STAT 518 --- Chapter 4 --- Contingency Tables

• Contingency tables are summaries (in matrix form) of categorical data, where the entries in the table are counts of how many observations fell into specific categories (and combinations of categories).

• A one-way contingency table summarizes data on a single categorical variable and has only one row.

• A two-way contingency table summarizes data on two categorical variables and may have several rows and several columns.

• Data on several categorical variables can be summarized by multi-way contingency tables.

• We begin with another goodness-of-fit test.

Section 4.5: Chi-Squared Goodness-of-Fit Test

• Suppose we have a single categorical variable with c categories. The cell counts can be arranged in a one-way table.

Example 1: 95 adults were randomly sampled and surveyed about their favorite sport. There were 6 categories. Their preferences are summarized:

Favorite Sport

Football	Baseball	Basketball	Auto	Golf	Other	N
37	12	17	8	5	16	95

p_1 = proportion of U.S. adults favoring football
p_2 = proportion of U.S. adults favoring baseball
p_3 = proportion of U.S. adults favoring basketball
p_4 = proportion of U.S. adults favoring auto racing
p_5 = proportion of U.S. adults favoring golf
p_6 = proportion of U.S. adults favoring "other"

- **It was hypothesized that the true proportions are $(p_1, p_2, p_3, p_4, p_5, p_6) = (.4, .1, .2, .06, .06, .18)$.**

- **We test our null hypothesis with the chi-squared goodness-of-fit test:**

H_0: $P(\text{class } j) = p_j^*$ for $j = 1, \ldots, c$
H_1: at least one of the hypothesized probabilities is wrong

The test statistic is:

$$T = \sum_{j=1}^{c} \frac{(O_j - E_j)^2}{E_j} = \left(\sum_{j=1}^{c} \frac{O_j^2}{E_j} \right) - N$$

where O_j is the observed "cell count" for category j and E_j is the expected cell count for category j if H₀ true **.**

- **Under H_0, T has an asymptotic χ^2 distribution with $c - 1$ d.f.**

Decision Rule: Reject H_0 if $T > \chi^2_{1-\alpha, c-1}$

(large values of $T \rightarrow$ observed counts are very different from the expected counts under H_0)

<u>Assumptions</u>: (1) The data are at least <u>nominal</u>.
(2) The random sample is sufficiently large. Koehler and Larntz's Rule of Thumb: Test is valid if

$N \geq 10$, $c \geq 3$, $\frac{N^2}{c} \geq 10$ and all $E_j \geq 0.25$

- **If H_0 is true, expected cell count E_j =** $p_j^* N$

Example 1 data:

j	1	2	3	4	5	6	
O_j	37	12	17	8	5	16	$\rightarrow N = 95$
E_j	38	9.5	19	5.7	5.7	17.1	

Test statistic value:

$$T = \frac{37^2}{38} + \frac{12^2}{9.5} + \frac{17^2}{19} + \frac{8^2}{5.7} + \frac{5^2}{5.7} + \frac{16^2}{17.1} - 95 = 1.98$$

Decision Rule:

Reject H_0 if $T > \chi^2_{.95,5} = 11.07$ ← Table A2

P-value $\approx .852$ from R.

Conclusion: Since $T \not> 11.07$, fail to reject H_0. The hypothesized probability distribution for the sports is reasonable.

- **See `chisq.test` function in R to perform this test.**

Chi-Squared Test with Unknown Parameters

• **If our null hypothesis specifies the distribution <u>except</u> for a certain number (say, *k*) of unknown parameters, we can adjust the chi-squared test to account for this.**

• **The main difference is that when *k* unknown parameters are estimated from the data, the asymptotic null distribution of *T* is χ^2 with** $c-1-k$ **d.f.**

• **The unknown parameters must be estimated using "good methods" (see pp. 243-245): Typically the method of moments or maximum likelihood estimators work well.**

Example 2: Page 244 lists data for the number of hits of 18 baseball players in their first 45 times at bat. Is it reasonable that these data all follow the same binomial distribution with *n* = 45 and some unspecified *p*?

