STAT 333
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, … , 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	Total
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
Total	$.c_1$	$.c_2$			$.c_c$	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: $p_{1j} = p_{2j} = \ldots\ldots\ldots\ldots = p_{rj}$	for all j
H_1: $p_{ij} \neq p_{kj}$	for some j and for some i, k

Test Statistic:

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

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

(get the value $\chi^2_{1-\alpha,(r-1)(c-1)}$ from chi-square 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

All E_{ij}'s are greater than 0.5 and at least half of the E_{ij}'s are greater than 1.

• 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	Total
Private	6	14	17	9	46
Public	30	32	17	3	82
Total	36	46	34	12	128 =N

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

Test statistic:

First calculate $E_{ij} = \frac{n_i c_i}{N}$

$$E_{11} = \frac{46x36}{128} = 12.94\,, E_{12} = \frac{46x46}{128} = 16.53\,, E_{13} = \frac{46x34}{128} = 12.22$$

$$E_{14} = \frac{46x12}{128} = 4.31\,, E_{21} = \frac{82x36}{128} = 23.06, E_{22} = \frac{82x46}{128} = 29.47$$

$$E_{23} = \frac{82x34}{128} = 21.78\,, E_{24} = \frac{82x12}{128} = 7.69$$

	Low	Marginal	Good	Excellent
Private	O_{11}=6 E_{11}= 12.94	O_{12}=14 E_{12}=16.53	O_{13}=17 E_{13}=12.22	O_{14}=9 E_{14}=4.31
Public	O_{21}=30 E_{21}=23.06	O_{22}=32 E_{22}=29.47	O_{23}=17 E_{23}=21.78	O_{24}=3 E_{24}=7.69

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

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

= 17.29

Decision rule and conclusion:

Reject H_0 if $T > \chi^2_{0.95,3}$ ($\chi^2_{0.95,3} = 7.815$ *from table A2*)

Since, 17.29 > 7.815

Then , we reject H_0 and conclude that the probability distribution differs for public and private school students

P-value = 0.006 (from R : P-value =1-pchisq(17.29,3)≈0.006)

Chi-Square Test for Independence

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

• Such data can also be presented in a two-way table.

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

Sport

Gender		Football	Baseball	Basket	Auto	Golf	other
	Male						
	Female						

• Two categorical variables: Gender and Sport .

Question: 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)

<u>Column Variable</u>

		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:

<u>Column Variable</u>

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

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

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

$$P_{ij} = (p_{rowi})\ (p_{col\ j})$$

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

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

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

(get the value of $\chi^2_{1-\alpha,(r-1)(c-1)}$ from chi- square 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 $\alpha = 0.05$) Random sample of 2484 adults taken; results given in a contingency table:

		Snoring Pattern			
		Never	Occasionally	Every Night 1	Total
Heart	Yes	24	35	51	110
Disease	No	1355	603	416	2374
Total		1379	638	467	2484=N

Expected Cell Counts: $E_{ij} = \frac{R_i C_j}{N}$

$$E_{11} = \frac{110x1379}{2484} = 61.07\,, E_{12} = \frac{110x638}{2484} = 28.25\,, E_{13} = \frac{110x467}{2484} = 20.68$$

$$E_{21} = \frac{2374x1379}{2484} = 1317.93\,, E_{22} = \frac{2374x638}{2484} = 609.75\,, E_{23} = \frac{2374x467}{2484} = 446.32$$

		Never	Occasionally	Every Night
Heart	Yes	$O_{11}=24$ $E_{11}= 61.07$	$O_{12}=35$ $E_{12}=28.25$	$O_{13}=51$ $E_{13}= 20.68$
Disease	No	$O_{21}=1355$ $E_{21}= 1317.93$	$O_{22}=603$ $E_{22}= 609.75$	$O_{23}=416$ $E_{23}= 446.32$

Test statistic:

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

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

Decision rule and conclusion:

Df = (r-1)(c-1)=(2-1)(3-1) =2 , 1-α = 0.95 so

Reject H_0 if T > $\chi^2_{0.95,2}$ (From table A2 : $\chi^2_{0.95,2} = 5.99$)

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: P-value =1-pchisq(71.75,2) ≈ 0)

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 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

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

P-value and conclusion:

P-value = 0.0566 from R
At α = 0.05 , 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 conditions 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	Total
>Grade Median	O_{11}	O_{12}	.	.	O_{1C}	a
≤ Grade Median	O_{21}	O_{22}	.	.	O_{2C}	b
Total	n_1	n_2	.	.	n_C	N

