2 Contingency Tables

In this section we will examine association among two or three categorical variables.

Example: College students were classified according both to frequency of marijuana use and parental use of alcohol and psychoactive drugs.

		Level of Marijuana Use		
		Never	Occasional	Regular
Parental Use	Neither	141	44	40
of Alcohol	One	68	44	51
and Drugs	Both	17	11	19

We have two discrete variables, X and Y:

X = parental use of alcohol and drugs

Y = level of marijuana use

2.1 Probability Structure for Contingency Tables

We have two categorical variables, X and Y, with I and J categories, respectively. We will let X define the rows of the table and Y define the columns of the table.

2.1.1 Parameters:

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

$\{\pi_{ij}\}$ gives the joint distribution of (X, Y).

$\{\pi_{i+}\}$ gives the marginal distribution of X, and $\{\pi_{+j}\}$ gives the marginal distribution of Y.

We obtain a table of cell probabilities:

$X\backslash Y$	1	2	$\cdots$	J	Sum
1	π_{11}	π_{12}	$\cdots$	π_{1J}	π_{1+}
2	π_{21}	π_{22}	$\cdots$	π_{2J}	π_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	π_{I1}	π_{I2}	$\cdots$	π_{ij}	π_{I+}
Sum	π_{+1}	π_{+2}	$\cdots$	π_{+J}	1

2.1.2 Data:

The sample data are the cell counts: n_{ij} = the number of responses where $X = i$ and $Y = j$.

We define row totals, column totals, and the grand total:

$$n_{i+} = \sum_{j=1}^{J} n_{ij}, \qquad n_{+j} = \sum_{i=1}^{I} n_{ij}, \qquad n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$$

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

The cell probabilities are estimated by the cell proportions:

$$p_{ij} = \frac{n_{ij}}{n}$$

2.1.3 Conditional Probabilities

Often we have a table where we view Y as a response variable and X as a predictor. In such a case, we are interested in the conditional distribution of Y given X. The conditional probability of Y given X is defined by

$$P(Y=j|X=i)=\frac{P[X=i,Y=j]}{P[X=i]}=\frac{\pi_{ij}}{\pi_{i+}}$$

2.1.4 Sensitivity and Specificity in Diagnostic Testing

Diagnostic tests are used to help detect certain medical conditions. A diagnostic test for a condition is said to be *positive* if it states that the condition is present and *negative* if it states that the condition is not present. We relate this to the actual presence of the condition using a two-way table. Let $X=1$ if the condition is present and $=2$ if the condition is not present. Let $Y=1$ if the test is positive and $=0$ if the test is negative.

The sensitivity of the test is the conditional probability that the test is positive given that the condition is present. The specificity of the test is the conditional probability that the test is negative given that the condition is not present. Thus,

$$\text{sensitivity} = P(Y=1|X=1) \quad \text{and} \quad \text{specificity} = P(Y=2|X=2).$$

In using mammograms to detect breast cancer, the estimated conditional probabilities of the test results are given in the following table:

	Diagnosis of Test		
Breast Cancer	Positive	Negative	Total
Yes	0.86	0.14	1.00
No	0.12	0.88	1.00

2.1.5 Independence

We say that X and Y are statistically independent if

$$\pi_{ij} = \pi_{i+} \times \pi_{+j}, \quad i = 1, \ldots, I,\ j = 1, \ldots, J$$

An equivalent condition for X and Y to be independent is

$$P(Y = j|X = i) = \frac{\pi_{ij}}{\pi_{i+}} = \frac{\pi_{i+} \times \pi_{+j}}{\pi_{i+}} = \pi_{+j}, \text{ for all } i, j.$$

Goal: Use $\{n_{ij}\}$ to make inferences about $\{\pi_{ij}\}$, $\{\pi_{i+}\}$, or $\{\pi_{+j}\}$. This will tell us how X and Y are associated.

2.1.6 Poisson, Binomial and Multinomial Sampling

Consider a 2×2 table of gender by response to a prescription under investigation:

	Responder		Nonresponder		Total	
Male	π_{11}	n_{11}	π_{12}	n_{12}	π_{1+}	n_{1+}
Female	π_{21}	n_{21}	π_{22}	n_{22}	π_{2+}	n_{2+}
Total	π_{+1}	n_{+1}	π_{+2}	n_{+2}	π_{++}	n

Parameters: $\{\pi_{ij}, i = 1, 2,\ j = 1, 2\}$

Data: $\{n_{ij}, i = 1, 2,\ j = 1, 2\}$

Notation: $\mu_{ij} = E(n_{ij})$

- Experiment 1: Observe $\{n_{ij}\}$ from all patients coming to the pharmacy for this prescription in the next 6 months.
  - n is random
  - each cell is an independent Poisson random variable: $n_{ij} \sim$ Poisson(μ_{ij})

- Experiment 2: Observe $\{n_{ij}\}$ from next 200 patients coming to the pharmacy for this prescription.
  - $n = 200$ is fixed
  - The four cell counts form a multinomial distribution with four categories:

$$\begin{pmatrix} n_{11} \\ n_{12} \\ n_{21} \\ n_{22} \end{pmatrix} \sim \text{Multinomial} \left(n, \begin{pmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{21} \\ \pi_{22} \end{pmatrix} \right)$$

- Experiment 3: Observe $\{n_{ij}\}$ from next 100 male and next 100 female patients coming to the pharmacy for this prescription.
  - n_{1+} and n_{2+} are fixed.
  - This leads to two independent binomial samples:

$$n_{11} \sim \text{Binomial}(n_{1+}, \frac{\pi_{11}}{\pi_{1+}}), \qquad n_{21} \sim \text{Binomial}(n_{2+}, \frac{\pi_{21}}{\pi_{2+}})$$

2.2 Comparing Proportions in 2×2 Tables

We will study methods of analyzing 2×2 tables to introduce methodology that can be extended to the analysis of $I \times J$ tables.

2.2.1 Difference of Two Proportions

In some two-way tables, we can consider the rows as the two groups and the columns as the binary response categories. For row 1 we let π_1 be the probability of success and for row 2 we let π_2 be the probability of success. Such data often result from independent binomial experiments.

Experiment 1: Observe n_{11} successes in $n_1 = n_{1+}$ trials with probability of success π_1

Experiment 2: Observe n_{21} successes in $n_2 = n_{2+}$ trials with probability of success π_2

We are interested in inference concerning the difference in the probabilities of success, $\pi_1 - \pi_2$. We will use the statistic:

$$\begin{aligned} p_1 - p_2 &= \tfrac{n_{11}}{n_1} - \tfrac{n_{21}}{n_2} \\ E[p_1 - p_2] &= \pi_1 - \pi_2 \\ \text{Var}[p_1 - p_2] &= \tfrac{\pi_1(1-\pi_1)}{n_1} + \tfrac{\pi_2(1-\pi_2)}{n_2} \end{aligned}$$

Wald Interval: The Wald $100(1-\alpha)\%$ confidence interval for $\pi_1 - \pi_2$ is

$$p_1 - p_2 \pm Z_{\alpha/2}\, SE,$$

where the estimated standard error is

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Newcombe Interval: An alternative confidence interval studied by Newcombe (1998) has improved performance relative to the Wald interval for the difference in proportions. It is based on the $100(1-\alpha)\%$ Wilson (score) intervals for π_1 and π_2: $[l_1, u_1]$ and $[l_2, u_2]$.

The $100(1-\alpha)\%$ Newcombe interval for $\pi_1 - \pi_2$ is given by $[L_{Newc}, U_{Newc}]$ where

$$\begin{aligned} L_{Newc} &= \hat{\pi}_1 - \hat{\pi}_2 - Z_{\alpha/2}\sqrt{\frac{l_1(1-l_1)}{n_1} + \frac{u_2(1-u_2)}{n_2}} \\ U_{Newc} &= \hat{\pi}_1 - \hat{\pi}_2 + Z_{\alpha/2}\sqrt{\frac{u_1(1-u_1)}{n_1} + \frac{l_2(1-l_2)}{n_2}} \end{aligned}$$

Agresti-Caffo Interval: Another alternative confidence interval for the difference in proportions proposed by Agresti and Caffo (2000) has improved performance relative to the Wald interval for the difference in proportions. It is similar to the Agresti-Coull interval for a single proportion in that there are added successes and failures in the estimate. We let

$$\tilde{\pi}_1 = \frac{n_{11}+1}{n_1+2} \quad \text{and} \quad \tilde{\pi}_2 = \frac{n_{21}+1}{n_2+2}.$$

The Agresti-Caffo $100(1-\alpha)\%$ confidence interval for $\pi_1 - \pi_2$ is

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

Example: Investigators examined a sample of 178 children who appeared to be in remission from leukemia using the standard criterion after undergoing chemotherapy. A new test (PCR) detected traces of cancer in 75 of these children. During 3 years of followup, 30 of these children suffered a relapse. Of the 103 children who did not show traces of cancer, 8 suffered a relapse.

