Chi-Square Tests

Chi-Square Test for Independence

❈ *Contingency Tables*

A *contingency table* is a cross-tabulation of *n* paired observations into categories

- Each cell shows the count of observations that fall into the category defined by its row (*r*) and column (*c*) heading.

	Variable A				
Variable B	1	2		c	Row Total
1	f_{11}	f_{12}	...	f_{1c}	R_1
2	f_{21}	f_{22}	...	f_{2c}	R_2
:	:	:	:	:	:
r	f_{r1}	f_{r2}	...	f_{rc}	R_r
Col Total	C_1	C_2	...	C_c	n

- For example:

TABLE 15.2 Privacy Disclaimer Location and Web Site Nationality WebSites

	Nationality of Web Site			
Location of Disclaimer	France	UK	USA	Row Total
Home page	56	68	35	159
Order page	19	19	28	66
Client page	6	10	16	32
Other page	12	9	13	34
Col Total	93	106	92	291

Source: Calin Gurau, Ashok Ranchhod, and Claire Gauzente, "To Legislate or Not to Legislate: A Comparative Exploratory Study of Privacy/Personalisation Factors Affecting French, UK, and US Web Sites," *Journal of Consumer Marketing* 20, no. 7 (2003), p. 659.

- In a test of independence for an *r* x *c* contingency table, the hypotheses are

H_0: Variable *A* is independent of variable *B*

H_1: Variable *A* is not independent of variable *B*

- Use the *chi-square test for independence* to test these hypotheses.
- This *non-parametric* test is based on *frequencies*.
- The *n* data pairs are classified into *c* columns and *r* rows and then the

observed frequency f_{jk} is compared with the *expected frequency* e_{jk}.

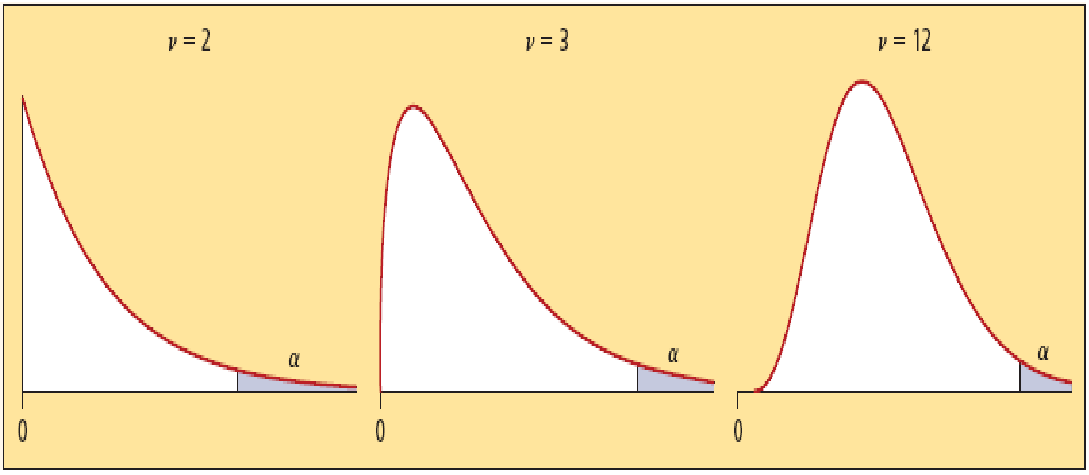
- The critical value comes from the *chi-square probability distribution* with n degrees of freedom.

n = degrees of freedom = $(r - 1)(c - 1)$

where r = number of rows in the table

c = number of columns in the table

- Appendix E contains critical values for right-tail areas of the chi-square distribution.
- The mean of a chi-square distribution is n with variance 2n.
- Consider the shape of the chi-square distribution:



- Assuming that H_0 is true, the expected frequency of row j and column k is:

$e_{jk} = R_j C_k / n$

where R_j = total for row j ($j = 1, 2, \ldots, r$)

C_k = total for column k ($k = 1, 2, \ldots, c$)

n = sample size

- The table of expected frequencies is:

	Variable A				
Variable B	1	2	. . .	c	Row Total
1	e_{11}	e_{12}	. . .	e_{1c}	R_1
2	e_{21}	e_{22}	. . .	e_{2c}	R_2
⋮	⋮	⋮	⋮	⋮	⋮
r	e_{r1}	e_{r2}	. . .	e_{rc}	R_r
Col Total	C_1	C_2	. . .	C_c	n

- The e_{jk} always sum to the same row and column frequencies as the observed frequencies.
- Step 1: State the Hypotheses

H_0: Variable A is independent of variable B

H_1: Variable A is not independent of variable B

- Step 2: State the Decision Rule

Calculate n = $(r – 1)(c – 1)$
For a given a, look up the right-tail critical value (χ^2_R)from Appendix E or

by using Excel.
Reject H_0 if χ^2_R > test statistic.

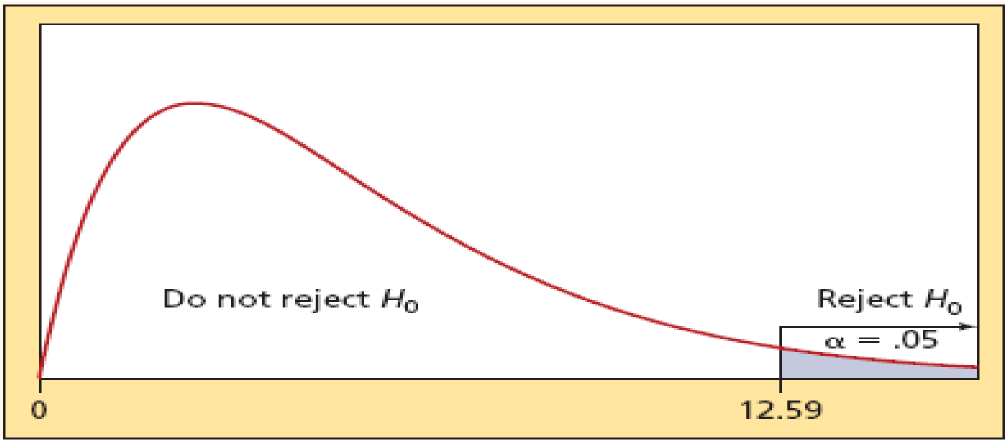
- For example, for n = 6 and a = .05, $\chi^2_{.05}$ = 12.59.