• **To estimate the unknown *p*, we use the estimate:**

$$\hat{p} = \frac{\text{total number of hits}}{\text{total number of at-bats}} = \frac{\sum_{i=1}^{18} X_i}{(18)(45)} = 0.2654$$

• **The expected cell counts can be found by the formula:**

$$E_j = \underset{\substack{\uparrow \\ \text{N, the number of players in the sample}}}{18}\; \underbrace{P(X=j)}_{\text{based on Binom}(45,\,0.2654)\text{ distribution}} \qquad \text{for } j = 0, 1, 2, \ldots, 45$$

• **Note that some E_j are very small; to alleviate this we should combine cells:**

j	≤ 7	8	9	10	11	12	13	14	15	16	17	≥ 18
O_j	1	1	1	5	2	1	1	2	1	1	1	1
E_j	1.10	1.06	1.57	2.04	2.35	2.40	2.20	1.82	1.36	0.92	0.57	0.61

Test statistic value:

$$T = \left(\sum \frac{O_j^2}{E_j} \right) - N = \frac{1^2}{1.10} + \frac{1^2}{1.06} + \cdots + \frac{1^2}{0.61} - 18 = 6.73$$

Decision Rule: $c = 12 \Rightarrow c - 1 - k = 10$

Reject H_0 if $T > \chi^2_{.95,\,10} = 18.31$

P-value ≈ 0.75 from R.

Conclusion: Since $6.73 \not> 18.31$, we fail to reject H_0. The binomial distribution provides a reasonable fit for these data.

• **While contingency tables describe <u>discrete</u> data, the chi-squared test can be used to check goodness of fit for <u>continuous</u> models as well.**

• **In that case, the continuous data must be discretized by grouping into intervals.**

• **How to form the intervals is somewhat arbitrary.**

Example 1 from Section 6.2: The data on page 445 consist of 50 observations. At $\alpha = 0.05$, is it reasonable to claim that the data follow a normal distribution?

We first estimate the two unknown parameters (μ and σ) of the normal distribution:

$\hat{\mu} = \bar{X} = 55.04$ $\quad\quad \hat{\sigma} = s = 19.00$

Let's choose 5 intervals:

Interval	[0, 20)	[20, 40)	[40, 60)	[60, 80)	[80, 100]
O_j	0	12	18	15	5
E_j	1.629	9.086	19.434	15.127	4.724

Test statistic value:

$$T = \frac{0^2}{1.629} + \frac{12^2}{9.086} + \frac{18^2}{19.434} + \frac{15^2}{15.127} + \frac{5^2}{4.724} - 50$$
$$= 2.69$$

Decision Rule: $c = 5 \Rightarrow c - 1 - k = 2$

Reject Ho if $T > \chi^2_{.95,2} = 5.991 \leftarrow$ Table A2

P-value ≈ 0.261 from R.

Conclusion: Since $2.69 \not> 5.991$, we fail to reject Ho. The normal distribution provides a reasonable fit for these data.

Section 4.1: Tests for 2 × 2 Tables

- **Consider the simplest form of two-way table:**

2 × 2 table (2 rows, 2 columns)

- **Such a table could summarize data arising from**

  - **Having a single sample in which** two binary variables **are measured on each individual**
  - **Having two samples in which** the same binary variable **is measured on each individual in each sample.**

Comparing Two Probabilities, Independent Samples

- **Suppose we have two independent samples, with respective sizes n_1 and n_2. We classify each individual in each sample into class 1 or class 2.**

- **Our data could be arranged in a 2 × 2 table as follows:**

	Class 1	Class 2	
Sample from Population 1	O_{11}	O_{12}	n_1
Sample from Population 2	O_{21}	O_{22}	n_2
	C_1	C_2	N

- **The total number of observations is $N = n_1 + n_2$.**

• **Our goal is to compare the probability of "success" (Class 1) across the two populations:**

p_1 = probability an observation from population 1 will be in class 1

p_2 = probability an observation from population 2 will be in class 1

Hypotheses:

Two-Tailed	Lower-Tailed	Upper-Tailed
$H_0: p_1 = p_2$	$H_0: p_1 \geq p_2$	$H_0: p_1 \leq p_2$
$H_1: p_1 \neq p_2$	$H_1: p_1 < p_2$	$H_1: p_1 > p_2$

Development of the Test Statistic

As estimators of p_1 and p_2, we have: $\hat{p}_1 = \frac{O_{11}}{n_1}$ and $\hat{p}_2 = \frac{O_{21}}{n_2}$

$$\hat{p}_1 - \hat{p}_2 = \frac{O_{11}}{n_1} - \frac{O_{21}}{n_2} = \frac{O_{11} n_2 - O_{21} n_1}{n_1 n_2}$$

$$= \frac{O_{11}(O_{21} + O_{22}) - O_{21}(O_{11} + O_{12})}{n_1 n_2}$$

$$= \frac{O_{11}O_{21} + O_{11}O_{22} - O_{21}O_{11} - O_{21}O_{12}}{n_1 n_2} = \frac{O_{11}O_{22} - O_{12}O_{21}}{n_1 n_2}$$