Traces of Cancer:	$n_1 = 75$	$n_{11} = 30$	$p_1 = \frac{30}{75} = 0.40$
Cancer Free:	$n_2 = 103$	$n_{21} = 8$	$p_2 = \frac{8}{103} = 0.078$

$$SE = \sqrt{\frac{(0.4)(0.6)}{75} + \frac{(0.078)(0.922)}{105}} = 0.0624$$

Traces of Cancer: $n_1 = 75 \quad n_{11} = 30 \qquad p_1 = \frac{30}{75} = 0.40$

Cancer Free: $n_2 = 103 \qquad n_{21} = 8 \quad p_2 = \frac{8}{103} = 0.078$

The 95% Wald interval is given by

$$0.40 - 0.078 \pm 1.96\sqrt{\frac{(0.4)(0.6)}{75} + \frac{(0.078)(0.922)}{103}}, \quad \text{or} \quad (0.200, 0.445).$$

The 95% score intervals for π_1 and π_2 are $(0.297, 0.513)$ and $(0.040, 0.146)$, respectively.

The 95% Newcombe interval is given by

$$\begin{aligned} L_{Newc} &= 0.400 - 0.078 - 1.96\sqrt{\frac{(0.297)(0.703)}{75} + \frac{(0.146)(0.854)}{103}} = 0.199 \\ U_{Newc} &= 0.400 - 0.078 + 1.96\sqrt{\frac{(0.513)(0.487)}{75} + \frac{(0.040)(0.960)}{103}} = 0.442 \end{aligned}$$

To form the Agresti-Caffo interval, we compute

$$\tilde{\pi}_1 = \frac{30+1}{75+2} = 0.403 \quad \text{and} \quad \tilde{\pi}_2 = \frac{8+1}{103+2} = 0.086.$$

The 95% Agrest-Caffo interval is given by

$$0.403 - 0.086 \pm 1.96\sqrt{\frac{(0.403)(0.597)}{77} + \frac{(0.086)(0.914)}{105}}, \quad \text{or} \quad (0.195, 0.439).$$

Example: In 1954, 401,974 children were participants in a large clinical trial for polio vaccine. Of these 201,229 were given the vaccine, and 200,745 were given a placebo. 110 of the children who received a placebo got polio and 33 of those given the vaccine got polio. Was the vaccine effective?

Vaccine:	$n_1 = 201229$	$n_{11} = 33$	$p_1 = 0.000164$
Placebo:	$n_2 = 200745$	$n_{21} = 110$	$p_2 = 0.000548$

$$SE = \sqrt{\frac{(0.000164)(.999836)}{201229} + \frac{(0.000548)(0.999452)}{200745}} = 0.0000595$$

$$0.000164 - 0.000548 \pm (1.96)(0.0000595)$$

$$-0.000384 \pm 0.000117 \qquad \text{or} \quad (-0.0005, -0.0003)$$

Remark: We could test the hypothesis $H_0 : \pi_1 - \pi_2 = 0$ by determining whether 0 is in the confidence interval for $\pi_1 - \pi_2$. Alternatively, we could use the test statistic:

$$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}} \quad \text{where} \quad p = \frac{n_{11} + n_{21}}{n_1 + n_2} = \frac{n_{+1}}{n_{++}}$$

2.2.2 Relative Risk

A difference between two proportions of a given size is more important when the proportions are near 0 or 1 than in the middle of the range.

Example:

$$\left[\begin{array}{l} \pi_1 = 0.460 \\ \pi_2 = 0.452 \end{array} \right. \quad \text{and} \quad \left[\begin{array}{l} \pi_1 = 0.010 \\ \pi_2 = 0.002 \end{array} \right.$$

In both cases the difference is 0.008, but the second difference seems more noteworthy.

The **relative risk** is the ratio of the success probabilities:

$$\frac{\pi_1}{\pi_2}$$

Example: $\frac{0.460}{0.452} = 1.018$ and $\frac{0.010}{0.002} = 5.$

Example: The sample relative risk in the leukemia example is

$$\frac{p_1}{p_2} = \frac{0.400}{0.078} = 5.150$$

Using PROC FREQ in SAS, a 95% confidence interval for the relative risk is $(2.505, 10.59)$.

Example: The sample relative risk in the vaccine example is

$$\frac{0.000164}{0.000548} = 0.2993$$

Using PROC FREQ in SAS, a 95% confidence interval for the relative risk is $(0.2028, 0.4416)$.

2.2.3 Relative Risk versus Absolute Risk

In medical applications, a large relative risk can be misleading in making interventions appear useful for rare outcomes. Physicians use the **number needed to treat** (NNT) which represents the number of patients who would need to receive a given intervention to prevent one case of the bad outcome:

$$NNT = \frac{1}{\pi_1 - \pi_2}.$$

Examples:

1. In both cases in the first example, $\pi_1 - \pi_2 = 0.008$ and $NNT = 125$.
2. If we change the first example to $\pi_1 = 0.46$ and $\pi_2 = 0.41$, we get $NNT = 1/0.05 = 20$.
3. In the polio example, $NNT = 1/0.000384 = 2604$.

2.3 Odds and Odds Ratio

We define the odds for success as the ratio of the probability of success to the probability of failure for each row:

$$\text{odds}_1 = \frac{\pi_1}{(1-\pi_1)} \qquad \text{odds}_2 = \frac{\pi_2}{(1-\pi_2)}$$

Example: $: \pi_1 = 0.8 \quad \longrightarrow \quad \text{odds}_1 = \frac{0.8}{0.2} = 4$

- $\text{odds} \geq 0$
- If $\pi_1 > (1-\pi_1)$, then $\text{odds}_1 > 1$.
- $\pi_1 = \frac{\text{odds}_1}{(1+\text{odds}_1)}$
- If $\pi_1 = \pi_2$, then $\text{odds}_1 = \text{odds}_2$.

Example: Leukemia Data

The odds of relapse for the group with traces of cancer are

$$\text{odds}_1 = \frac{0.40}{0.60} = 0.667$$

and for the group with no trace of cancer are

$$\text{odds}_2 = \frac{0.078}{0.922} = 0.084$$

Example: Vaccination Data

The odds of having polio for the vaccine group are

$$\text{odds}_1 = \frac{0.000164}{0.999836} = 0.000164$$

and for the non-vaccine group are

$$\text{odds}_2 = \frac{0.000548}{0.999458} = 0.000548$$

Statistics 659

The **odds ratio** is defined as

$$\theta = \frac{\text{odds}_1}{\text{odds}_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

Note: The odds ratio is sometimes known as the *cross-product ratio*. Consider the two way table:

	S	F
1	π_1	$1-\pi_1$
2	π_2	$1-\pi_2$

The odds ratio is the ratio of the product of the diagonal elements to the product of the off-diagonal elements.

Properties of the Odds Ratio

- $\theta \geq 0$
- $\theta = 1 \quad \Longleftrightarrow \quad \pi_1 = \pi_2$

- $\theta > 1 \quad \Longleftrightarrow \quad \pi_1 > \pi_2$
- $\theta < 1 \quad \Longleftrightarrow \quad \pi_1 < \pi_2$
- Larger values of θ indicate a stronger association.
- The odds ratio is also called the *cross-product ratio*.
- Often the log odds ratio, $\log(\theta)$, is used.
- If $\theta = 1$, then $\log(\theta) = 0$.
- Let $\theta' =$ odds ratio for 2 relative to 1. Then

$$\theta' = \frac{\pi_2(1-\pi_1)}{\pi_1(1-\pi_2)} = \frac{1}{\theta}$$

and

$$\log(\theta') = \log\left(\frac{1}{\theta}\right) = -\log(\theta).$$

Copyright ©2016 by Thomas E. Wehrly

Sample Odds Ratio:

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

Example: Leukemia Example:

$$\hat{\theta} = \frac{(0.40)/(0.60)}{(0.078)/(0.922)} = \frac{30 \times 95}{8 \times 45} = 7.92$$

Example: Vaccine Example:

$$\hat{\theta} = \frac{(0.000164)/(0.999836)}{(0.000548)/(0.999458)} = 0.29916$$

2.3.1 Inference for Odds Ratios and Log Odds Ratios

The sampling distribution of $\hat{\theta}$ can be highly skewed. The sampling distribution of $\log(\hat{\theta})$ is better behaved; that is, we can use a normal approximation.

- First find a confidence interval for $\log\theta$.
- Exponentiate the endpoints to form a confidence interval for θ.

