prepared by Professor Jenny Baglivo

5 MATH448001 Notebook 5

This notebook considers methods for analyzing data structured as I-by-J tables of frequencies. The row and column designations in the tables correspond to levels of two factors, and analyses focus on relationships between these factors. The notes include topics from Chapter 13 (analysis of categorical data) of the Rice textbook.

5.1 Multinomial Distributions

A *multinomial experiment* is an experiment with exactly k outcomes. We often write the probabilities of these outcomes as a vector of proportions

$$(p_1, p_2, \ldots, p_k),$$

where p_i is the probability that the i^{th} outcome occurs. The outcomes of a multinomial experiment are often called "categories" or "groups".

Let X_i be the number of occurrences of the i^{th} outcome in n independent trials of the multinomial experiment. Then the random k-tuple of frequencies

$$\underline{X} = (X_1, X_2, \ldots, X_k)$$

is said to have a *multinomial distribution* with parameters n and $(p_1, p_2, \ldots, p_k)$.

The joint PDF of the random k-tuple is

$$p(x_1, x_2, \ldots, x_k) = \binom{n}{x_1, x_2, \ldots, x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

when the frequencies x_i satisfy $0 \le x_i \le n$ and $\sum_{i=1}^{k} x_i = n$; otherwise, the PDF equals zero.

Example. Consider a population of individuals who were involved in bicycle accidents in a given time period. The multinomial experiment consists of randomly choosing an individual from this population and determining if the individual suffered a head injury (event A) or not, and if the individual was wearing a helmet at the time of the accident (event B) or not.

The outcomes and probabilities are as follows:

Head Injury		*No Head Injury*	
Helmet	*No Helmet*	*Helmet*	*No Helmet*
$p_1 = P(A \cap B)$	$p_2 = P(A \cap B^c)$	$p_3 = P(A^c \cap B)$	$p_4 = P(A^c \cap B^c)$

If we choose a simple random sample of size n from this population, and if the total population size is large, then the multinomial distribution with parameters n and (p_1, p_2, p_3, p_4) can be used to model the frequencies of individuals in each group.

5.1.1 Maximum Likelihood Estimation

The likelihood function for a multinomial distribution is the joint PDF thought of as a function of the probability vector, $(p_1, p_2, \ldots, p_k)$. In the general case,

$$Lik(p_1, p_2, \ldots, p_k) = \binom{n}{X_1, X_2, \ldots, X_k} p_1^{X_1} p_2^{X_2} \cdots p_k^{X_k},$$

and the set of all possible parameters is

$$\Omega = \left\{ (p_1, p_2, \ldots, p_k) \mid p_i \in (0,1) \text{ for all } i, \ \sum_{i=1}^{k} p_i = 1 \right\}.$$

The following theorem gives the maximum likelihood (ML) estimators of the p_i's.

> ***Theorem (Parameter Estimation).*** Under the conditions above, the maximum likelihood estimators of the parameters of the multinomial model are
>
> $$\widehat{p_i} = \frac{X_i}{n}, \quad \text{for } i = 1, 2, \ldots, k.$$
>
> That is, the sample proportions are the ML estimators in the general case.
>
> [*Note:* The method of Lagrange multipliers makes this proof easy to do.]

Models with fewer parameters. In many situations, we are interested in models with fewer parameters. For example,

1. *Grouped Geometric Model:* Let $\theta \in (0,1)$, and let $k > 1$ be an integer. Then the probabilities the grouped geometric model with parameters θ and k has probabilities

   $$(p_1, p_2, \ldots, p_k) = ((1-\theta), \theta(1-\theta), \theta^2(1-\theta), \ldots, \theta^{k-2}(1-\theta), \theta^{k-1}).$$

2. *Independence Model:* Let $\alpha = P(A)$ and $\beta = P(B)$. Then the independence model for events A and B has probabilities

   $$(p_1, p_2, p_3, p_4) = (\alpha\beta, \alpha(1-\beta), (1-\alpha)\beta, (1-\alpha)(1-\beta)),$$

   where $p_1 = P(A \cap B)$, $p_2 = P(A \cap B^c)$, $p_3 = P(A^c \cap B)$, $p_4 = P(A^c \cap B^c)$.

Application (Survival Analysis):

A typical application of the grouped geometric model is to survival analysis. In this application, the parameter θ is the probability of surviving one time period, and independence is assumed between time periods. The probability of dying in the first time period is $p_1 = 1 - \theta$. The probability of surviving the first time period but dying in the second time period is $p_2 = \theta(1-\theta)$, and so forth.

The last outcome corresponds to surviving $k-1$ or more time periods.

Exercise. Assume that (X_1, X_2, X_3) has a multinomial distribution with parameters n and

$$(p_1, p_2, p_3) = ((1-\theta), \theta(1-\theta), \theta^2),$$

where θ is a proportion, $\theta \in (0, 1)$. Find the general form of the ML estimator of θ.

5.1.2 Likelihood Ratio Goodness-of-Fit

Likelihood ratio tests can be used to determine if a given family of models fits a given list of observed frequencies.

The usual setup is as follows:

1. *Parameter Space:*

   The parameter space, denoted by Ω, is the set of all possible parameters. For multinomial distributions, the parameter space is

   $$\Omega = \left\{ (p_1, p_2, \ldots, p_k) \mid p_i \in (0,1) \text{ for all } i, \ \sum_{i=1}^{k} p_i = 1 \right\}.$$

   There are $k-1$ free parameters in the parameter space.

2. *Null Space:*

   The null space, denoted by ω_o, is the subset of parameters corresponding to the model you would like to fit. For example,

   (a) The null space for the *grouped geometric model* is

   $$\omega_o = \left\{ ((1-\theta), \theta(1-\theta), \ldots, \theta^{k-2}(1-\theta), \theta^{k-1}) \mid \theta \in (0,1) \right\} \subset \Omega.$$

   There is 1 free parameter for this model.

   (b) The null space for the *independence model* is

   $$\omega_o = \{ (\alpha\beta, \alpha(1-\beta), (1-\alpha)\beta, (1-\alpha)(1-\beta)) \mid \alpha, \beta \in (0,1) \} \subset \Omega.$$

   There are two free parameters for this model.