• **This estimates how far apart p_1 and p_2 are.**

• **Scaling this by dividing by the estimated standard error (see Eq. 5, p. 187), we get the test statistic**

$$T_1 = \frac{\sqrt{N}\,(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{n_1 n_2 C_1 C_2}}$$

which has a standard normal **distribution when H_0 is true.**

for large samples

- If T_1 is far from zero, this indicates that $p_1 \neq p_2$
- If T_1 is far below zero, this indicates that $p_1 < p_2$
- If T_1 is far above zero, this indicates that $p_1 > p_2$

Decision Rules

H_1: $p_1 \neq p_2$	H_1: $p_1 < p_2$	H_1: $p_1 > p_2$
Reject H_0 if $\lvert T_1 \rvert > Z_{1-\alpha/2}$	Reject H_0 if $T_1 < Z_\alpha = -Z_{1-\alpha}$	Reject H_0 if $T_1 > Z_{1-\alpha}$
P-value: $2\left[\min\{P(Z < T_1^{obs}), P(Z > T_1^{obs})\}\right]$	$P(Z < T_1^{obs})$	$P(Z > T_1^{obs})$

- **Note: The normal approximation for T_1 is valid for large samples, say, if**

each of O_{11}, O_{12}, O_{21}, O_{22} are at least 5.

Example 1: A survey was conducted of 160 rural households and 261 urban households with Christmas trees. Of interest was whether the tree was natural or artificial. Is the probability of natural trees different for rural and urban households? Use $\alpha = 0.05$.

Data:

Population		Tree: Natural	Tree: Artificial	
	Rural	64	96	160
	Urban	89	172	261
		153	268	421

H₀: $p_1 = p_2$ **H₁:** $p_1 \neq p_2$

Test statistic:

$$T_1 = \frac{\sqrt{421}\,[(64)(172)-(96)(89)]}{\sqrt{(160)(261)(153)(268)}} = 1.22$$

Reject Ho if $|T_1| > z_{.975} = 1.96$ (top, Table A1). Since $|1.22| \not> 1.96$, fail to reject Ho. Cannot conclude the probability of natural tree differs for urban and rural households. P-value ≈ 0.2218 from R.

Example 2: Page 184 gives data from a study to determine whether a new lighting system worsened midshipmen's vision.

Data:

		Vision		
		Good	**Poor**	
Lighting	**Old**	**714**	**111**	825
	New	**662**	**154**	816
		1376	265	1641

H₀: $p_1 \leq p_2$ **H₁:** $p_1 > p_2$

Test statistic:

$$T_1 = \frac{\sqrt{1641}\,[(714)(154)-(111)(662)]}{\sqrt{(825)(816)(1376)(265)}} = 2.982$$

Reject Ho if $T_1 > z_{.95} = 1.645$ (Table A1, top).

⇒ Reject Ho. Conclude the old lighting produced a better chance of good vision than new lighting.

P-value = .0014 from R.

Fisher's Exact Test

• **In the previous examples, the row totals were the sizes of the two samples, which are <u>fixed</u> before the data are examined (i.e., they are not random).**

• **When we have a single sample in which two** binary **variables are measured on each individual, the resulting 2 × 2 table has** random **row totals and** random **column totals.**

• **We will cover that scenario in Section 4.2.**

• **In other situations, both the row totals and the column totals may be** fixed **prior to the data being examined.**

• **In this case of "** fixed **margins", Fisher's Exact Test is ideal.**

Data setup:

	Column 1	Column 2	
Row 1	x	$r-x$	r
Row 2	$c-x$	$N-r-c+x$	$N-r$
	c	$N-c$	N

• **We again wish to compare:**

p_1 = probability of an observation in row 1 being classified into column 1

p_2 = probability of an observation in row 2 being classified into column 1

Test statistic T_2 = x = number of observations in (1, 1) cell

Null Distribution

- **Let p = probability an observation is in Column 1.**
- **Under H_0, this probability is the same whether the observation is in Row 1 or Row 2. Then:**

P(table results | row totals) = $\binom{r}{x}\binom{N-r}{c-x}p^c(1-p)^{N-c}$

P(column totals) = $\binom{N}{c}p^c(1-p)^{N-c}$

→ P(table results | row totals & column totals) =

$$\frac{\binom{r}{x}\binom{N-r}{c-x}p^c(1-p)^{N-c}}{\binom{N}{c}p^c(1-p)^{N-c}} = \frac{\binom{r}{x}\binom{N-r}{c-x}}{\binom{N}{c}}$$

- **The decision is based on the P-value, which is found differently depending on the alternative hypothesis:**

H_1: $p_1 \neq p_2$	**H_1:** $p_1 < p_2$	**H_1:** $p_1 > p_2$
P-val $= 2\left[\min\{P(T_2 \le T_2^{obs}), P(T_2 \ge T_2^{obs})\}\right]$	P-val $= P(T_2 \le T_2^{obs})$	P-val $= P(T_2 \ge T_2^{obs})$

- **In all cases, reject H_0 if the p-value $\le \alpha$.**

Example 3: Fourteen new hires (10 male and 4 female) are being assigned to bank positions (there are 4 account representative positions open and 10 (less desirable) teller positions open. The data on page 190 summarize the assignments. If all new employees are equally qualified, is there evidence that female hires were more likely to get the account representative jobs?

