STAT 518 --- Section 4.2 --- Tests for $r \times c$ Tables

• **We now consider more general two-way tables:**

• **In Sec. 4.1 we had two samples in which a two-category variable is measured on each individual in each sample.**

• **Now suppose we have r samples in which the same c -category variable is measured on each individual in each sample.**

Comparing Multinomial Probabilities Across Several Independent Samples

• **Suppose we have r independent samples, with respective sizes $n_1, n_2, \ldots, n_r$. We classify each individual in each sample into class 1, 2, …, c.**

• **Our data (which could be nominal or ordinal) could be arranged in an $r \times c$ table as follows:**

	Class 1	Class 2	…	Class c	
Sample 1	O_{11}	O_{12}	…	O_{1c}	n_1
Sample 2	O_{21}	O_{22}	…	O_{2c}	n_2
⋮	⋮	⋮		⋮	⋮
Sample r	O_{r1}	O_{r2}	…	O_{rc}	n_r
	C_1	C_2	…	C_c	N

Chi-Square Test for Homogeneity in a Two-Way Table

- **This is a basic extension of the two-tailed z-test comparing p_1 and p_2.**

Hypotheses: $H_0: p_{1j} = p_{2j} = \cdots = p_{rj}$ for all j.

$H_1: p_{ij} \neq p_{kj}$ for some j and for some i, k.

Test Statistic

$$T = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}} - N, \quad \text{where } E_{ij} = \frac{n_i C_j}{N}$$

which has an asymptotic χ^2 **distribution with** $(r-1)(c-1)$ **degrees of freedom when H_0 is true.**

- **Note if H_0 is true and all the populations have the same set of class probabilities, the expected count in cell (i, j) is the** size of the i-th sample **times** the proportion of observations (of all N) falling in category j.

- **If $r = c = 2$, this $T =$** T_1^2 **from Section 4.1.**

- **If T is far from zero, this indicates that** H_0 is false and that the probability distribution differs among the r populations.

Decision Rule:

Reject H_0 if $T > \chi^2_{1-\alpha,\,(r-1)(c-1)}$

found in Table A2.

• The P-value is found through interpolation in Table A2 or using R.

• Note: The χ^2 approximation for T is valid for large samples, say, if <u>all</u> E_{ij}'s are greater than 0.5 and <u>at least half</u> of the E_{ij}'s are greater than 1.0.

• If some expected cell counts are too small, two or more categories could be combined, as long as this is sensible.

Example 1: Page 202 gives test score category counts from a sample of public school students and from a sample of private school students. Is the probability distribution of scores equal for public and private school students? Use $\alpha = 0.05$.

$E_{11} = \frac{(46)(36)}{128} = 12.94$, $E_{12} = \frac{(46)(46)}{128} = 16.53$, $E_{13} = \frac{(46)(34)}{128} = 12.22$, etc.

Data:

	Low	Marginal	Good	Excellent	
Private	6	14	17	9	46
Public	30	32	17	3	82
	36	46	34	12	128

(Score)

H₀: $p_{1j} = p_{2j}$ for all $j = 1, 2, 3, 4$ **H₁:** $p_{1j} \neq p_{2j}$ for some j

Test statistic:

$$T = \frac{6^2}{12.94} + \frac{14^2}{16.53} + \frac{17^2}{12.22} + \frac{9^2}{4.31} + \frac{30^2}{23.06} + \frac{32^2}{29.47} + \frac{17^2}{21.78} + \frac{3^2}{7.69} - 128 = \boxed{17.29}$$

Decision rule and conclusion:

Reject H₀ if $T > \chi^2_{.95,3} = 7.815$ (Table A2). Since $17.29 > 7.815$, we reject H₀ and conclude the probability distribution differs for public and private school students.

P-value

$\approx .0006$ from R

Chi-Square Test for Independence

• Now we consider observations in a single sample of size *N* that are classified according to <u>two</u> categorical variables.

• Such data can also be presented in a <u>two-way</u> table.

Example: Suppose the people in the "favorite-sport" survey had been further classified by gender:

		Sport					
		Football	Bsbl	Bskbl	Auto	Golf	Other
Gender	Male						
	Female						

• Two categorical variables: Gender **and** Sport

<u>Question</u>: Are the two classifications independent or dependent?

• For instance, does people's favorite sport depend on their gender? Or does gender have no association with favorite sport?

• Unlike the *r*-sample problem, in this situation both column totals <u>and</u> row totals are random (only *N* is fixed).