3. *Null and Alternative Hypotheses:*

   The null and alternative hypotheses for a test to determine if the proposed model fits a given list of frequencies are set up as follows:

   $$H_0 : (p_1, \ldots, p_k) \in \omega_o \qquad \text{versus} \qquad H_A : (p_1, \ldots, p_k) \in \Omega \backslash \omega_o.$$

4. *Likelihood Ratio Test:*

The likelihood ratio test statistic, denoted by Λ, is the ratio of

- the maximum value of the likelihood function for all models satisfying H_0 to
- the maximum value of the likelihood function for all possible models.

A likelihood ratio test of size α is a test with decision rule

$$\text{Reject } H_0 \text{ in favor of } H_A \text{ when } \Lambda \leq c,$$

where c is chosen so that $P(\Lambda \leq c) = \alpha$ if the model family is correct.

In the multinomial setting,

(a) The numerator of the likelihood ratio statistic is

$$Lik(\widehat{p_1}, \widehat{p_2}, \ldots, \widehat{p_k}) = \binom{n}{X_1, \ldots, X_k} (\widehat{p_1})^{X_1} \cdots (\widehat{p_k})^{X_k},$$

where each $\widehat{p_i}$ is obtained by substituting the ML estimators for the unknown parameters under the null hypothesis, and

(b) The denominator of the likelihood ratio is statistic is

$$Lik(X_1/n, X_2/n, \ldots, X_k/n) = \binom{n}{X_1, \ldots, X_k} (X_1/n)^{X_1} \cdots (X_k/n)^{X_k},$$

using the ML estimators from the theorem stated earlier.

Analyses use the following transformed form of Λ:

$$-2\log(\Lambda) = \sum_{i=1}^{k} 2X_i \log\left(\frac{X_i}{n\widehat{p_i}}\right),$$

where log() is the natural logarithm function, and an equivalent decision rule:

$$\text{Reject } H_0 \text{ in favor of } H_A \text{ when } -2\log(\Lambda) \geq c^*,$$

where c^* is chosen so that $P(-2\log(\Lambda) \geq c^*) = \alpha$ if the model family is correct.

Exercise. Demonstrate that $-2\log(\Lambda)$ has the form given above for multinomial models.

Large sample theory. In the 1930's, Wilks proved that the statistic $-2\log(\Lambda)$ has an approximate chi-square distribution when n is large enough. Specifically,

> ***Wilks' Theorem.*** Under the conditions above, the goodness-of-fit test statistic
>
> $-2\log(\Lambda)$ has an approximate chi-square distribution with $(k-1-k_o)$ *df*,
>
> where k_o is the number of free parameters in ω_o, when n is large.

Using Wilks' theorem, the form of the decision rule for a $100\alpha\%$ large sample test is

$$\text{Reject } H_0 \text{ in favor of } H_A \text{ when } -2\log(\Lambda) \geq \chi^2_{k-1-k_o}(\alpha),$$

where the cutoff is the $100(1-\alpha)\%$ point of the chi-square distribution with $(k-1-k_o)$ *df*.

Rule of Thumb:
The approximation is adequate if each cell expectation is 5 or more.

Exercise from page 5, continued. Consider again the grouped geometric model with three categories, and assume that $(x_1, x_2, x_3) = (52, 19, 29)$ has been observed. Conduct a likelihood ratio goodness-of-fit test at the 5% significance level. Clearly state your conclusion.

5.1.3 Pearson's Goodness-of-Fit

Pearson's statistic is the most commonly used statistic for testing goodness-of-fit of multinomial models. The form of Pearson's statistic is as follows:

$$\boldsymbol{X^2} = \sum_{i=1}^{k} \frac{(X_i - n\widehat{p_i})^2}{n\widehat{p_i}}.$$

Here are some important facts about this statistic:

1. *Second Order Taylor Approximation:* Pearson's statistic is a second order Taylor approximation to the $-2\log(\Lambda)$ statistic. The approximation improves as n increases.

   Further, when n is large,

   $$-2\log(\Lambda) \approx \boldsymbol{X^2}$$

   and each statistic has an approximate chi-square distribution with $(k - 1 - k_o)$ *df.*

2. *Estimated Standardized Residuals:* For a given frequency list, $(x_1, x_2, \ldots, x_k)$, estimated standardized residuals can be defined as follows:

   $$r_i = \frac{x_i - n\widehat{p_i}}{\sqrt{n\widehat{p_i}}}, \quad i = 1, 2, \ldots, k.$$

   If the proposed model fits the data, then most r_i's will lie in the interval $(-2, 2)$.

   Cells with $|r_i| \geq 2$ deserve mention in your analysis.

Exercise, continued. Continuing with the data from the grouped geometric exercise on the previous page, find the standardized residuals and the observed value of Pearson's statistic.

5.1.4 Example: Independence of Events

Let A and B be events of interest, and let $\alpha = P(A)$ and $\beta = P(B)$. The joint distribution of the two events can be viewed as a 2-by-2 table of probabilities:

	Event B:	*Event B^c:*	
Event A:	$p_{1,1} = P(A \cap B)$	$p_{1,2} = P(A \cap B^c)$	α
Event A^c:	$p_{2,1} = P(A^c \cap B)$	$p_{2,2} = P(A^c \cap B^c)$	$1 - \alpha$
	β	$1 - \beta$	1

If events A and B are independent, the table of probabilities becomes

	Event B:	*Event B^c:*	
Event A:	$p_{1,1} = \alpha\beta$	$p_{1,2} = \alpha(1 - \beta)$	α
Event A^c:	$p_{2,1} = (1 - \alpha)\beta$	$p_{2,2} = (1 - \alpha)(1 - \beta)$	$1 - \alpha$
	β	$1 - \beta$	1

If $(X_{1,1}, X_{1,2}, X_{2,1}, X_{2,2})$ has a multinomial distribution with parameters n and $(p_{1,1}, p_{1,2}, p_{2,1}, p_{2,2})$, then we can examine the frequencies by examining a corresponding 2-by-2 table of frequencies:

	Event B:	*Event B^c:*	
Event A:	$X_{1,1}$	$X_{1,2}$	$X_{1,\bullet}$
Event A^c:	$X_{2,1}$	$X_{2,2}$	$X_{2,\bullet}$
	$X_{\bullet,1}$	$X_{\bullet,2}$	n

The following theorem gives the maximum likelihood estimators of α and β:

> ***Theorem (Parameter Estimation).*** If $(X_{1,1}, X_{1,2}, X_{2,1}, X_{2,2})$ satisfies the independence model above, then the ML estimators of α and β are as follows:
>
> $$\widehat{\alpha} = \frac{X_{1,\bullet}}{n} \quad \text{and} \quad \widehat{\beta} = \frac{X_{\bullet,1}}{n}.$$
>
> That is, the sample proportions for the marginal probabilities are the ML estimators of these probabilities.