Data:

	Account Rep	Teller	
Males	1	9	10
Females	3	1	4
	4	10	14

H₀: $p_1 \geq p_2$ **H₁:** $p_1 < p_2$

Test statistic: $T_2^{obs} = 1$

P-value: $P(T_2 \leq 1) = P(T_2 = 0) + P(T_2 = 1)$

$r = 10$
$N = 14$
$c = 4$

$$= \frac{\binom{10}{0}\binom{4}{4-0}}{\binom{14}{4}} + \frac{\binom{10}{1}\binom{4}{4-1}}{\binom{14}{4}} = .041$$

Since $.041 \leq .05$, we reject Ho and conclude female hires more likely to get acct. rep. jobs.

• **See `fisher.test` function in R to perform this test.**

• **Fisher's Exact Test may be used if the row totals and/or column totals are random, but in this case it is** more conservative **than the z-test.**

• **Fisher's Exact Test can also be viewed as an alternative to the z-test when the large-sample rule is not met, but the Exact Test** lacks power **when the sample size is very small.**

• **Suppose we have several related (but not identical) conditions in which sub-experiments are conducted, each of which produces a 2 × 2 table.**

• **It is of interest to see whether rows and columns are independent in <u>each</u> table.**

Mantel-Haenszel Test

- **We assume we have $k \geq 2$ such 2×2 tables, each with fixed row and column totals (although the test can be done even with random totals).**

Let p_{1i} = probability of an observation in row 1 being classified into column 1, in the i-th table.

and p_{2i} = probability of an observation in row 2 being classified into column 1, in the i-th table.

Hypotheses:

$H_0: p_{1i} = p_{2i}$ for $i=1,\dots,k$	$H_0: p_{1i} \geq p_{2i}$ for all i	$H_0: p_{1i} \leq p_{2i}$ for all i
H_1: Either $p_{1i} > p_{2i}$ for some i, or $p_{1i} < p_{2i}$ for some i, but not both	$H_1: p_{1i} \leq p_{2i}$ for all i and $p_{1i} < p_{2i}$ for some i	$H_1: p_{1i} \geq p_{2i}$ for all i, and $p_{1i} > p_{2i}$ for some i

Test statistic

$$T_4 = \frac{\sum_{i=1}^{k} x_i - \sum_{i=1}^{k} \frac{r_i c_i}{N_i}}{\sqrt{\sum_{i=1}^{k} \frac{r_i c_i (N_i - r_i)(N_i - c_i)}{N_i^2 (N_i - 1)}}}$$

- **The null distribution is approximately standard normal, tabulated in Table A1.**

Decision Rules and P-value:

Two-Tailed	Lower-Tailed	Upper-Tailed
Reject Ho if	Reject Ho if	Reject Ho if
$T_4 > Z_{1-\alpha/2}$ or $T_4 < Z_{\alpha/2}$	$T_4 < Z_\alpha$	$T_4 > Z_{1-\alpha}$
P-value = $2\left[\min\{P(Z \le T_4^{obs}), P(Z \ge T_4^{obs})\}\right]$	P-value = $P[Z \le T_4^{obs}]$	P-value = $P[Z \ge T_4^{obs}]$

Example 4: Three groups of cancer patients were given either a drug treatment or a control, and for each patient, whether the outcome was successful was recorded. Is there evidence that <u>in at least one group</u>, the treatment produces a better chance of success than the control? (Use $\alpha = 0.05$.)

Data:

	Group 1		Group 2		Group 3	
	Success	Failure	Success	Failure	Success	Failure
Treatment	10	1	9	0	8	0
Control	12	1	11	1	7	3

H0: $p_{1i} \le p_{2i}$ for all i **H1:** $p_{1i} > p_{2i}$ for some i
(and $p_{1i} \ge p_{2i}$ for all i)

Test statistic: $T_4 = 1.0057$ (R reports T_4^2)
P-value: 0.157 from R

Conclusion: There is not evidence that the success probability is better for the treatment than for the control, in any group.

- **See `mantelhaen.test` function in R to perform this test.**