The *asymptotic standard error* of the sampling distribution of $\log(\hat{\theta})$ is

$$SE = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

We approximate the distribution of $\log(\hat{\theta})$ using a normal distribution with mean $\log(\theta)$ and standard deviation given by the above SE.

A large sample level $1-\alpha$ confidence interval for $\log\theta$ is given by

$$\log(\hat{\theta}) \pm Z_{\alpha/2} SE.$$

After exponentiating the endpoints, the large sample level $1-\alpha$ confidence interval for θ is

$$\left(\hat{\theta}/\exp\left[Z_{\alpha/2}SE\right], \hat{\theta} \times \exp\left[Z_{\alpha/2}SE\right]\right).$$

Example: Leukemia:

$$SE = \sqrt{\frac{1}{30} + \frac{1}{45} + \frac{1}{8} + \frac{1}{95}} = 0.437$$

$$\log(\hat{\theta}) \pm Z_{\alpha/2} SE = \log(7.92) \pm (1.960)(0.437) = 2.069 \pm 0.857$$

The 95% confidence interval for $\log \theta$ is $(1.212, 2.926)$. The 95% c.i. for θ is $(3.36, 18.66)$.

Example: Vaccine:

$$SE = \sqrt{\frac{1}{201196} + \frac{1}{33} + \frac{1}{200674} + \frac{1}{110}} = 0.1985$$

$$\begin{aligned} \log(\hat{\theta}) \pm Z_{\alpha/2} SE &= \log(0.29916) \pm (1.960)(0.1985) \\ &= -1.2068 \pm 0.3891 \end{aligned}$$

The 95% confidence interval for $\log \theta$ is $(-1.5958, -0.8177)$.

The 95% c.i. for θ is $(0.2027, 0.4415)$ or $(2.2653, 4.9324)$ for the reciprocal of θ.

2.3.2 Relationship Between Odds Ratio and Relative Risk

$$\text{Odds Ratio} = \hat{\theta} = \frac{p_1(1-p_2)}{p_2(1-p_1)} = RR \times \frac{1-p_2}{1-p_1}$$

- If p_1 and p_2 are both close to zero, then $\hat{\theta} \approx RR$.

- For some data sets, due to their study designs, it is not possible to estimate RR. Then we use $\hat{\theta}$ to approximate RR.

2.3.3 The Odds Ratio and Case-Control Studies

In a case-control study, patients with the condition of interest are matched to one or more other patients without the condition who are similar to the case patients in other respects. Then each patient in the study is classified according to a variable thought to affect the condition.

Table 2.4 Cross-classification of Smoking Status and Myocardial Infarction (MI)

Ever	Myocardial	
Smoker	Infarction	Controls
Yes	172	173
No	90	346

- First column refers to 262 women with acute MI.
- Second column refers to sets of two patients with other acute conditions matched to each patient in column 1.
- All patients were classified according to smoking status.
- This is a *case-control study*.
- We cannot estimate $P(MI)$ or $P(MI|\text{Smoker})$ from this study.
- We can estimate $P(\text{Smoker}|MI)$ or $P(\text{Smoker}|\text{Other Acute Disease})$

2.3.4 Types of Studies

1. Prospective Study—subjects followed through time
   - Cohort Study—individuals stratified according to some variable (perhaps choice)
   - Clinical Trial—individuals assigned at random to groups of interest
2. Retrospective Study—the past of current subjects is studied
   - Case-control Study—match subjects with the condition (cases) with others free of the condition (controls)
   - Cross-sectional Study—classify individuals simultaneously on group and condition

All the above are **observational studies** except for the clinical trial which is an **experimental study**.

Examples: Patients are classified according to two variables:

- Myocardial infarction, case or control
- Smoker, yes or no

1. **Cross-Sectional Study**—Classify a random sample of n patients according to MI and to smoker. We obtain the following table of expected frequencies:

	MI	C
Y	$n \times \pi_{YMI}$	$n \times \pi_{YC}$
N	$n \times \pi_{NMI}$	$n \times \pi_{NC}$

$$OR = \frac{\pi_{YMI} \times \pi_{NC}}{\pi_{NMI} \times \pi_{YC}}$$

We can estimate all the conditional probabilities: $\pi_{Y|MI}$, $\pi_{Y|C}$, $\pi_{MI|Y}$, and $\pi_{MI|N}$.

2. **Case-Control Study**—For each case, match a given number of control subjects. We can estimate $\pi_{Y|MI}$ and $\pi_{Y|C}$, but not $\pi_{MI|Y}$, $\pi_{MI|N}$.

	MI	C
Y	$n_{MI} \times \pi_{Y\|MI}$	$n_C \times \pi_{Y\|C}$
N	$n_{MI} \times \pi_{N\|MI}$	$n_C \times \pi_{N\|C}$

$$\begin{aligned} OR &= \frac{\pi_{Y|MI} \times \pi_{N|C}}{\pi_{N|MI} \times \pi_{Y|C}} \\ &= \frac{\left(\frac{\pi_{YMI}}{\pi_{MI}}\right) \times \left(\frac{\pi_{NC}}{p_C}\right)}{\left(\frac{\pi_{NMI}}{\pi_{MI}}\right) \times \left(\frac{\pi_{YC}}{\pi_C}\right)} \\ &= \frac{\pi_{YMI} \times \pi_{NC}}{\pi_{NMI} \times \pi_{YC}} \end{aligned}$$

3. **Clinical Trial**—Randomly assign n_Y smokers and n_N nonsmokers. Classify according to MI. We can estimate $\pi_{MI|Y}$ and $\pi_{MI|N}$, but not $\pi_{Y|MI}$, $\pi_{Y|C}$.

	MI	C
Y	$n_Y \times \pi_{MI\|Y}$	$n_Y \times \pi_{C\|Y}$
N	$n_N \times \pi_{MI\|N}$	$n_N \times \pi_{C\|N}$

$$\begin{aligned} OR &= \frac{\pi_{MI|Y} \times \pi_{C|N}}{\pi_{MI|N} \times \pi_{C|Y}} \\ &= \frac{\pi_{YMI} \times \pi_{NC}}{\pi_{NMI} \times \pi_{YC}} \end{aligned}$$

2.4 Chi-Squared Tests of Independence

We consider tests of the null hypothesis (H_0) that the cell probabilities equal certain fixed values $\{\pi_{ij}\}$. When H_0 is true, the expected frequencies are

$$\mu_{ij} = n\pi_{ij} = E(n_{ij})$$

.

The test statistics measure the closeness of the observed cell frequencies $\{n_{ij}\}$ to the expected cell frequencies $\{\mu_{ij}\}$.

1. Pearson's Chi-Squared Statistic

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

2. Likelihood Ratio Statistic

$$G^2 = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log\left(\frac{n_{ij}}{\mu_{ij}}\right).$$

- Both statistics have an approximate chi-squared distribution when H_0 is true.
- We typically require that all $\mu_{ij} \geq 5$.
- Both test statistics equal zero if all $n_{ij} = \mu_{ij}$.
- We reject H_0 for large values of the test statistic.
- Usually the two statistics have similar numerical values and will lead to the same conclusion.
- Both statistics stay the same if the rows or columns are reordered.
- If the cell means depend on unknown parameters, we estimate the parameters using maximum likelihood and modify the degrees of freedom.

2.4.1 Tests of Independence

We now consider testing whether X and Y are independent:

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}, \text{ all } i, j$$

Under H_0, $\pi_{ij} = \pi_{i+}\pi_{+j}$ and $\mu_{ij} = E(n_{ij}) = n\pi_{ij} = n\pi_{i+}\pi_{+j}$. This is estimated by

$$\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = n\frac{n_{i+}}{n}\frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}$$

For testing independence in an $I \times J$ table, Pearson's and the LR statistics are

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \quad \text{and} \quad G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right).$$

The degrees of freedom equal $(I-1)(J-1)$.

In general the d.f. of the test statistic equals the difference in the numbers of parameters between the null and alternative hypotheses.

Null hypothesis: $I-1$ row probabilities and $J-1$ column probabilities for a total of $I+J-2$.