Example (Source: Pagano & Gauvreau, 1993). The following 2-by-2 table displays the results of a study designed to study the relationship between wearing a bicycle helmet (event H) and sustaining a head injury (event I). The data consist of a simple random sample of 793 individuals who were involved in bicycle accidents during a specified one year period.

	Event H:	*Event H^c:*	
Event I:	17	218	235
Event I^c:	130	428	558
	147	646	793

Note that 29.6% (235/793) of the individuals sustained a head injury, 18.5% (147/793) wore a helmet, and 2.1% (17/793) both sustained a head injury and wore a helmet.

To determine if there is a relationship between head injury and helmet use, goodness-of-fit tests of the independence model will be conducted at the 5% significance level.

The *left* table below gives the estimated group means assuming the independence model holds, and the *right* table gives the estimated standardized residuals:

	Event H:	*Event H^c:*
Event I:	43.56	191.44
Event I^c:	103.44	454.56

	Event H:	*Event H^c:*
Event I:	−4.02	1.92
Event I^c:	2.61	−1.25

The following table summarizes the goodness-of-fit tests:

LR Statistic:	*Pearson's Statistic:*	*P Value based on Pearson's Statistic:*	*Sampling Distribution:*
32.5432	28.2555	≈ 0	Chi-Square, 1 *df*

These results suggest (please complete)

Footnote on Positive and Negative Association:

Assume that A and B are events with positive probability and that A and B are *not* independent. Then either

$$P(A \cap B) > P(A)P(B) \quad \text{or} \quad P(A \cap B) < P(A)P(B).$$

If the first case is true, we say that A and B are *positively associated*; otherwise, we say that A and B are *negatively associated*.

5.2 Independence Analysis for I-by-J Tables

This section considers methods for samples of size n, structured as tables with IJ *cells*:

	$j=1$	$j=2$	$\cdots$	$j=J$	
$i=1$	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,J}$	$x_{1,\bullet}$
$i=2$	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,J}$	$x_{2,\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i=I$	$x_{I,1}$	$x_{I,2}$	$\cdots$	$x_{I,J}$	$x_{I,\bullet}$
	$x_{\bullet,1}$	$x_{\bullet,2}$	$\cdots$	$x_{\bullet,J}$	n

For each (i, j), $x_{i,j}$ is the number of observations at

- the i^{th} level of the first factor (the *row factor*), and
- the j^{th} level of the second factor (the *column factor*).

Factors are considered to be unordered. For example, the levels of a factor could represent different political party affiliations or different professions.

5.2.1 Large Sample Goodness-of-Fit

The data are assumed to summarize the results of n independent trials of a multinomial experiment with IJ outcomes, and with probabilities

	$j=1$	$j=2$	$\cdots$	$j=J$	
$i=1$	$p_{1,1}$	$p_{1,2}$	$\cdots$	$p_{1,J}$	α_1
$i=2$	$p_{2,1}$	$p_{2,2}$	$\cdots$	$p_{2,J}$	α_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i=I$	$p_{I,1}$	$p_{I,2}$	$\cdots$	$p_{I,J}$	α_I
	β_1	β_2	$\cdots$	β_J	1

Of interest is a test of the null hypothesis that the row and column factors are independent.

Equivalently, we are interested in testing the null hypothesis that

$$p_{i,j} = \alpha_i \beta_j, \quad \text{for } i = 1, 2, \ldots, I, \ j = 1, 2, \ldots, J,$$

versus the general alternative that equality does not hold in at least one case.

The following theorem gives the ML estimators for the unknown parameters.

> ***Theorem (Parameter Estimation).*** Let $\underline{X}$ be a multinomial IJ-tuple structured as follows:
>
> | | $j=1$ | $j=2$ | $\cdots$ | $j=J$ | |
> |---|---|---|---|---|---|
> | $i=1$ | $X_{1,1}$ | $X_{1,2}$ | $\cdots$ | $X_{1,J}$ | $X_{1,\bullet}$ |
> | $i=2$ | $X_{2,1}$ | $X_{2,2}$ | $\cdots$ | $X_{2,J}$ | $X_{2,\bullet}$ |
> | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
> | $i=I$ | $X_{I,1}$ | $X_{I,2}$ | $\cdots$ | $X_{I,J}$ | $X_{I,\bullet}$ |
> | | $X_{\bullet,1}$ | $X_{\bullet,2}$ | $\cdots$ | $X_{\bullet,J}$ | n |
>
> If $\underline{X}$ satisfies the independence model above, then the maximum likelihood estimators of $\underline{\alpha}$ and $\underline{\beta}$ are as follows:
>
> $$\widehat{\alpha}_i = \frac{X_{i,\bullet}}{n} \quad \text{and} \quad \widehat{\beta}_j = \frac{X_{\bullet,j}}{n}$$
>
> for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. That is, the sample proportions for the marginal probabilities are the ML estimators of these probabilities.

Notes:

1. *Cell Means:*

   For each (i, j), the expected cell frequency is $E(X_{i,j}) = np_{i,j}$.

   Under the independence model, the ML estimator of $E(X_{i,j})$ is

   $$n\widehat{p}_{i,j} = n\widehat{\alpha}_i\widehat{\beta}_j = n\left(\frac{X_{i,\bullet}}{n}\right)\left(\frac{X_{\bullet,j}}{n}\right) = \frac{(X_{i,\bullet})(X_{\bullet,j})}{n} \quad \text{for all } i, \ j.$$

2. *Large Sample Analysis:*

   If cell expectations are 5 or more, then the chi-square approximation to the sampling distributions of the goodness-of-fit statistics is adequate. There are a total of

   $$(k-1) - k_o = (IJ-1) - ((I-1) + (J-1)) = (I-1)(J-1)$$

   degrees of freedom.