Observed Counts for a $r \times c$ Contingency Table
(r = # of rows, c = # of columns)

Column Variable

		1	2	...	c	Row Totals
Row Variable	1	O_{11}	O_{12}	...	O_{1c}	R_1
	2	O_{21}	O_{22}	...	O_{2c}	R_2
	⋮	⋮	⋮		⋮	⋮
	r	O_{r1}	O_{r2}	...	O_{rc}	R_r
Col. Totals		C_1	C_2	...	C_c	N

Probabilities for a $r \times c$ Contingency Table:

Column Variable

		1	2	...	c	
Row Variable	1	p_{11}	p_{12}	...	p_{1c}	$p_{\text{row }1}$
	2	p_{21}	p_{22}	...	p_{2c}	$p_{\text{row }2}$
	⋮	⋮	⋮		⋮	⋮
	r	p_{r1}	p_{r2}	...	p_{rc}	$p_{\text{row }r}$
		$p_{\text{col }1}$	$p_{\text{col }2}$	...	$p_{\text{col }c}$	1

- **Note: If the two classifications are <u>independent</u>, then: $p_{11} = (p_{\text{row }1})(p_{\text{col }1})$ and $p_{12} = (p_{\text{row }1})(p_{\text{col }2})$, etc.**

- **So under the hypothesis of independence, we expect the cell probabilities to be the product of the corresponding <u>marginal probabilities</u>:**

$$p_{ij} = (p_{\text{row }i})(p_{\text{col }j})$$

Hence if H_0 is true, the (estimated) expected count in cell (i, j) is simply:

$$N p_{ij} = N(p_{row\, i})(p_{col\, j}) \approx N\left(\frac{R_i}{N}\right)\left(\frac{C_j}{N}\right) = \frac{R_i C_j}{N}$$

χ^2 test for independence

H_0: The classifications are independent
H_a: The classifications are dependent

Test statistic:

$$T = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \left(\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}}\right) - N$$

where the expected count in cell (i, j) is $E_{ij} = \frac{R_i C_j}{N}$

Decision Rule:

Reject H_0 if $T > \chi^2_{1-\alpha,\ (r-1)(c-1)}$

↑ Table A2.

- **The P-value is found through interpolation in Table A2 or using R.**

Note: The same large-sample rule of thumb applies as in the previous χ^2 test.

Example: Does the incidence of heart disease depend on snoring pattern? (Test using $\alpha = .05$.) Random sample of 2484 adults taken; results given in a <u>contingency table</u>:

		Snoring Pattern			
		Never	Occasionally	≈Every Night	
Heart Disease	Yes	24	35	51	110
	No	1355	603	416	2374
		1379	638	467	2484

Expected Cell Counts:

$$E_{11} = \frac{(110)(1379)}{2484} = 61.07, \quad E_{12} = \frac{(110)(638)}{2484} = 28.25,$$

$$\ldots, \quad E_{23} = \frac{(2374)(467)}{2484} = 446.32$$

Test statistic:

$$T = \frac{24^2}{61.07} + \frac{35^2}{28.25} + \frac{51^2}{20.68} + \frac{1355^2}{1317.93} + \frac{603^2}{609.75} + \frac{416^2}{446.32} - 2484 = \boxed{71.75}$$

Decision rule and conclusion:

$(r-1)(c-1) = 2$, so

Reject H_0 if $T > \chi^2_{.95,2} = 5.991$. Since $71.75 > 5.99$, we reject H_0 and conclude the incidence of heart disease is associated with snoring pattern.

P-value ≈ 0 from R.

Tests for $r \times c$ Tables with Fixed Marginal Totals

- **If the table has r rows and c columns and both the row totals and column totals are fixed, an extended version of the Exact Test is available.**

- **In this case, there are no one-tailed alternatives possible – the hypotheses are simply** the same as for the χ^2 test for homogeneity or the χ^2 test for independence, depending on the sampling setup.

- **The P-value are obtained using `fisher.test` in R, as the exact null distribution is cumbersome.**

- **The exact P-value is obtained by considering all possible tables resulting in the given margins, and sorting these by how favorable to H_1 they are.**

- **The exact P-value is the proportion of possible tables that are** as or more **favorable to H_1 as the table we observed.**

Example Data (alteration of bank data to a 3 × 3 table):

		Position			
		Acct. Rep	Teller	Data Analyst	
Race	White	0	5	1	6
	Black	2	3	0	5
	Asian	2	0	1	3
		4	8	2	14

P-value and conclusion: P-value = .0566 from R.
At $\alpha = .05$, cannot conclude the probabilities of the various jobs differ across the races.

Section 4.3 --- Median Test

• **We return to the situation in which we want to know whether several (c) populations have the same median.**

• **For $c > 2$, this is similar to the setup of the** Kruskal-Wallis **test.**

• **For $c = 2$, this is similar to the setup of the** Mann-Whitney **test.**

• **The difference is in the <u>conditions</u> of the tests:**
The M-W and K-W tests assume that under H_0,

the c populations have identical distributions.

while the Median Test assumes only that under H_0,

the c populations have the same median.