Alternative hypothesis: $IJ-1$ cell probabilities

$$df = IJ - 1 - (I + J - 2) = IJ - I - J + 1 = (I-1)(J-1)$$

Other Test Statistics on SAS Output:

- Continuity Adjusted χ^2 Test:

$$Q_C = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{[max(0, |n_{ij} - \hat{\mu}_{ij}| - 0.5)]^2}{\hat{\mu}_{ij}}$$

- Mantel-Haenszel χ^2 Test:

$$Q_{MH} = \frac{(n_{11} - \hat{e}_{11})^2}{v_{11}}$$

  where

$$v_{11} = \widehat{\mathrm{Var}}(n_{11}|H_0) = \frac{n_{1+}n_{2+}n_{+1}n_{+2}}{n^2(n-1)}$$

  - This is based on the hypergeometric distribution discussed with Fisher's exact test.
  - For a 2×2 table, $Q_{MH} = \frac{(n-1)}{n} X^2$

2.4.2 Odds Ratios in $I \times J$ Tables

Odds ratios can be formed for each pair of rows with each pair of columns resulting in $\binom{I}{2} \times \binom{J}{2}$ possible odds ratios. Often the *local odds ratios* between adjacent rows and columns are computed.

Example: Cross-Classification of Aspirin Use and Myocardial Infarction

	Myocardial Infarction		
	Fatal Attack	Nonfatal Attack	No Attack
Placebo	18	171	10845
Aspirin	5	99	10933

The local sample odds ratios are

$$\begin{aligned} \hat{\theta}_{12} &= \tfrac{18\times 99}{171\times 5} = 2.08 \\ \hat{\theta}_{23} &= \tfrac{171\times 10933}{10845\times 99} = 1.74 \end{aligned}$$

The other odds ratio between columns 1 and 3 can be computed by taking the product, $\hat{\theta}_{13} = 2.08 \times 1.74 = 3.63.$

2.4.3 Residual Analysis for Tests of Independence

We often wish to follow up a test of independence that results in a significant association by studying the nature of the association. One approach is to examine the observed and estimated expected frequencies of the various cells. Since larger differences tend to occur in cells with larger expected frequencies, it is useful to use adjusted residuals rather than the raw difference.

Two commonly used cell residuals are the Pearson's residual

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

and the adjusted Pearson's residual

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

When the null hypothesis of independence is true, the adjusted residuals have approximately a standard normal distribution. Large values of these residuals indicate the cells that are discrepant with independence.

In many tables, there will be a general interaction between the variables that cause us to reject independence. In other tables, the assumption of independence may be reasonable except for a few unusual cells. In this case we can use the residuals to identify these *outlier cells*.

Example: College students were classified according both to frequency of marijuana use and parental use of alcohol and psychoactive drugs.

The SAS output for the test of independence follows:

```
Chi-Square                      4      24.4171    <.0001
Likelihood Ratio Chi-Square     4      24.3571    <.0001
```

We conclude that there is strong evidence of association between parental use of alcohol and student drug use. The standardized residuals are in the following table:

		Level of Marijuana Use		
		Never	Occasional	Regular
Parental Use	Neither	4.63	-1.64	-3.73
of Alcohol	One	-3.31	1.63	2.23
and Drugs	Both	-2.29	0.11	2.53

2.4.4 Partitioning Chi-Square

We refer to a couple of facts from statistical distribution theory.

- If Z has a standard normal distribution, then Z^2 has a chi-squared distribution with 1 degree of freedom.
- If $Z_1, Z_2, \ldots, Z_\nu$ are independent standard normal random variables, then $Z_1^2 + \cdots + Z_\nu^2$ has a chi-squared distribution with ν degrees of freedom.
- These facts imply that a random variable that has a chi-squared distribution with ν degrees of freedom can be represented as a sum of ν independent chi-squared random variables each with 1 degree of freedom.
- We can also partition the chi-squared statistic as a sum of independent chi-squared random variables whose degrees of freedom sum to the degrees of freedom for the original statistic.
- In some cases we can partition a chi-squared statistic so that the components reflect some important properties of the dependence structure. This partitioning may indicate that the association is due to differences between certain categories or groups of categories.
- We can use either G^2 or Pearson's statistic to test the individual tables. However, the sum of the Pearson's statistics for the individual tables does not equal the value of the overall test statistic.

- For a $2 \times J$ table, the G^2 statistic has $J - 1$ d.f. We can partition it into $J - 1$ components.

- A simple method of partitioning the table follows:

  The j^{th} component is the G^2 statistic for testing independence in a 2×2 table where column one combines columns 1 through j in the original table and column two is column $j + 1$ in the original table. Each G^2 value has 1 d.f.

- The "obvious" method of comparing each of the first $j - 1$ columns in turn to the last does not result in independent components.

- For general $I \times J$ tables, more sophisticated approaches to forming the partition must be used. We can use methods analogous to the above to form $J - 1$ statistics for the columns or $I - 1$ statistics for the rows.

Example: **Aspirin and Myocardial Infarction**

On the SAS output, we find the following values of G^2:

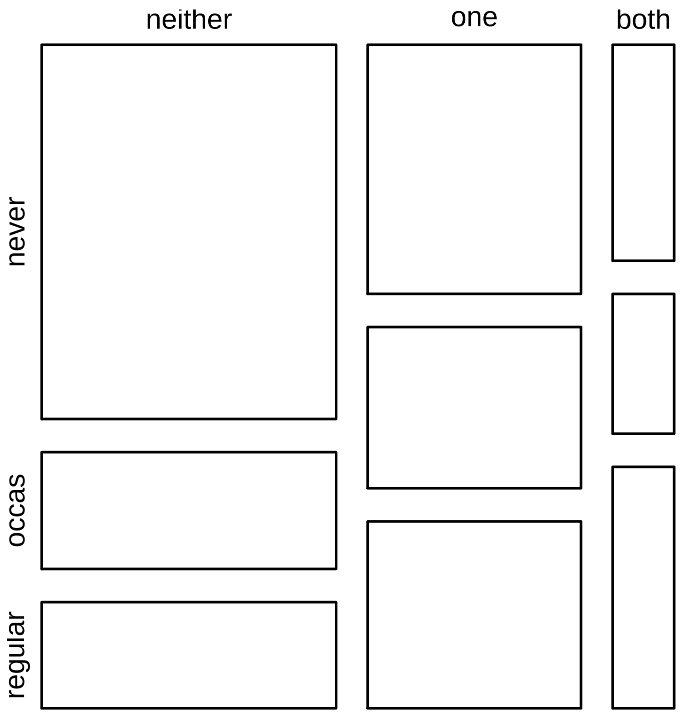
First two columns	2.2173
First two columns vs. column 3	25.3720
Overall table	27.5893

2.4.5 Mosaic Plots

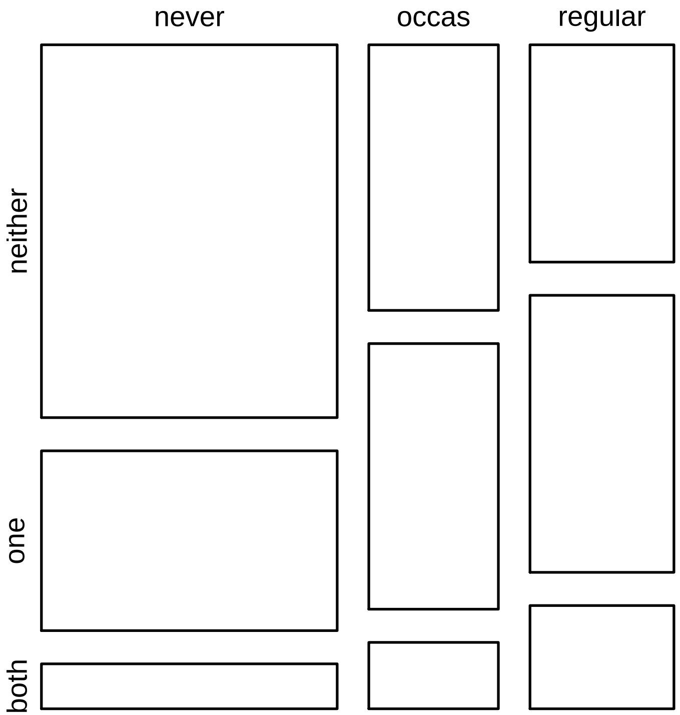
A graphical representation of an $I \times J$ table that helps to visualize the deviations from independence is the **mosaic plot**.

- The width of the box is determined by the percentage of observations that fall in the column.
- The heights of the boxes are determined by the column percents for each of the row categories.
- In the case of independence, all the boxes in a given row would be the same height.
- Here the column variable is viewed as an explanatory variable with the row variable as the response.
- The roles of the variables can be exchanged by forming a mosaic plot for the table with rows and columns interchanged.

Parent’s Use in Columns



Children’s Use in Columns



2.5 Testing Independence for Ordinal Data

The methods for two-way tables earlier in this chapter are most appropriate for nominal data. We now look at methods that are more powerful when data are ordinal.

2.5.1 Review of Linear Correlation

When X and Y are both **numerical** random variables, a measure of their linear relationship is the population covariance:

$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

The corresponding sample measure of their relationship is the sample covariance:

$$\frac{1}{n-1}\sum(x_i-\bar{x})(y_i-\bar{y})$$

.