Example (Source: Rice textbook, Chapter 13). Overfield and Klauber (1980) published the following data on the incidence of tuberculosis in relation to blood types in a sample of Eskimos.

	O	*A*	*AB*	*B*	
Moderate-Advanced Disease (M-AD):	7	5	3	13	28
Minimal Disease (MinD):	27	32	9	18	86
Disease Not Present (NoD):	55	50	7	24	136
	89	87	19	55	250

The row factor corresponds to disease group (moderate to advanced, minimal, not present). The column factor corresponds to blood type in the *ABO system*.

Assume these data summarize the values of a simple random sample of 250 individuals from a large Eskimo population. To determine if there is a relationship between disease group and blood type using the ABO system, goodness-of-fit tests of independence will be conducted at the 5% significance level.

The *left* table below gives the estimated cell means assuming the independence model holds, and the *right* table gives the estimated standardized residuals:

	O	*A*	*AB*	*B*
M-AD:	9.97	9.74	2.13	6.16
MinD:	30.62	29.93	6.54	18.92
NoD:	48.42	47.33	10.34	29.92

	O	*A*	*AB*	*B*
M-AD:	-0.94	-1.52	0.60	2.76
MinD:	-0.65	0.38	0.96	-0.21
NoD:	0.95	0.39	-1.04	-1.08

The following table summarizes the goodness-of-fit tests:

LR Statistic:	*Pearson's Statistic:*	*P Value based on Pearson's Statistic:*	*Sampling Distribution:*
14.798	15.9842	0.0138	Chi-Square, 6 *df*

These results suggest (please complete)

Example (Source: Moore & McCabe, 1999). As part of a study on factors affecting early childhood development, information on daily alcohol use and on daily nicotine use was collected in a simple random sample of 452 young mothers:

	No Nicotine	*1-15 mg*	*16+ mg*	
No Alcohol:	105	7	11	123
0.01-0.10 oz:	58	5	13	76
0.11-0.99 oz:	84	37	42	163
1.00+ oz:	57	16	17	90
	304	65	83	452

The row factor corresponds to alcohol use prior to pregnancy (none, 0.01-0.10 ounces/day, 0.11-0.99 ounces/day, 1.00 or more ounces/day), and the column factor corresponds to nicotine intake during pregnancy (none, 1-15 milligrams/day, 16 or more milligrams/day).

To determine if there is a relationship between alcohol use prior to pregnancy and nicotine use during pregnancy, goodness-of-fit tests of the independence model will be conducted at the 5% significance level.

The *left* table below gives the estimated cell means assuming the independence model holds, and the *right* table gives the estimated standardized residuals:

	0 mg	*1-15 mg*	*16+ mg*
0 oz:	82.73	17.69	22.59
0.01-0.10 oz:	51.12	10.93	13.96
0.11-0.99 oz:	109.63	23.44	29.93
1.00+ oz:	60.53	12.94	16.53

	0 mg	*1-15 mg*	*16+ mg*
0 oz:	2.45	-2.54	-2.44
0.01-0.10 oz:	0.96	-1.79	-0.26
0.11-0.99 oz:	-2.45	2.8	2.21
1.00+ oz:	-0.45	0.85	0.12

The following table summarizes the goodness-of-fit tests:

LR Statistic:	*Pearson's Statistic:*	*P Value based on Pearson's Statistic:*	*Sampling Distribution:*
44.6527	42.2521	≈ 0	Chi-Square, 6 *df*

These results suggest (please complete)

5.2.2 Permutation Analysis: Goodness-of-Fit

When some cell expectations are small ($E(X_{i,j}) < 5$), permutation methods can be used to estimate p values for goodness-of-fit tests of the independence model.

The idea is to think of the observations as a list of n pairs

$$x_{1,1} \text{ copies of (1,1), followed by } x_{1,2} \text{ copies of (1,2), and so forth.}$$

The paired data are ordered by the levels of the row factor. Then,

1. Let $\underline{v}$ be the list of first coordinates and $\underline{w}$ be the list of second coordinates.
2. For each matching of a permutation of the $\underline{w}$ list to the ordered $\underline{v}$ list, an I-by-J table is constructed, and the value of the goodness-of-fit statistic is computed.
3. The permutation p value is the proportion of statistics that are greater than or equal to the observed value for the given sample.

To illustrate the permutation method, the 2-by-2 table shown on the *left* below leads to the $\underline{v}$ and $\underline{w}$ vectors on the *right*:

	$j=1$	$j=2$	
$i=1$	2	1	3
$i=2$	3	3	6
	5	4	9

$\Longrightarrow$

$\underline{v}$: 1, 1, 1, 2, 2, 2, 2, 2, 2

$\underline{w}$: 1, 1, 2, 1, 1, 1, 2, 2, 2

The random permutation of $\underline{w}$ shown on the *left* leads to the 2-by-2 table shown on the *right*:

$\underline{v}$: 1, 1, 1, 2, 2, 2, 2, 2, 2

$\underline{w^*}$: 2, 1, 2, 2, 1, 1, 1, 1, 2

$\Longrightarrow$

	$j=1$	$j=2$	
$i=1$	1	2	3
$i=2$	4	2	6
	5	4	9

The margins of the table are fixed in the permutation process.

Permutation goodness-of-fit tests are appropriate under both population and randomization models. The null hypothesis is that any observed association between the factors is due to chance alone versus the general alternative that the association is due to something other than chance.

Monte Carlo analysis is used to estimate p values in most situations.

Example from page 15, continued. Consider again the data from the study on the relationship between tuberculosis disease group and blood type using the ABO system. Since one of the estimated cell means was less than 5, a permutation test will be conducted at the 5% significance level using Pearson's statistic.

The following table summarizes the results of a Monte Carlo analysis using 5000 random permutations (including the observed matching of disease group to blood type):

Pearson's Statistic	*Estimated p-Value*	*99% CI for p-Value*
15.9842	0.0128 (64/5000)	[0.0091, 0.0175]

Our conclusions are the same as before.