Appendix E: Critical Values for Chi-Square

This table shows the critical value that defines the specified area for the stated degrees of freedom (ν).

ν	Left Tail Area					Right Tail Area				
	0.005	0.01	0.025	0.05	0.10	0.10	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.60
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28	14.86
5	0.412	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09	16.75
6	0.676	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81	18.55
7	0.989	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48	20.28
8	1.344	1.646	2.180	2.733	3.490	13.36	15.51	17.53	20.09	21.95
9	1.735	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67	23.59
10	2.156	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21	25.19

- Here is the rejection region.



- Step 3: Calculate the Expected Frequencies $e_{jk} = R_j C_k / n$

- For example,

Location	Expected Frequencies France	UK	USA	Row Total
Home	(159 × 93)/291 = 50.81	(159 × 106)/291 = 57.92	(159 × 92)/291 = 50.27	159
Order	(66 × 93)/291 = 21.09	(66 × 106)/291 = 24.04	(66 × 92)/291 = 20.87	66
Client	(32 × 93)/291 = 10.23	(32 × 106)/291 = 11.66	(32 × 92)/291 = 10.12	32
Other	(34 × 93)/291 = 10.87	(34 × 106)/291 = 12.38	(34 × 92)/291 = 10.75	34
Col Total	93	106	92	291

- Step 4: Calculate the Test Statistic

The chi-square test statistic is

$$\chi^2 = \sum_{j=1}^{r} \sum_{k=1}^{c} \frac{[f_{jk} - e_{jk}]^2}{e_{jk}}$$

- Step 5: Make the Decision

Reject H_0 if χ^2_R > test statistic or if the

p-value $\leq$ α.

Chi-Square Test for Goodness-of-Fit

❋***Purpose of the Test***

- The *goodness-of-fit* (*GOF*) test helps you decide whether your sample resembles a particular kind of population.
- The chi-square test will be used because it is versatile and easy to understand.
- The hypotheses are:

H_0: The population follows a ………..distribution

H_1: The population does not follow a ………….distribution

- The blank may contain the name of any theoretical distribution (e.g., uniform, Poisson, normal).
- Assuming *n* observations, the observations are grouped into *c* classes and then

the *chi-square test statistic* is found using:

$$\chi^2 = \sum_{j=1}^{c} \frac{[f_j - e_j]^2}{e_j}$$

where f_j = the observed frequency of observations in class *j*
e_j = the expected frequency in class *j* if

H_0 were true

- If the proposed distribution gives a good fit to the sample, the test statistic will be near zero.
- The test statistic follows the chi-square distribution with degrees of freedom
  n = $c - m - 1$
  where *c* is the no. of classes used in the test *m* is the no. of parameters estimated

Uniform: $\nu = c - m - 1 = \nu = c - 0 - 1 = c - 1$ (since no parameters are estimated)

Poisson: $\nu = c - m - 1 = \nu = c - 1 - 1 = c - 2$ (since λ is estimated)

Normal: $\nu = c - m - 1 = \nu = c - 2 - 1 = c - 3$ (since μ and σ are estimated)

Uniform Goodness-of-Fit Test

❈*Multinomial Distribution*

- A *multinomial distribution* is defined by any *k* probabilities $p_1, p_2, \ldots, p_k$ that sum to unity.
- For example, consider the following "official" proportions of M&M colors.

Color	Official π_j	Observed f_j	Expected e_j	$f_j - e_j$	$(f_j - e_j)^2/e_j$
Brown	0.30	58	66	−8	0.9697
Red	0.20	40	44	−4	0.3636
Blue	0.10	34	22	12	6.5455
Orange	0.10	22	22	0	0.0000
Green	0.10	30	22	8	2.9091
Yellow	0.20	36	44	−8	1.4545
Sum	1.00	220	220	0	$\chi^2 = 12.2424$

- The hypotheses are

H_0: $\pi_1 = .30$, $\pi_2 = .20$, $\pi_3 = .10$, $\pi_4 = .10$, $\pi_5 = .10$, $\pi_6 = .20$

H_1: At least one of the π_j differs from the hypothesized value

- No parameters are estimated ($m = 0$) and there are $c = 6$ classes, so the degrees of freedom are

$n = c - m - 1 = 6 - 0 - 1$

- The *uniform goodness-of-fit* test is a special case of the multinomial in which every value has the same chance of occurrence.
- The chi-square test for a uniform distribution compares all *c* groups simultaneously.
- The hypotheses are:

H_0: $\pi_1 = \pi_2 = \ldots, \pi_c = 1/c$

H_1: Not all π_j are equal

- The test can be performed on data that are already tabulated into groups.
- Calculate the expected frequency e_{ij} for each cell.
- The degrees of freedom are n = c – 1 since there are no parameters for the uniform distribution.
- Obtain the critical value χ^2_a from Appendix E for the desired level of significance

α.

- The *p*-value can be obtained from Excel.
- Reject H_0 if *p*-value $\leq$ α.