A measure not affected by the units of measurement is the population correlation coefficient:

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}$$

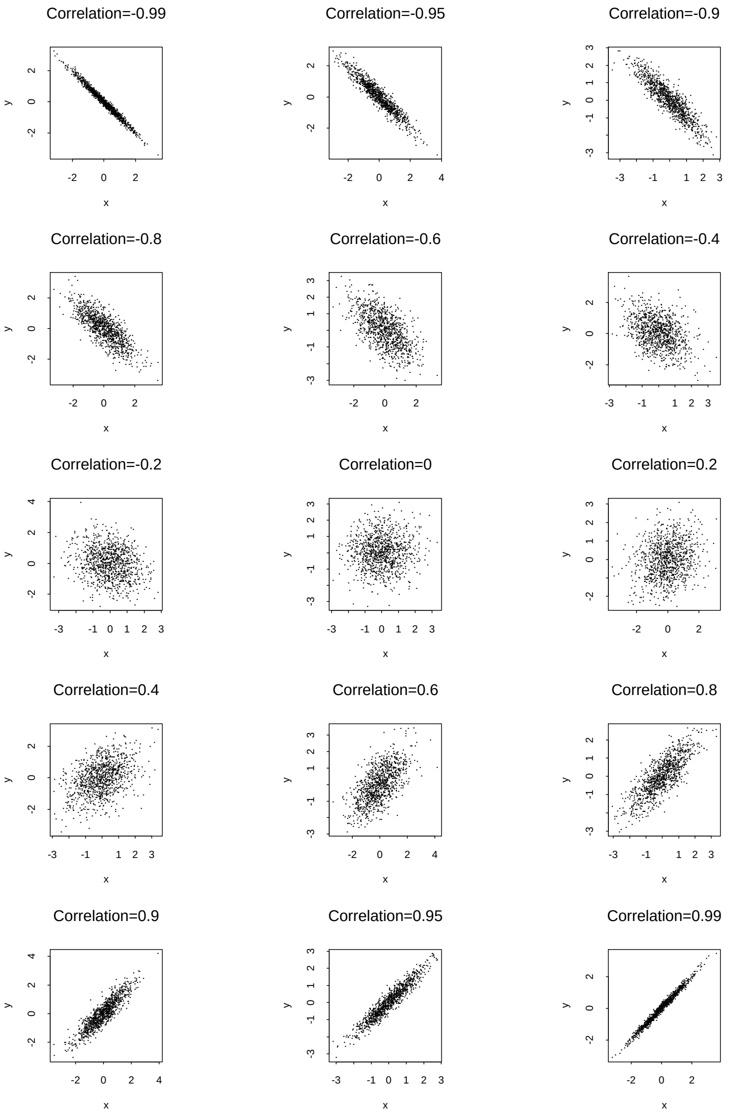
The corresponding sample measure of linear relations is the sample (Pearson's) correlation coefficient:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Remarks:

1. $-1 \leq \rho \leq 1$
2. $\rho = \pm 1$ if all the distribution of (X, Y) is concentrated on a straight line.
3. ρ near zero indicates no linear relationship.
4. $\rho > 0$ indicates that Y has a tendency to increase as X increases.
5. $\rho < 0$ indicates that Y has a tendency to decrease as X increases.
6. r has a similar interpretation for the scatter plot of n (x, y) pairs.

Random Samples from Bivariate Normal Distributions



2.5.2 Tests and Measures of Association for $I \times J$ Contingency Tables

- The previous tests and measures of association treated both categorical variables X and Y as nominal.

- If either the rows or columns are ordinal, test statistics that use the order in X or Y are more appropriate.

- If the data (X and Y) have an interval numerical scale (scores that are evenly spaced), the Pearson correlation coefficient is an appropriate measure of association.

- The Spearman rank correlation coefficient is obtained by replacing the variables with their ranks in the Pearson correlation coefficient.

- To test for linear trends, one must assign **scores** to the categories.

  The scores should be **monotone**, that is, have the same ordering as the category levels.

  The differences in scores represent the distances between the categories. For Spearman's correlation, the scores are 1,2,3, etc.

Let $u_1 \leq u_2 \leq \cdots \leq u_I$ denote the scores for the rows.

Let $v_1 \leq v_2 \leq \cdots \leq v_J$ denote the scores for the columns.

row$\backslash$ column scores	v_1	v_2	$\cdots$	v_J	
u_1	n_{11}	n_{12}	$\cdots$	n_{1J}	n_{1+}
u_2	n_{21}	n_{22}	$\cdots$	n_{2J}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
u_I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	n_{I+}
Sum	n_{+1}	n_{+2}	$\cdots$	n_{+J}	n

Letting $\bar{u} = \sum_i u_i p_{i+}$ be the sample mean of the row scores and $\bar{v} = \sum_j v_j p_{+j}$ be the sample mean of the column scores, the formula for Pearson's correlation between X and Y is:

$$r = \frac{\sum_{i,j}(u_i - \bar{u})(v_j - \bar{v})p_{ij}}{\sqrt{(\sum_i (u_i - \bar{u})^2 p_{i+})\left(\sum_j (v_j - \bar{v})^2 p_{+j}\right)}}$$

- A statistic for testing the null hypothesis of independence against the two-sided alternative of nonzero correlation is

$$M^2 = (n-1)r^2$$

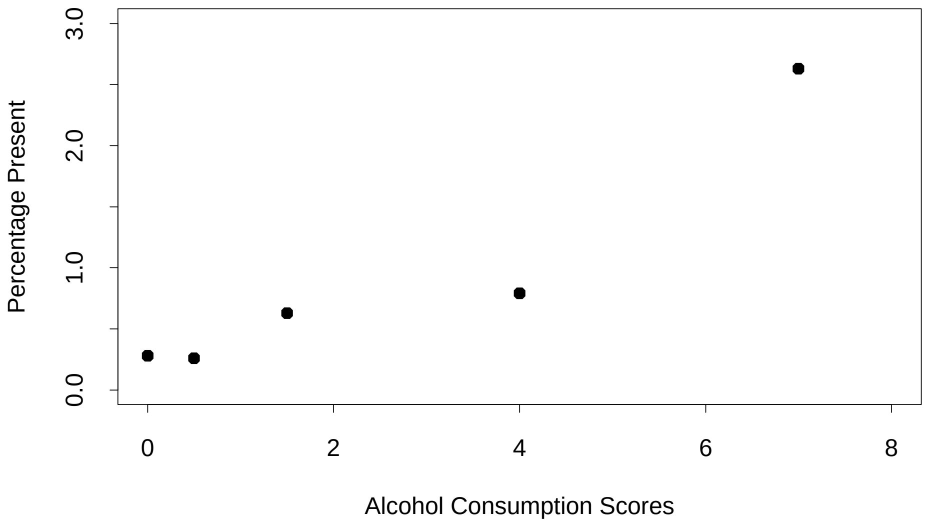
  - For large samples, M^2 has approximately a chi-squared distribution with $df = 1$. We reject independence for large values of M^2.
  - This statistic is sensitive to positive (or negative) trends in the relationship of two ordinal categorical variables. When there is truly a “linear” trend, M^2 will tend to have more power than the Pearson and likelihood ratio chi-squared tests, which are designed for a general alternative with $df = (I-1)(J-1)$.

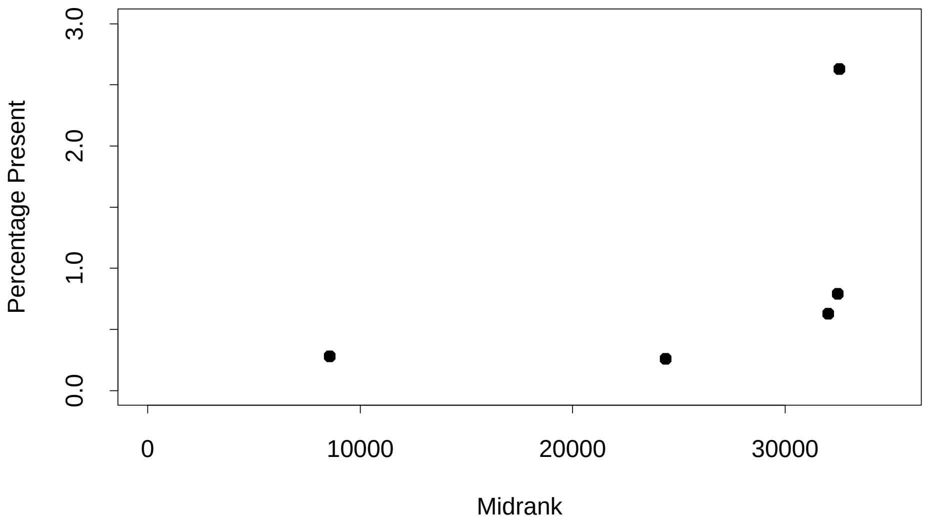
- Choice of Scores — For most data sets, the choice of scores has little effect on the results. However, it the data are unbalanced, the choice of monotone scores can affect the value of M^2.
  - Integer Scores: These are useful when the ordered categories can be considered as equally spaced.
  - Rank-based Scores: For a two-way table, the use of ranks and midranks produces Spearman’s rank correlation coefficient.
  - It is sometimes useful to carry out a *sensitivity analysis* to see how sensitive the conclusions are to the choice of scores.