Additional analysis: The researchers also classified the 250 individuals by disease group (row factor) and blood type using the *MN system* (column factor):

	MM	*MN*	*NN*	
Moderate-Advanced Disease (M-AD):	21	6	1	28
Minimal Disease (MinD):	54	27	5	86
Disease Not Present (NoD):	74	51	11	136
	149	84	17	250

Consider testing goodness-of-fit to the independence model using Pearson's statistic and the 5% significance level. The observed value of the statistic is 4.73397.

The *left* table below gives the estimated cell means assuming the independence model holds, and the *right* table gives the estimated standardized residuals:

	MM	*MN*	*NN*
M-AD:	16.69	9.41	1.90
MinD:	51.26	28.90	5.85
NoD:	81.06	45.70	9.25

	MM	*MN*	*NN*
M-AD:	1.05	-1.11	-0.65
MinD:	0.38	-0.35	-0.35
NoD:	-0.78	0.78	0.58

Since there is one very small estimated cell mean, permutation methods will be used to find the p value. The following table summarizes the results of a Monte Carlo analysis using 5000 random permutations (including the observed matching of disease group to blood type):

Pearson's Statistic	*Estimated p-Value*	*99% CI for p-Value*
4.73397	0.3092 (1546/5000)	[0.2925, 0.3263]

These results suggest that disease group is not associated with blood type using the MN system.

5.2.3 Ordered Categories: Rank Correlation Test

In some studies, the levels of the row and column factors have a natural ordering. For example, in the alcohol-nicotine study (page 16), the levels of the row factor correspond to increasing use of alcohol and the levels of the column factor correspond to increasing use of nicotine.

If the levels of the row and column factors are ordered, then the table can be analyzed using Spearman's rank correlation statistic, R_s.

1. For the first factor, $x_{1\bullet}$ observations are assumed to be tied at the lowest level, $x_{2\bullet}$ at the next level, and so forth.
2. Similarly, for the second factor, $x_{\bullet 1}$ observations are assumed to be tied at the lowest level, $x_{\bullet 2}$ at the next level, and so forth.
3. R_s is the sample correlation of the paired ranks.

For example, consider again the 2-by-2 table from the last section:

	$j=1$	$j=2$	
$i=1$	2	1	3
$i=2$	3	3	6
	5	4	9

$\Longrightarrow$

$\underline{v}$: 1, 1, 1, 2, 2, 2, 2, 2, 2

$\underline{w}$: 1, 1, 2, 1, 1, 1, 2, 2, 2

We no longer consider the elements of $\underline{v}$ and $\underline{w}$ as placeholders for the (unordered) levels of the two factors. Instead, we convert the numbers to ranks for further analysis:

v	w	r_v	r_w
1	1	2.0	3.0
1	1	2.0	3.0
1	2	2.0	7.5
2	1	6.5	3.0
2	1	6.5	3.0
2	1	6.5	3.0
2	2	6.5	7.5
2	2	6.5	7.5
2	2	6.5	7.5

The rank correlation is 0.158114.

Rank correlation tests are appropriate under both population and randomization models. The null hypothesis is that any observed association between the factors is due to chance alone versus alternatives that the factors under study are positively or negatively associated. Tests are usually conducted using the large sample normal approximation to the permutation distribution of R_s.

Example from page 16, continued. Consider again the data from the alcohol-nicotine use study. The following table summarizes the results of a two-sided test of the null hypothesis that the correlation between the factors is zero:

Spearman Statistic:	*Two-Sided P Value:*	*Sampling Distribution:*
0.218215	≈ 0	Normal ($\mu = 0$, $\sigma = 0.047$)

Since the p value is virtually zero and the observed rank correlation is positive, the results suggest a positive association between the alcohol intake before pregnancy and nicotine intake during pregnancy.

5.3 Homogeneity Analysis for I-by-J Tables

This section considers methods for I samples of sizes n_i, for $i = 1, 2, \ldots, I$.

The samples are structured as I-by-J tables with fixed row sums:

	$j=1$	$j=2$	$\cdots$	$j=J$	
$i=1$	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,J}$	n_1
$i=2$	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,J}$	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i=I$	$x_{I,1}$	$x_{I,2}$	$\cdots$	$x_{I,J}$	n_I
	$x_{\bullet,1}$	$x_{\bullet,2}$	$\cdots$	$x_{\bullet,J}$	n

In this table,

1. The total number of observations is $n = \sum_{i=1}^{I} n_i$,

2. The I samples correspond to levels of the first factor (the *row factor*), and

3. The value $x_{i,j}$ is the number of observations in the i^{th} sample at the j^{th} level of the second factor (the *column factor*).

General examples include

- Comparing blood type distributions in different ethnic communities,
- Comparing distributions of professions in different urban communities.

5.3.1 Maximum Likelihood Estimation

The data are assumed to summarize I independent random samples from multinomial distributions with J outcomes, and whose probabilities are as follows:

	$j=1$	$j=2$	$\cdots$	$j=J$	
$i=1$	$p_{1,1}$	$p_{1,2}$	$\cdots$	$p_{1,J}$	1
$i=2$	$p_{2,1}$	$p_{2,2}$	$\cdots$	$p_{2,J}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i=I$	$p_{I,1}$	$p_{I,2}$	$\cdots$	$p_{I,J}$	1

Of interest is a test of the null hypothesis that the I row models are equal.

The null hypothesis can be written as

$$p_{1,j} = p_{2,j} = \cdots = p_{I,j} \text{ for } j = 1, 2, \ldots, J.$$

The general alternative hypothesis of interest is that, for at least one level of the second factor, some probabilities differ.

Product likelihood function. To analyze the homogeneity model, we take the product of the likelihoods for the I samples.

1. *Sample i:*

   Let $\underline{X}_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,J})$ be the multinomial random vector for the i^{th} row, and

   $$Lik_i(p_{i,1}, p_{i,2}, \ldots, p_{i,J}) = \binom{n_i}{X_{i,1}, X_{i,2}, \ldots, X_{i,J}} p_{i,1}^{X_{i,1}} p_{i,2}^{X_{i,2}} \cdots p_{i,J}^{X_{i,J}}$$

   be the likelihood function for the i^{th} row.[1]

[1] If there were no restrictions on the parameters for the i^{th} model, then the parameter estimation theorem from page 4 implies that the ML estimators are the sample proportions $\widehat{p}_{i,j} = (X_{i,j}/n_i)$.

