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 <u>two</u> samples in which a <u>two-category</u> variable is measured on each individual in each sample.

• Now suppose we have ___ samples in which the same <u>___-category</u> 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:

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:

Test Statistic

which has an asymptotic ______ distribution with _____ 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 _____________________________________ times ____

• If $r = c = 2$, this $T =$ from Section 4.1.

• If T is far from zero, this indicates that

Decision Rule:

• 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

• 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 $\alpha = 0.05$.

Data:

			Score	
	Low	Marginal	Good	Excellent
Private	6	14	17	9
Public	30	32	17	3

H_0: **H_1:**

Test statistic:

Decision rule and conclusion:

P-value

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:

• Two categorical variables: ________ and ___________

<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
	1	O_{11}	O_{12}	…	O_{1c}	R_1
Row	2	O_{21}	O_{22}	…	O_{2c}	R_2
	⋮	⋮	⋮		⋮	⋮
Variable	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	
	1	p_{11}	p_{12}	…	p_{1c}	$p_{\text{row }1}$
Row	2	p_{21}	p_{22}	…	p_{2c}	$p_{\text{row }2}$
	⋮	⋮	⋮		⋮	⋮
Variable	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>:**

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

χ^2 test for independence

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

Test statistic:

where the expected count in cell (*i, j*) is

Decision Rule:

• 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	Yes	24	35	51	110
Disease	No	1355	603	416	2374
		1379	638	467	2484

Expected Cell Counts:

Test statistic:

Decision rule and conclusion:

P-value

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 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 ________________ favorable to H_1 as the table we observed.

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

P-value and conclusion:

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 __________ test.
• For $c = 2$, this is similar to the setup of the __________ test.

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

while the Median Test assumes only that under H_0,

• So the Median Test can be applied ______ _________.

• 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:

Hypotheses:

• **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**

for the top-row cells, and

for the bottom-row cells.

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

which can be simplified into

since

• **The asymptotic null distribution of *T* is**

Decision rule:

• 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: **c = ____. 2 × c table:**

Test statistic T =

Decision Rule and Conclusion:

P-value

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: **c = ____. 2 × c table:**

Test statistic

Decision Rule and Conclusion:

P-value

Comparison of Median Test to Competing Tests

- **The classical parametric approach for comparing the centers of several populations is the _________________.**

- **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 ______ with normal populations and _______ with double exponential (heavy-tailed) populations.**