Example: Infant Mortality Data from Agresti, Chapter 2

A prospective study was carried out of maternal drinking and congenital malformations. A survey was completed on alcohol consumption at three weeks of pregancy. Following childbirth observations were made concerning the presence or absence of congenital sex organ malformations.

Alcohol	Malformation	Malformation	Total
Consumption	Absent	Present	
0	17066	48	17144
< 1	14464	38	14502
$1 - 2$	788	5	793
$3 - 5$	126	1	127
≥ 6	37	1	38


Infant Malformation and Mother’s Alcohol Consumption


Plot Using Midranks

2.5.3 Trend Tests for $I \times 2$ and $2 \times J$ Tables

We suppose that X is an explanatory variable and Y is a response variable.

- $2 \times J$ Table: Tables like this represent two groups where the rows represent two treatments. The M^2 statistic is directed toward detecting differences in the row mean scores. When we use the midrank scores, the test is sensitive to testing differences in average ranks for the two rows. This is a form of the *Wilcoxon* or *Mann-Whitney test*.

- $I \times 2$ Table: The response variable is binary. Here we look for a trend in the proportion of success across the rows. This test is called the *Cochran-Armitage trend test*.

2.5.4 Some Ordinal Measures of Association on SAS Output

- Other measurements are based on the number of concordant and discordant pairs in the data.

  A pair is **concordant** if the category ranking higher on the row variable also ranks higher on the column variable ($x_i < x_j$ and $y_i < y_j$).

  A pair is **discordant** if the category ranking higher on the row variable also ranks lower on the column variable ($x_i < x_j$ and $y_i > y_j$).

- Let $C =$ the number of concordant pairs.
- Let $D =$ the number of discordant pairs.
- Kendall's $\tau = \frac{C-D}{n(n-1)/2}$
- $\tau = 1$ indicates that $x_i < x_j$ whenever $y_i < y_j$.
- $\tau = -1$ indicates that $x_i < x_j$ whenever $y_i > y_j$.

- The measures, gamma, Kendall’s Tau-b, Stuart’s Tau-c, and Somer’s D, are all based on the numbers of concordant and discordant pairs.

  These measures take values between -1 and 1.

  These measures differ mainly in their strategies for adjusting for ties and sample sizes.

- $\text{Gamma} = \hat{\gamma} = \frac{C-D}{C+D}$
- Kendall’s Tau-b (τ_b) is similar except it uses a correction for ties.
- Stuart’s Tau-c (τ_c) also makes an adjustment for table size.
- Somer’s D (C$\vert$R) is an asymmetric modification of Tau-b where the column variable is the dependent or response variable.
- Somer’s D (R$\vert$C) is an asymmetric modification of Tau-b where the row variable is the dependent or response variable.
- Gamma will have the largest value of these measures.

2.5.5 Some Numerical Measures of Nominal Association on SAS Output

- The first three measures are based on Pearson's chi-squared statistic, Q_p.
- Phi coefficient: $\phi = \sqrt{Q_p/n}$
- Contingency coefficient: $P = \sqrt{\frac{Q_p}{Q_p+n}}$
- Cramer's V: $V = \sqrt{\frac{Q_p}{n}\frac{1}{\min(I-1,J-1)}}$
- Asymmetric lambda, $\lambda(C|R)$, is interpreted as the probable improvement in predicting the column variable Y given knowledge of the row variable X. The nondirectional lambda is the average of the two asymmetric lambdas, $\lambda(C|R)$ and $\lambda(R|C)$.

2.6 Exact Tests for Small Samples

2.6.1 Fisher's Exact Test

- Consider first the 2×2 table. We will consider the exact distribution of counts for the sets of tables have the same row and column totals as the observed data.

- Under Poisson, independent binomial, or multinomial sampling for the cell counts, the distribution that applies to the set of tables with fixed row and column totals is the **hypergeometric distribution**.

- H_0 : independence of two categorical variables

Review of the Hypergeometric Distribution

Consider a population of N objects of which M are labelled as defective. Take a sample of size n without replacement from the population. Define the random variable X to be the number of defective items in the sample. The r.v. X has p.m.f.

$$p(x) = P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

where max$(0, n - N + M) \leq x \leq$ min(n, M).

The mean and variance of a hypergeometric r.v. are

$$\begin{aligned} E[X] &= \mu &&= n\tfrac{M}{N} = n\pi \text{ where } \pi = \tfrac{M}{N} \\ \mathrm{Var}[X] &= \sigma^2 &&= (\tfrac{N-n}{N-1})n\tfrac{M}{N}(1 - \tfrac{M}{N}) \\ & &&= (\tfrac{N-n}{N-1})\, n\pi(1 - \pi) \end{aligned}$$

- When the row and column totals in a 2×2 table are fixed, we can express probabilities for the 4 cell counts in terms of n_{11} only.

n_{11}		n_{1+}
		n_{2+}
n_{+1}	n_{+2}	

- To test H_0, the p-value is the sum of the hypergeometric probabilities for outcomes at least as favorable to the alternative hypothesis as the observed outcome.

- When H_0 holds, the probability of a particular value for n_{11} is

$$p[n_{11}] = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}} = \frac{n_{1+}!n_{2+}!n_{+1}!n_{+2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}$$

- The mean and variance of the cell counts are

$$E[n_{ij}|H_0] = \frac{n_{i+}n_{+j}}{n} \text{ and } \mathrm{Var}[n_{ij}|H_0] = \frac{n_{1+}n_{2+}n_{+1}n_{+2}}{n^2(n-1)}$$

2.6.2 Fisher’s Tea Taster

A colleague of R. A. Fisher claimed she could determine if milk were added to the cup of tea first. Fisher designed an experiment where she tasted eight cups of tea. Four of these had milk added first, and four had tea added first. The results of the experiment were:

	Guess Poured First		
Poured First	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	8

In this experiment both column and row totals are fixed since she knew that 4 cups had milk added first. The distribution of n_{11} is hypergeometric with potential values $\{0, 1, 2, 3, 4\}$.

The possible tables are

4	0
0	4

3	1
1	3

2	2
2	2

1	3
3	1

0	4
4	0

Use the hypergeometric distribution to find probabilities. For example,

$$P(3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = 0.229$$

Hypergeometric Distribution for Example

n_{11}	Probability	P-value	mid $P-$value	X^2
0	0.0143	1.0000	0.9929	8.0
1	0.2286	0.9857	0.8714	2.0
2	0.5143	0.7571	0.5000	0.0
3	0.2286	0.2429	0.1286	2.0
4	0.0143	0.0143	0.0071	8.0

The table gives P-values for testing $H_0 : \theta = 1$ versus $H_a : \theta > 1$. The only table more extreme than the one observed is the last one. Thus, the P-value and mid P-value are

$$\begin{aligned} P-\text{value} &= P(3) + P(4) = .2286 + .0143 = 0.2429, \\ \text{mid } P-\text{value} &= 0.5P(3) + P(4) = 0.5(.2286) + .0143 = 0.1286. \end{aligned}$$

Two-Sided Alternative: Consider testing $H_0 : \theta = 1$ versus $H_a : \theta \neq 1$. The exact P-value is defined as the sum of probabilities of tables no more likely than the observed table.

Example: Fisher's Tea Taster

$P -$ value $= 2(.014) + 2(.229) = 0.486$

Remarks

- The `EXACT` statement in the `FREQ` procedure is used to carry out exact inference for contingency tables.
- We can use SAS to carry out exact tests for independence for $I \times J$ tables. The computations may be slow for very large tables. A *Monte Carlo* approach using randomly generated tables may be necessary to compute $P-$values.
- The `RISKDIFF` option in the `EXACT` statement provides an exact confidence interval for the difference in proportions.
- The `OR` option in the `EXACT` statement provides an exact confidence interval for the odds ratio.

2.7 Three-Way Contingency Tables

In the previous sections we examined association between two categorical variables, X and Y. In many situations, there are other variables that can affect the relationship between the two variables of interest.