2. *Product likelihood:*

   Now, let $(p_1, p_2, \ldots, p_J)$ be the model probabilities under the null hypothesis of homogeneity of row models. By independence, the likelihood function is the product of the likelihood functions for the I row models, using the common proportions for each row:

$$Lik(p_1, p_2, \ldots, p_J) = \prod_{i=1}^{I} Lik_i(p_1, p_2, \ldots, p_J).$$

The following theorem gives the ML estimators of the parameters.

Theorem (Parameter Estimation). Let $\underline{X}_i$ be a multinomial model with J categories, for $i = 1, 2, \ldots, I$, where the IJ-frequencies are structured as follows:

	$j=1$	$j=2$	$\cdots$	$j=J$	
$i=1$	$X_{1,1}$	$X_{1,2}$	$\cdots$	$X_{1,J}$	n_1
$i=2$	$X_{2,1}$	$X_{2,2}$	$\cdots$	$X_{2,J}$	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i=I$	$X_{I,1}$	$X_{I,2}$	$\cdots$	$X_{I,J}$	n_i
	$X_{\bullet,1}$	$X_{\bullet,2}$	$\cdots$	$X_{\bullet,J}$	n

   If the collection $\{\underline{X}_i\}$ satisfies the homogeneity model above, then the maximum likelihood estimators of the common probabilities $\underline{p}$ are as follows:

$$\widehat{p}_j = \frac{X_{\bullet,j}}{n}$$

   for $j = 1, 2, \ldots, J$. That is, the sample proportions for the marginal probabilities are the ML estimators of the common probabilities.

Notes:

1. *Combining Models:*

   If the homogeneity model holds, then we can combine the frequencies in each category to produce a single multinomial model with n observations.

2. *Cell Means:*

   For each (i, j), the expected cell frequency is $E(X_{i,j}) = n_i p_{i,j}$.

   Under the homogeneity model, the ML estimator of $E(X_{i,j})$ is

$$n_i \widehat{p}_{i,j} = n_i \widehat{p}_j = n_i \left(\frac{X_{\bullet,j}}{n} \right) = \frac{n_i (X_{\bullet,j})}{n} \quad \text{for all } i,\ j.$$

3. *Free Parameters:*

   If there are no restrictions on the $p_{i,j}$'s, then there are

   _____ free parameters.

   If the homogeneity model holds, then there are

   _____ free parameters.

   And, the difference between these values is _____.

4. *Likelihood Ratio Statistic:*

   By independence, the likelihood ratio statistic is the product of likelihood ratio statistics for the I models:

$$\Lambda = \prod_{i=1}^{I} \Lambda_i = \prod_{i=1}^{I} \frac{Lik_i(\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_J)}{Lik_i(X_{i,1}/n_i, X_{i,2}/n_i, \ldots, X_{i,J}/n_i)},$$

   and the equivalent form we use for multinomial models is

$$-2\log(\Lambda) = \sum_{i=1}^{I} -2\log(\Lambda_i) = \sum_{i=1}^{I}\sum_{j=1}^{J} 2X_{i,j}\log\left(\frac{X_{i,j}}{n_i\widehat{p}_j}\right).$$

5.3.2 Large Sample Goodness-of-Fit

Wilks' theorem implies that $-2\log(\Lambda)$ has an approximate chi-square distribution when the totals n_i are all large enough. There are $(I-1)(J-1)$ degrees of freedom.

Using Wilks' theorem, the form of the decision rule for a $100\alpha\%$ large sample test is

$$\text{Reject } H_0 \text{ in favor of } H_A \text{ when } -2\log(\Lambda) \geq \chi^2_{(I-1)(J-1)}(\alpha),$$

where

- the null hypothesis is that the homogeneity model holds,
- the alternative hypothesis is the models do not satisfy homogeneity,
- log() is the natural logarithm function, and
- the cutoff is the $100(1-\alpha)\%$ point of the chi-square distribution with $(I-1)(J-1)$ *df.*

Notes:

1. *Rule of Thumb:*

   The chi-square approximation is adequate when each cell expectation is 5 or more.

2. *Comparison with Independence Model:*

   The form of the likelihood ratio test for homogeneity models is similar to the form for independence models. If we compare their estimated cell means, we see that

$$\widehat{E}(X_{i,j}) = n\ \widehat{\alpha_i}\ \widehat{\beta_j} = \frac{(X_{i,\bullet})(X_{\bullet,j})}{n} \quad \text{for independence models and}$$

$$\widehat{E}(X_{i,j}) = n_i \widehat{p_j} = \frac{(n_i)(X_{\bullet,j})}{n} \quad \text{for homogeneity models.}$$

   (In the independence model the row sums are random variables, while in the homogeneity model the row sums are fixed.) In addition, the degrees of freedom are the same for both types of analyses.

   Thus, although the sampling models are different in the two cases, and the interpretations are different, we can use the same basic program to analyze resulting tables.

Pearson's statistic; standardized residuals. As before, Pearson's statistic,

$$\boldsymbol{X^2} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(X_{i,j} - n_i\widehat{p}_j)^2}{n_i\widehat{p}_j},$$

is the second-order Taylor approximation to the likelihood ratio statistic.

Estimated standardized residuals,

$$r_{i,j} = \frac{x_{i,j} - n_i\widehat{p}_j}{\sqrt{n_i\widehat{p}_j}}$$

are used to help diagnose the fit of a given collection of frequencies to the homogeneity model.

Example (Source: Rice textbook, Chapter 13). "When Jane Austen died, she left the novel *Sanditon* only partially completed, but she left a summary of the remainder. A highly literate admirer finished the novel, attempting to emulate Austen's style, and the hybrid was published.

"[In a 1978 study,] Morton counted the occurrences of various words in several works:

1. Chapters 1 and 3 of *Sense and Sensibility*;
2. Chapters 1, 2, and 3 of *Emma*;
3. Chapters 1 and 6 of *Sanditon* (written by Austen); and
4. Chapters 12 and 24 of *Sanditon* (written by the admirer).