• **So the Median Test can be applied** more generally.

• **Suppose from each of c populations, we have a random sample, with sizes $n_1, n_2, \ldots, n_c$.**

• **We assume that the c samples are independent and that the data are at least ordinal, so that the "median" is a meaningful measure.**

• **Calculate the grand median of all $N = n_1 + n_2 + \ldots + n_c$ observations, and arrange the data into a $2 \times c$ table:**

Sample

	1	2	....	c	
> Grand Median	O_{11}	O_{12}	. . .	O_{1c}	a
≤ Grand Median	O_{21}	O_{22}	. . .	O_{2c}	b
	n_1	n_2	....	n_c	N

Hypotheses:

H_0: All c populations have the same median

H_1: At least 2 populations have different medians

- **The null hypothesis implies that being in the top row or bottom row is independent of which column (population) an observation is in.**

- **Note that the expected cell count under H_0 is**

$E_{1i} = \dfrac{n_i a}{N}$ **for the top-row cells, and**

$E_{2i} = \dfrac{n_i b}{N}$ **for the bottom-row cells.**

So the test statistic, as in the χ^2 test for independence, is

$$T = \sum_{i=1}^{c} \frac{\left(O_{1i} - \frac{n_i a}{N}\right)^2}{\frac{n_i a}{N}} + \sum_{i=1}^{c} \frac{\left(O_{2i} - \frac{n_i b}{N}\right)^2}{\frac{n_i b}{N}}$$

which can be simplified into

$$T = \frac{N^2}{ab} \sum_{i=1}^{c} \frac{\left(O_{1i} - \frac{n_i a}{N}\right)^2}{n_i} = \left(\frac{N^2}{ab} \sum_{i=1}^{c} \frac{O_{1i}^2}{n_i}\right) - \frac{Na}{b}$$

since

$$O_{2i} = n_i - O_{1i}$$

- **The asymptotic null distribution of *T* is** χ^2_{c-1}

Decision rule:

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

• **The P-value is found through interpolation in Table A2 or using R.**

Note: The same large-sample rule of thumb applies as in the previous χ^2 test.

• **The median test may be generalized to test about any particular quantile – in that case, the appropriate "grand quantile" is used instead of the "grand median".**

Example 1: Bidding/Buy-It-Now Data from Section 5.1 notes. At $\alpha = .05$, are the median selling prices significantly different for the two groups?
Data:
Bidding: 199, 210, 228, 232, 245, 246, 246, 249, 255
BIN: 210, 225, 225, 235, 240, 250, 251

Grand Median: 237.5 **$c =$** 2 **. $2 \times c$ table:**

	Bidding	BIN	
> Grand Median	5	3	8
≤ Grand Median	4	4	8
	9	7	16

Test statistic $T =$ $\dfrac{16^2}{(8)(8)}\left(\dfrac{5^2}{9} + \dfrac{3^2}{7}\right) - \dfrac{(16)(8)}{(8)} = 0.254$

Decision Rule and Conclusion: Reject H_0 if $T > \chi^2_{.95,1} = 3.84$ ← Table A2

Since $0.254 \not> 3.84$, we fail to reject H_0. The two methods may have the same median price.

P-value ≈ 0.614 from R.

Example 2: Data on page 221 gives corn yields for four different growing methods. At $\alpha = .05$, are the median yields significantly different for the four methods?

Grand Median: 89 $c =$ 4. $2 \times c$ table:

	Method 1	2	3	4	
> Grand Median	6	3	7	0	16
≤ Grand Median	3	7	0	8	18
	9	10	7	8	34

Test statistic

$$T = \frac{34^2}{(16)(18)}\left(\frac{6^2}{9} + \frac{3^2}{10} + \frac{7^2}{7} + \frac{0^2}{8}\right) - \frac{(34)(16)}{18} = 17.54$$

Decision Rule and Conclusion:

Reject Ho if $T > \chi^2_{.95,3} = 7.815$. Since $17.54 > 7.815$, we reject Ho and conclude the median yields differ among the 4 methods.

P-value $\approx .0005$ from R.

Comparison of Median Test to Competing Tests

- **The classical parametric approach for comparing the centers of several populations is the** ANOVA F-test.

- **In Sec. 5.1 we examined the efficiency of the Mann-Whitney test relative to the median test when $c = 2$.**

- **Of these options, the median test is the most flexible since it makes the fewest assumptions about the data.**

- **The A.R.E. of the median test relative to the F-test is** 0.64 **with normal populations and** 2.00 **with double exponential (heavy-tailed) populations.**