Example: A study was carried out to compare two treatments for a respiratory disorder. The goal was to compare the proportions of patients responding favorably to test and placebo. A confounding factor is that the study was carried out at two centers which had different patient populations. We wish to examine the association between treatment and response while adjusting for the effects of the centers.

		Respiratory Improvement		
Center	Treatment	Yes	No	Total
1	Test	29	16	45
1	Placebo	14	31	45
Total		43	47	90
2	Test	37	8	45
2	Placebo	24	21	45
Total		61	29	90

- In studying the effect of an explanatory variable X on a response variable Y, one should take into account other variables (covariates) that may influence the relationship. Otherwise, an observed effect of X on Y may simply reflect the associations of the covariates on X and Y.

- One strategy for examining the association between two variables while adjusting for the effect of others is *stratified analysis*.
  - In some cases the stratification can result from the study design. This is the case in a multicenter clinical trial.
  - In other cases, it may arise from a post-study stratification to control for the effects of certain explanatory variables. This is often used in observational studies where randomization cannot be used.

- Each two-way table corresponds to one stratum. The strata are determined by the unique combinations of the explanatory variables.

- The analysis of sets of tables addresses many of the same questions asked in the analysis of a single table:
  - Is there an association between rows and columns?
  - What is the strength of that association (odds ratio)?

  Here we are looking at the overall association instead of the association in just one table.

2.8 Partial Association

Multivariate Categorical Data: (X, Y, Z).

$X =$ explanatory variable (predictor)

$Y =$ response

$Z =$ control variable (covariate)

2.8.1 Partial Tables

- A **partial table** is a two-way cross-sectional table that classifies X and Y at the different levels of the control variable Z.
  - They display the relationship between X and Y at fixed levels of Z.
  - They control (remove) the effect of Z by holding its value constant.
- The X–Y marginal table is formed by combining the partial tables.
  - It contains no information about Z.
  - It ignores Z rather than controlling for it.

Example: The marginal table for the respiratory disease data follows:

	Respiratory Improvement		
Treatment	Yes	No	Total
Test	66	24	90
Placebo	38	52	90
Total	104	76	180

- The associations in partial tables are called *conditional associations*.

- The association in the marginal table is called the *marginal association*.

- One can describe these associations using odds ratios.

2.8.2 Conditional and Marginal Odds Ratios

Suppose we have a $2 \times 2 \times K$ table. This corresponds to K different 2×2 tables, one for each of the K levels of Z,

- The *conditional odds ratios* are the odds ratios for the partial tables.
  - Let n_{ijk} denote the observed count of $X = i,\ Y = j$, and $Z = k$, and let $E(n_{ijk}) = \mu_{ijk}$.
  - The odds ratio and its estimate for the k^{th} partial table are

$$\theta_{XY(k)} = \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}}, \qquad \text{and} \qquad \hat{\theta}_{XY(k)} = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}.$$

  - $\theta_{XY(k)}$ measures the conditional X–Y association when $Z = k$.

- The marginal odds ratio is the odds ratio for the marginal X–Y table

$$\theta_{XY} = \frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}}, \qquad \text{and} \qquad \hat{\theta}_{XY} = \frac{n_{11+}n_{22+}}{n_{12+}n_{21+}}.$$

Example: Compute the conditional and marginal odds ratio for the respiratory disease data set.

$$\hat{\theta}_{XY(1)} = \frac{n_{111}n_{221}}{n_{121}n_{211}} = \frac{29\cdot 31}{14\cdot 16} = 4.013$$

$$\hat{\theta}_{XY(2)} = \frac{n_{112}n_{222}}{n_{122}n_{212}} = \frac{37\cdot 21}{24\cdot 8} = 4.047$$

$$\hat{\theta}_{XY} = \frac{n_{11+}n_{22+}}{n_{12+}n_{21+}} = \frac{66\cdot 52}{38\cdot 24} = 3.763$$

2.8.3 Simpson's Paradox

We will illustrate Simpson's paradox with an example.

Example: Applicants can apply for either a sales position or an office position. For each position, each applicant was classified according to gender and according to whether the applicant was offered the position. The combined table follows:

	Result		
Gender	Offer	Deny	Total
Male	160	115	275
Female	160	165	325
Total	320	280	600

The $2 \times 2 \times 2$ table follows:

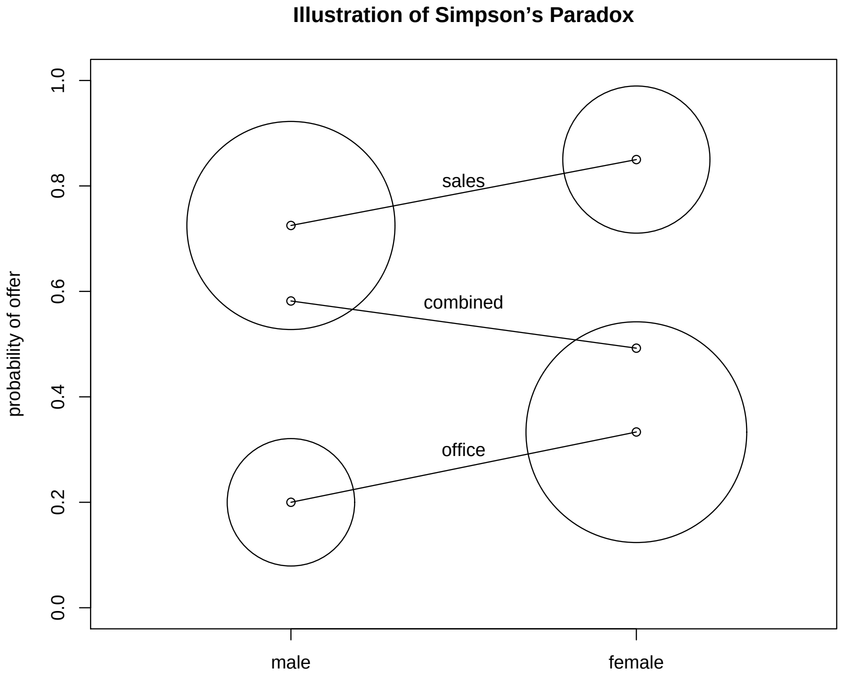
		Result		
Type of Position	Gender	Offer	Deny	Total
Office	Male	15	60	75
Office	Female	75	150	225
Total		90	210	300
Sales	Male	145	55	200
Sales	Female	85	15	100
Total		230	70	300

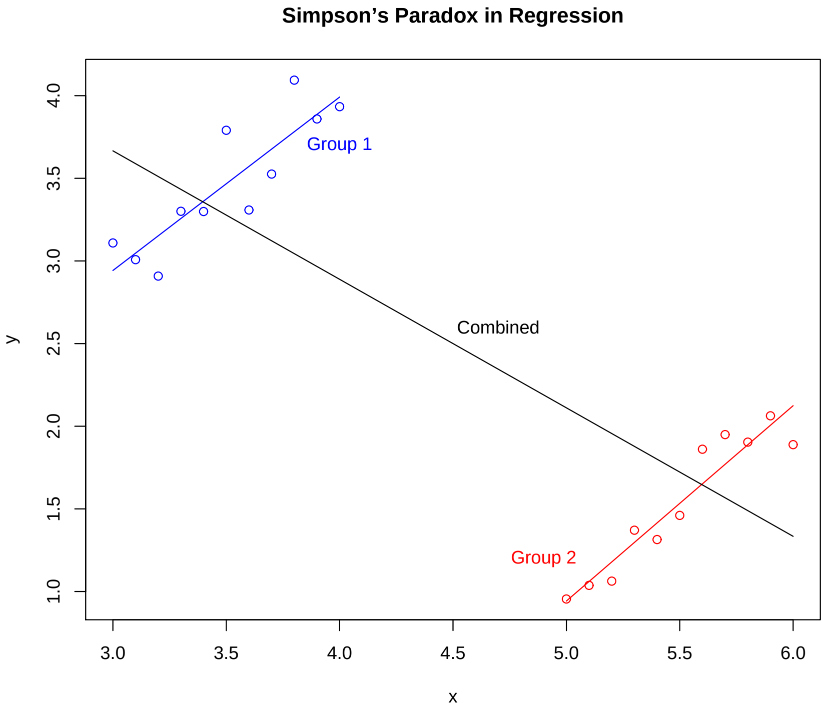
The resulting odds ratios are

$$\begin{aligned} \hat{\theta}_{XY} &= \frac{165 \cdot 160}{115 \cdot 160} &= 1.435 \\ \hat{\theta}_{XY(1)} &= \frac{150 \cdot 15}{60 \cdot 75} &= 0.5 \\ \hat{\theta}_{XY(2)} &= \frac{15 \cdot 145}{85 \cdot 55} &= 0.465 \end{aligned}$$

Thus, the marginal association has a different direction than the conditional associations. This phenomenon is called *Simpson's paradox*.