The counts Morton obtained for six words are given in the following table:"

	a	*an*	*this*	*that*	*with*	*without*	
1. Sense & Sensibility:	147	25	32	94	59	18	375
2. Emma:	186	26	39	105	74	10	440
3. Sanditon I:	101	11	15	37	28	10	202
4: Sanditon II:	83	29	15	22	43	4	196

A homogeneity analysis will be conducted at the 5% significance level to determine if Jane Austen was consistent in her use of these six common words.

If Jane Austen was consistent in her use of these words, then a second analysis will be conducted, at the 5% significance level, to determine if her admirer was successful in emulating Austen's use of the words.

Analysis I: To determine if Jane Austen was consistent in her use of the six words, we will test the null hypothesis of homogeneity of row models at the 5% significance level using the following 3-by-6 contingency table:

	a	*an*	*this*	*that*	*with*	*without*	
1. S & S:	147	25	32	94	59	18	375
	(160.03)	*(22.86)*	*(31.71)*	*(87.02)*	*(59.37)*	*(14.01)*	
2. Emma:	186	26	39	105	74	10	440
	(187.77)	*(26.82)*	*(37.21)*	*(102.10)*	*(69.66)*	*(16.44)*	
3. Sanditon I:	101	11	15	37	28	10	202
	(86.20)	*(12.31)*	*(17.08)*	*(46.88)*	*(31.98)*	*(7.55)*	
	434	62	86	236	161	38	1017

Estimated cell means for the homogeneity model are given in parentheses in the table above, and the table below gives the estimated standardized residuals:

	a	*an*	*this*	*that*	*with*	*without*
1. S & S:	-1.03	0.45	0.05	0.75	-0.05	1.07
2. Emma:	-0.13	-0.16	0.29	0.29	0.52	-1.59
3. Sanditon I:	1.59	-0.37	-0.50	-1.44	-0.70	0.89

The following table summarizes the goodness-of-fit tests:

LR Statistic:	*Pearson's Statistic:*	*P Value based on Pearson's Statistic:*	*Sampling Distribution:*
12.5873	12.2714	0.2670	Chi-Square, 10 *df*

The results suggest (please complete)

Analysis II: To determine if the admirer was successful in emulating Austen's use of these six words, we will test the null hypothesis of homogeneity of row models at the 5% significance level using the combined counts from the works written by Austen (row 1) and the counts from the work written by her admirer (row 2):

	a	*an*	*this*	*that*	*with*	*without*	
Austen:	434	62	86	236	161	38	1017
	(433.46)	*(76.30)*	*(84.68)*	*(216.31)*	*(171.04)*	*(35.21)*	
Admirer:	83	29	15	22	43	4	196
	(83.54)	*(14.70)*	*(16.32)*	*(41.69)*	*(32.96)*	*(6.79)*	
	517	91	101	258	204	42	1213

Estimated cell means for the homogeneity model are given in parentheses in the table above, and the table below gives the estimated standardized residuals:

	a	*an*	*this*	*that*	*with*	*without*
Austen:	0.03	-1.64	0.14	1.34	-0.77	0.47
Admirer:	-0.06	3.73	-0.33	-3.05	1.75	-1.07

The following table summarizes the goodness-of-fit tests:

LR Statistic:	*Pearson's Statistic:*	*P Value based on Pearson's Statistic:*	*Sampling Distribution:*
31.7368	32.8096	≈ 0	Chi-Square, 5 *df*

The results suggest (please complete)

5.3.3 Permutation Analysis: Goodness-of Fit

When some cell expectations are small ($E(X_{i,j}) < 5$), permutation methods can be used to estimate p values for goodness-of-fit tests of the homogeneity model.

The idea is to think of the i^{th} sample as a list of n_i values:

$$x_{i,1} \text{ copies of 1, followed by } x_{i,2} \text{ copies of 2, and so forth.}$$

Then,

1. For each partition of the I samples into distinguishable groups of sizes n_1, n_2, ..., n_I, an I-by-J table is formed, and the value of the goodness-of-fit statistic is computed.
2. The permutation p value is the proportion of statistics that are greater than or equal to the observed value for the given sample.

To illustrate the permutation method, suppose that there are three samples of sizes 4, 6, 3, and that there are four possible responses. Then, the 3-by-4 table shown on the *left* below would lead to the three samples shown on the *right*:

	$j=1$	$j=2$	$j=3$	$j=4$	
$i=1$	2	0	1	1	4
$i=2$	1	2	1	2	6
$i=3$	0	2	0	1	3
	3	4	2	4	13

$\Longrightarrow$

	Samples:
$i=1$	{1,1,3,4}
$i=2$	{1,2,2,3,4,4}
$i=3$	{2,2,4}

The random partition shown on the *left* below would lead to the table shown on the *right*:

	Random Partition:
$i=1$	{3,4,4,1}
$i=2$	{4,2,2,2,4,1}
$i=3$	{1,3,2}

$\Longrightarrow$

	$j=1$	$j=2$	$j=3$	$j=4$	
$i=1$	1	0	1	2	4
$i=2$	1	3	0	2	6
$i=3$	1	1	1	0	3
	3	4	2	4	13

The margins of the table are fixed in the permutation process.

Permutation goodness-of-fit tests are appropriate under both population and randomization models. The general null hypothesis is that observed differences in the I samples are due to chance alone versus the alternative that observed differences are due to something other than chance.

Monte Carlo analysis is used to estimate p values in most situations.

Example (Source: Fienberg, 1980). As part of a study to determine if the food additive known as *Red Dye #2* was a carcinogen, 88 rats were randomly assigned to two different dosage groups: 44 were fed a low dosage of the food additive and 44 were fed a high dosage.

The *left* table below corresponds to the 53 animals who died before the end of the study period, and the *right* table to the remaining 35 animals who were sacrificed at the end of the study.

	Present:	*Absent:*
Low Dose:	4	26
High Dose:	7	16

	Present:	*Absent:*
Low Dose:	0	14
High Dose:	7	14

In each table, the levels of the row factor correspond to dosage level (low dosage, high dosage), and the levels of the column factor correspond to the presence or absence of tumors.