Hypotheses:

H_0: All C populations have the same medians.
H_A: 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

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

$E_{2i} = \frac{n_i\, b}{N}$ 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**

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

since

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

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

<u>Decision rule:</u>

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

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

<u>Note:</u> 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 $\alpha = 0.05$, 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 (From data) c = 2 , 2× c table:

	Bidding	BIN	Total
>Grade Median	$5 = O_{11}$	$3 = O_{12}$	8 = a
≤ Grade Median	4	4	8= b
Total	$9 = n_1$	$7 = n_2$	16 =N

Test statistic :

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

$$= \frac{16^2}{8x8}\, x \left(\frac{5^2}{9} + \frac{3^2}{7}\right) - \frac{16x8}{8} = 0.254$$

Decision Rule and Conclusion:

df= c-1 =2-1=1 , $1-\alpha = 1-0.05 = 0.95$ (Get chi value from A2)

Reject H_0 if $T > \chi^2_{0.95,1}$

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: P-value =1-pchisq(0.254 ,1)≈ 0.614)

Example 2: Data on page 104 gives corn yields for four different growing methods. At α = 0.05, are the median yields significantly different for the four methods?
Data:

Method 1	Method 2	Method 3	Method 4
83	91	101	78
91	90	100	82
94	81	91	81
89	83	93	77
89	84	96	79
96	83	95	81
91	88	94	80
92	91		81
90	89		
	84		

Grand Median: = 89 (From data in page 104 in book) $c = 4$, $(2 \times c)$ table:

	1	2	3	4	Total
>Grade Median	$6 = O_{11}$	$3 = O_{12}$	$7 = O_{13}$	$0 = O_{14}$	16 = a
≤ Grade Median	3	7	0	8	18 = b
Total	$9 = n_1$	$10 = n_2$	$7 = n_3$	$8 = n_4$	34 = N

Test statistic:

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

$$= \frac{34^2}{16x18} \, x \left(\frac{6^2}{9} + \frac{3^2}{10} + \frac{7^2}{7} + \frac{0^2}{8} \right) - \frac{34x16}{18} = 17.54$$

Decision Rule and Conclusion:

(df= c-1= 4 -1=3 , 1-α =1- 0.05 = 0.95)

Reject H_0 $T > \chi^2_{0.95,3}$ (from table A2 , $\chi^2_{0.95,3} = 7.815$)

Since, 17.54 > 7.815

We reject H_0 and conclude that the median yields differ among the 4 methods

P-value = 0.005 (from R: P-value =1-pchisq(17.54,3)≈ 0.005

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.

****

TABLE A2 Chi-Squared Distribution[a]

	p = 0.750	0.900	0.950	0.975	0.990	0.995	0.999
k = 1	1.323	2.706	3.841	5.024	6.635	7.879	10.83
2	2.773	4.605	5.991	7.378	9.210	10.60	13.82
3	4.108	6.251	7.815	9.348	11.34	12.84	16.27
4	5.385	7.779	9.488	11.14	13.28	14.86	18.47
5	6.626	9.236	11.07	12.83	15.09	16.75	20.51
6	7.841	10.64	12.59	14.45	16.81	18.55	22.46
7	9.037	12.02	14.07	16.01	18.48	20.28	24.32
8	10.22	13.36	15.51	17.53	20.09	21.96	26.13
9	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	13.70	17.28	19.68	21.92	24.73	26.76	31.26
12	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	23.83	28.41	31.41	34.17	37.57	40.00	45.32
21	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	28.24	33.20	36.42	39.37	42.98	45.56	51.18
25	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	45.62	51.81	55.76	59.34	63.69	66.77	73.40
50	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	77.58	85.53	90.53	95.02	100.4	104.2	112.3
80	88.13	96.58	101.9	106.6	112.3	116.3	124.8
90	98.65	107.6	113.1	118.1	124.1	128.3	137.2
100	109.1	118.5	124.3	129.6	135.8	140.2	149.4
z_p	0.675	1.282	1.645	1.960	2.326	2.576	3.090

For $k > 100$ use the approximation $w_p = (\frac{1}{2})(z_p + \sqrt{2k-1})^2$, or the more accurate $w_p = k\left(1 - \frac{2}{9k} + z_p\sqrt{\frac{2}{9k}}\right)^3$, where z_p is the value from the standardized normal distribution shown in the bottom of the table.

SOURCE: Abridged from Table 8, Vol. I of Pearson and Hartley (1976), with permission from the *Biometrika*, Trustees.