-control study, the marginal totals for "MI or not" are fixed, we can estimate

$$\tau_1 = \mathrm{P}(\text{smoker} \mid \text{heart attack}) \text{ and}$$
$$\tau_2 = \mathrm{P}(\text{smoker} \mid \text{no heart attack})$$

		heart attack (Y)	
	$\mathrm{P}(X\|Y)$	Yes	No
smoker (X)	Yes	τ_1	τ_2
	No	$1-\tau_1$	$1-\tau_2$

but

$$\pi_1 = \mathrm{P}(\text{heart attack} \mid \text{smoker}) \text{ and}$$
$$\pi_2 = \mathrm{P}(\text{heart attack} \mid \text{nonsmoker}).$$

are not estimable from such a study.

- (π_1, π_2) cannot be computed from (τ_1, τ_2)
- If we just want to know if heart attack is independent of smoking, testing $\pi_1 = \pi_2$ is equivalent to testing $\tau_1 = \tau_2$.

Most Important Property of the Odds Ratio

		Y (e.g., disease)		
		Yes	No	X margin
X	1 (smoker)	π_{11}	π_{12}	π_{1+}
	2 (nonsmoker)	π_{21}	π_{22}	π_{2+}
	Y margin	π_{+1}	π_{+2}	π_{++}

Prospective study:

		Y	
	$\mathrm{P}(Y\|X)$	Yes	No
X	1	π_1	$1-\pi_1$
	2	π_2	$1-\pi_2$

Retrospective study:

		Y	
	$\mathrm{P}(X\|Y)$	Yes	No
X	1	τ_1	τ_2
	2	$1-\tau_1$	$1-\tau_2$

$$\boxed{\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{\tau_1/(1-\tau_1)}{\tau_2/(1-\tau_2)}}$$

Odds ratio treats the rows and columns symmetrically, i.e., it does not distinguish X and Y.

Case-Control Study About Smoking & Heart Attack Revisit

	Heart Attack (Y)	
Smoker (X)	Cases	Controls
Yes	172	173
No	90	346
Total	262	519

Recall

$$\pi_1 = \mathrm{P}(\text{heart attack} \mid \text{smoker}),$$
$$\pi_2 = \mathrm{P}(\text{heart attack} \mid \text{nonsmoker}),$$
$$\tau_1 = \mathrm{P}(\text{smoker} \mid \text{heart attack}),$$
$$\tau_2 = \mathrm{P}(\text{smoker} \mid \text{no heart attack}),$$

Want π_1, π_2, but only got $\widehat{\tau}_1 = \frac{172}{262}$, $\widehat{\tau}_2 = \frac{173}{519}$. Neither $\pi_1 - \pi_2$ nor π_1/π_2 is estimable.

However, the odds ratio $\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$ is estimable from $\widehat{\tau}_1$ and $\widehat{\tau}_2$ since

$$\widehat{\theta} = \frac{\widehat{\pi}_1/(1-\widehat{\pi}_1)}{\widehat{\pi}_2/(1-\widehat{\pi}_2)} = \frac{\widehat{\tau}_1/(1-\widehat{\tau}_1)}{\widehat{\tau}_2/(1-\widehat{\tau}_2)} = \frac{172 \times 346}{173 \times 90} \approx 3.82$$

Conclusion: Odds of heart attack for smokers estimated to be about 3.8 times the odds for non-smokers.

If $\pi_1, \pi_2 \approx 0$ (heart attack was rare), then $\theta \approx$ relative risk, can conclude that risk of heart attack is ≈ 3.8 times as high for smokers as for non-smokers.