A natural first step is to determine if these two tables can be combined.

Analysis I. One way to determine if the information in the tables can be combined is to conduct a test of the homogeneity of row models in the following 2-by-4 table using Pearson's test and the 5% significance level.

	LD-P	*LD-A*	*HD-P*	*HD-A*	
Left Table:	4	26	7	16	53
	(2.41)	*(24.09)*	*(8.43)*	*(18.07)*	
Right Table:	0	14	7	14	35
	(1.59)	*(15.91)*	*(5.57)*	*(11.93)*	
	4	40	14	30	88

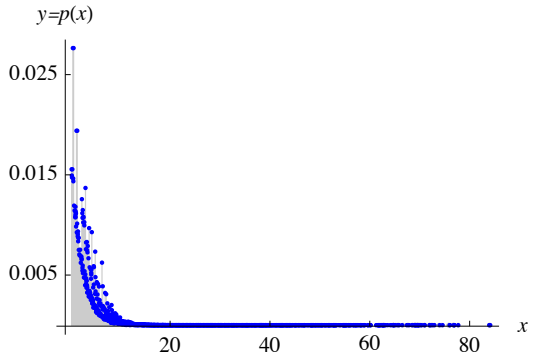
In this table, rows correspond to the left and right tables above and the numbers in parentheses are the estimated cell means under the homogeneity model.

Since the estimated cell means in the first column are quite small, a permutation test was conducted at the 5% significance level using Pearson's statistic.

The observed value of Pearson's statistic is 4.23. Since the permutation p value is

$$P(\boldsymbol{X^2} \geq 4.23) \approx 0.242,$$

the results suggest that the tables can be combined in order to analyze the relationship between dosage level and presence of tumors.



Analysis II. Working with the combined data,

	Present:	*Absent:*	
Low Dose:	4	40	44
High Dose:	14	30	44
	18	70	88

we next concentrate on the relationship between dosage level and the presence or absence of tumors, using the 5% significance level. Notice that 9.1% (4/44) of the rats in the low dosage group developed tumors, compared to 35.0% (14/44) in the high dosage group.

The *left* table gives the estimated group means assuming the homogeneity model holds, and the *right* table gives the estimated standardized residuals.

	Present:	*Absent:*
Low Dose:	9.00	35.00
High Dose:	9.00	35.00

	Present:	*Absent:*
Low Dose:	-1.67	0.85
High Dose:	1.67	-0.85

The following table summarizes the goodness-of-fit tests:

LR Statistic:	*Pearson's Statistic:*	*P Value based on Pearson's Statistic:*	*Sampling Distribution:*
7.31735	6.98413	0.00822	Chi-Square, 1 *df*

These results suggest (please complete)

5.3.4 Ordered Categories: Kruskal-Wallis Test

In some studies, the levels of the column factor have a natural ordering. If the levels of the column factor are ordered, then the table can be analyzed using the Kruskal-Wallis statistic. In the test,

- $x_{\bullet 1}$ observations are tied at the lowest level of the second factor,
- $x_{\bullet 2}$ observations are tied at the next lowest level,

and so forth.

For example, consider again the 3-by-4 table from the last section:

	$j=1$	$j=2$	$j=3$	$j=4$	
$i=1$	2	0	1	1	4
$i=2$	1	2	1	2	6
$i=3$	0	2	0	1	3
	3	4	2	4	13

$\Longrightarrow$

	Samples:
$i=1$	{1,1,3,4}
$i=2$	{1,2,2,3,4,4}
$i=3$	{2,2,4}

We no longer consider the elements of the 3 samples as placeholders for the (unordered) levels of the second factor. Instead, we rank all numbers for further analysis.

	Ranked Samples:
$i=1$	{2.0, 2.0, 8.5, 11.5}
$i=2$	{2.0, 5.5, 5.5, 8.5, 11.5, 11.5}
$i=3$	{5.5, 5.5, 11.5}

The observed value of the Kruskal-Wallis statistic is 0.41.

Kruskal-Wallis tests are appropriate under both population and randomization models. The null hypothesis is that observed differences in the I samples are due to chance alone versus the alternative that, for at least two samples, the values in one sample tend to be larger or smaller than the values in some other sample.

Example (Source: Hand, et al, 1993). As part of a study designed to compare treatments for small cell lung cancer, a simple random sample of 253 male patients were randomized to one of two treatments: either the same combination of drugs was given at regular intervals during the study period, or two different combinations of drugs were alternately given during the period. There were four levels of response:

(1) the tumor increased in size (called *disease progression*),
(2) there was no change in the tumor,
(3) the tumor decreased in size (called *partial remission*), and
(4) the tumor was not detectable (called *complete remission*).

The following table summarizes the results:

	Progression	*No Change*	*Partial Remission*	*Complete Remission*	
Same Combination:	28	45	29	26	128
Alternating Combinations:	41	44	20	20	125

Since about 43% (55/128) of the men in the first treatment group had partial or complete remission, compared to about 32% (40/125) in the second treatment group, the first treatment strategy appears to be better.

Analysis I: We first analyze the data using goodness-of-fit tests of the homogeneity model and the 5% significance level.

Estimated cell means (*left*) and standardized residuals (*right*) are shown below:

	P	*NC*	*PR*	*CR*
SC:	34.91	45.03	24.79	23.27
AC:	34.09	43.97	24.21	22.73

	P	*NC*	*PR*	*CR*
SC:	-1.17	0.00	0.85	0.57
AC:	1.18	0.00	-0.86	-0.57

The following table summarizes the goodness-of-fit tests:

LR Statistic:	*Pearson's Statistic:*	*P Value based on Pearson's Statistic:*	*Sampling Distribution:*
4.88696	4.86129	0.1820	Chi-Square, 3 *df*

Since the p value is greater than 0.05, there is insufficient evidence to conclude that the treatment protocols differ.

Analysis II: We next analyze the data using the Kruskal Wallis test and the 5% significance level. The following table summarizes the results of the test

KW Statistic:	*P Value:*	*Sampling Distribution:*
4.26064	0.039	Chi-Square, 1 *df*

The results suggest (please complete):