Statistics 659





2.8.4 Marginal versus Conditional Independence

- If X and Y are independent in each partial table, they are said to be *conditionally independent*. In this case,

$$\theta_{XY(k)} = 1, \text{ for all } k = 1, \ldots, K$$

- If X and Y are independent in the marginal table, they are said to be *marginally independent*. In this case,

$$\theta_{XY} = 1$$

- Conditional independence does not imply marginal independence. This is shown in Table 2.11 from Agresti:

		Response		
Clinic	Treatment	Success	Fail	
1	A	18	12	$\hat{\theta}_{XY(1)} = 1$
	B	12	8	
2	A	2	8	$\hat{\theta}_{XY(2)} = 1$
	B	8	32	
Total	A	20	20	$\hat{\theta}_{XY} = 2$
	B	20	40	

2.8.5 Homogeneous Association

- A $2 \times 2 \times K$ table has *homogeneous association* if

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

- Conditional independence is a special case of homogeneous association.

$$\text{All } \theta_{XY(k)} = 1$$

- A general $I \times J \times K$ table has *homogeneous association* if any conditional odds ratio formed using two levels of X and two levels of Y is the same at any level of Z.

- Homogeneous association is a symmetric property. For variables $X,\ Y$, and Z, if there is a homogeneous X–Y association, there are also homogeneous X–Z and Y–Z associations.

- When homogeneous association occurs, there is no *interaction* between two variables in their effect on a third.

2.9 Cochran-Mantel-Haenszel Methods

We consider K sets of 2×2 tables where the k^{th} table, $k = 1, \ldots, K$, has counts

$X \backslash Y$	1	2	Total
1	n_{11k}	n_{12k}	n_{1+k}
2	n_{21k}	n_{22k}	n_{1+k}
Total	n_{+1k}	n_{+2k}	n_{++k}

2.9.1 Testing Homogeneity of the Odds Ratios

We consider testing the null hypothesis of equality of the odds ratios across the K strata:

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_K$$

The *Breslow-Day* test statistic has the Pearson's chi-squared form:

$$Q_{BD} = \sum_{k=1}^{K} \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ijk} - e_{ijk})^2}{e_{ijk}}$$

where e_{ijk} is the estimated expected cell count in the k^{th} partial table under the null hypothesis of equal odds ratios.

The Breslow-Day statistic is calculated assuming

1. The Mantel-Haenszel estimator $\hat{\theta}_{MH}$ of the common odds ratio (on the next slide) is the odds ratio for each stratum.
2. The estimated expected cell frequencies for each partial table have the same row and column totals as the observed tables.
3. The Breslow-Day statistic is approximately chi-squared with $df = K - 1$. The sample sizes need to be relatively large in each partial table with $e_{ijk} \geq 5$ in at least 80% of the cells.

2.9.2 Estimation of a Common Odds Ratio

In a $2 \times 2 \times K$ table, when the association seems stable across the K partial tables, we can estimate an assumed common value for the K true odds ratio.

- Recall that the odds ratio for the k^{th} partial table is estimated by

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

The standard error of $\log(\hat{\theta}_k)$ is estimated by

$$\sqrt{\frac{1}{n_{11k}} + \frac{1}{n_{12k}} + \frac{1}{n_{21k}} + \frac{1}{n_{22k}}}$$

- Suppose that $\theta_1 = \theta_2 = \cdots = \theta_K$. The *Mantel-Haenszel estimator* of the common odds ratio is

$$\hat{\theta}_{MH} = \frac{\displaystyle\sum_{k=1}^{K} \frac{n_{11k} n_{22k}}{n_{++k}}}{\displaystyle\sum_{k=1}^{K} \frac{n_{12k} n_{21k}}{n_{++k}}}$$

The formula for the standard error of $\log(\hat{\theta}_{MH})$ is complicated.

- Another estimator for the common odds ratio is the *logit estimator*:

$$\hat{\theta}_L = \exp\left\{ \frac{\displaystyle\sum_{k=1}^{K} w_k \log(\hat{\theta}_k)}{\displaystyle\sum_{k=1}^{K} w_k} \right\}$$

The weights are given by

$$w_k = \left(\frac{1}{n_{11k}} + \frac{1}{n_{12k}} + \frac{1}{n_{21k}} + \frac{1}{n_{22k}} \right)^{-1}$$

- A 100$(1-\alpha)$% confidence interval for θ based on $\hat{\theta}_L$ is

$$\exp\left\{\log(\hat{\theta}_L) \pm Z_{\alpha/2}\left[\sum_{k=1}^{K} w_k\right]^{-\frac{1}{2}}\right\}.$$

- The logit estimator is also a reasonable estimator, but it requires adequate sample sizes (all $n_{ijk} > 5$). The Mantel-Haenszel estimator is not as sensitive to sample size.
- When the counts are small and you want to find an exact confidence interval for the common odds ratio, you need to use logistic regression.
- If the true odds ratios are not identical but do not vary drastically (in a single direction), $\hat{\theta}_{MH}$ or $\hat{\theta}_L$ provides a useful summary of the K conditional associations.
- If the odds ratios are not homogeneous, the the common odds ratio should be viewed cautiously. One should emphasize the within-strata odds ratios.

2.9.3 Testing Conditional Independence in Sets of 2×2 Tables

- We consider the null hypothesis that X and Y are conditionally independent; that is

$$H_0 : \theta_{XY(k)} = 1,\ k = 1, \ldots, K$$

- Under the null hypothesis,

$$E[n_{11k}|H_0] = e_{11k} = \frac{n_{1+k}n_{+1k}}{n_{++k}} \quad \text{and} \quad \mathrm{Var}[n_{11k}|H_0] = v_{11k} = \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n_{++k}^2(n_{++k}-1)}.$$

- The Cochran-Mantel Haenszel statistic summarizes the information from the K partial tables:

$$CMH = Q_{CMH} = \frac{\left[\sum_{k=1}^{K}(n_{11k} - e_{11k})\right]^2}{\sum_{k=1}^{K} v_{11k}} = \frac{\left[\sum_{k=1}^{K} \frac{n_{1+k}n_{2+k}}{n_{++k}}(p_{11k} - p_{21k})\right]^2}{\sum_{k=1}^{K} v_{11k}}$$

where $p_{i1k} = n_{i1k}/n_{i+k}$ is the proportion of subjects from the k^{th} stratum and i^{th} group to have a favorable response.

- When there is only one stratum ($K = 1$), the CMH statistic reduces to $(n-1)Q_p/n$ where Q_p is Pearson's chi-squared statistic. In this case, CMH equals the Mantel-Haenszel chi-squared statistic in SAS `PROC FREQ`. (See p. 29 of the notes for this chapter.)

- When there is more than one stratum, CMH becomes a stratum-adjusted chi-squared statistic.

- The CMH statistic is large when $(n_{11k} - e_{11k})$ is consistently positive or consistently negative for all the tables.

- The CMH statistic has approximately a chi-squared distribution with $df = 1$ when H_0 is true. This approximation holds when the combined row sample sizes ($\sum_{k=1}^{K} n_{i+k} = n_{i++}$) are large enough ($> 30$).

- The CMH statistic potentially removes the confounding effect of the explanatory variables that define the strata and can provide an increase in the power for detecting association by comparing like subjects with like subjects. This strategy is similar to the adjustment of using blocks in the randomized block design.

- The CMH test works best when the X–Y association is similar in each partial table. It may not work well when the association varies dramatically among the partial tables.
  - Usually when there is an association, it is usually in the same direction across tables, although to a varying degree.
  - The CMH test has good power against the alternative of consistent patterns of association. It has low power for detecting association in opposite directions.
  - Regardless of power, the method always has the desired level of significance under H_0, so it is always a valid method.
  - Always examine the partial tables to determine if you have a situation in which the association is inconsistent and the CMH statistic is not very powerful.

- The CMH and Breslow-Day tests are large sample tests. Exact tests and confidence intervals are available in SAS using the `EXACT` statement with `PROC FREQ`.
  - The `COMOR` option provides confidence limits for the common OR and a test for a common OR.
  - The `EQOR` option provides a test for equal ORs.
  - The `MHCHI` provides the CMH test.

- There are extensions of the CMH test to $I \times J \times K$ tables. The appropriate procedure will depend on whether
  - both X and Y are nominal,
  - the responses are ordinal
  - both the responses and the groups are ordinal

See *Categorical Data Analysis Using the SAS System* for more details in Chapter 6 covering general $I \times J \times K$ tables.