CATEGORICAL DATA ANALYSIS AND GOODNESS OF FIT TESTS

UNIT-IV

"If you haven't observed it you don't know what you are talking about."

Learning Objectives

After completing this chapter, you should be able to

- Test a distribution for goodness of fit, using chi-square.
- Test two variables for independence, using chi-square.
- Test proportions for homogeneity, using chi-square

**** Applications***

Statistics and Heredity

- An Austrian monk, Gregor Mendel (1822–1884), studied genetics, and his principles are the foundation for modern genetics. Mendel used his spare time to grow a variety of peas at the monastery.
- One of his many experiments involved crossbreeding peas that had smooth yellow seeds with peas that had wrinkled green seeds. He noticed that the results occurred with regularity.
- That is, some of the offspring had smooth yellow seeds, some had smooth green seeds, some had wrinkled yellow seeds, and some had wrinkled green seeds.

Statistics and Heredity…

- Furthermore, after several experiments, the percentages of each type seemed to remain approximately the same.
- Mendel formulated his theory based on the assumption of dominant and recessive traits and tried to predict the results.
- He then crossbred his peas and examined 556 seeds over the next generation.
- Finally, he compared the actual results with the theoretical results to see if his theory was correct. To do this, he used a “simple” chi-square test, which is explained in this chapter.

Example: Titanic

- The ship Titanic sank in 1912 with the loss of most of its passengers
- 809 of the 1,309 passengers and crew died

= 61.8%

- **Research question:** Did class (of travel) affect survival?

Chi-Square as a Statistical Test

- ***Chi-square test:*** an **inferential statistics** technique designed to test for **significant relationships** between two variables organized in a bivariate table.

- Chi-square requires **no assumptions** about the shape of the population distribution from which a sample is drawn.

The Chi Square Test

- A statistical method used to determine goodness of fit
  - Goodness of fit refers to how close the observed data are to those predicted from a hypothesis

- Note:
  - The chi square test does not prove that a hypothesis is correct
    - It evaluates to what extent the data and the hypothesis have a good fit

Limitations of the Chi-Square Test

- The chi-square test does **not** give us much information about the ***strength*** of the relationship or its ***substantive significance*** in the population.
- The chi-square test is **sensitive** to ***sample size***. The size of the calculated chi-square is **directly proportional** to the size of the sample, independent of the strength of the relationship between the variables.
- The chi-square test is also **sensitive** to **small expected frequencies** in one or more of the cells in the table.

Statistical Independence

- *Independence (statistical):* the **absence of association** between two cross-tabulated variables. The percentage distributions of the dependent variable within each category of the independent variable are **identical**.

Hypothesis Testing with Chi-Square

Chi-square follows five steps:

1. Making assumptions (**random sampling**)
2. Stating the research and null hypotheses
3. Selecting the sampling distribution and specifying the test statistic
4. Computing the test statistic
5. Making a decision and interpreting the results

The Assumptions

- The chi-square test requires **no assumptions** about the **shape of the population distribution** from which the sample was drawn.

- However, like all inferential techniques it assumes **random sampling**.

Stating Null and Alternative Hypotheses

- The **null hypothesis** (H_0) states that **no association exists** between the two cross-tabulated variables in the population, and therefore the variables are **statistically independent**.
- The **Alternative hypothesis** (H_1) proposes that the two variables are **related** in the population.

The Chi-square Test-Goodness of Fit

- In addition to being used to test a single variance, the chi-square statistic can be used to see whether a frequency distribution fits a specific pattern.
- For example, to meet customer demands, a manufacturer of running shoes may wish to see whether buyers show a preference for a specific style. A traffic engineer may wish to see whether accidents occur more often on some days than on others, so that she can increase police patrols accordingly.
- An emergency service may want to see whether it receives more calls at certain times of the day than at others, so that it can provide adequate staffing.
- When you are testing to see whether a frequency distribution fits a specific pattern, you can use the chi-square **goodness-of-fit test.**

The Chi-square Test-Goodness of Fit-Procedure

The Chi-Square Goodness-of-Fit Test

Step 1 State the hypotheses and identify the claim.

Step 2 Find the critical value. The test is always right-tailed.

Step 3 Compute the test value.

Find the sum of the values.

$$\chi^2 = \sum \frac{(Oi - Ei)^2}{Ei}$$

Step 4 Make the decision.

Step 5 Summarize the results.

The Chi-square Test-Goodness of Fit-Procedure

- When there is perfect agreement between the observed and the expected values, Chi-square=0. Also, chi-square can never be negative.

“Ho: Good fit” and

“H1: Not a good fit” mean that chi-square will be small in the first case and large in the second case.

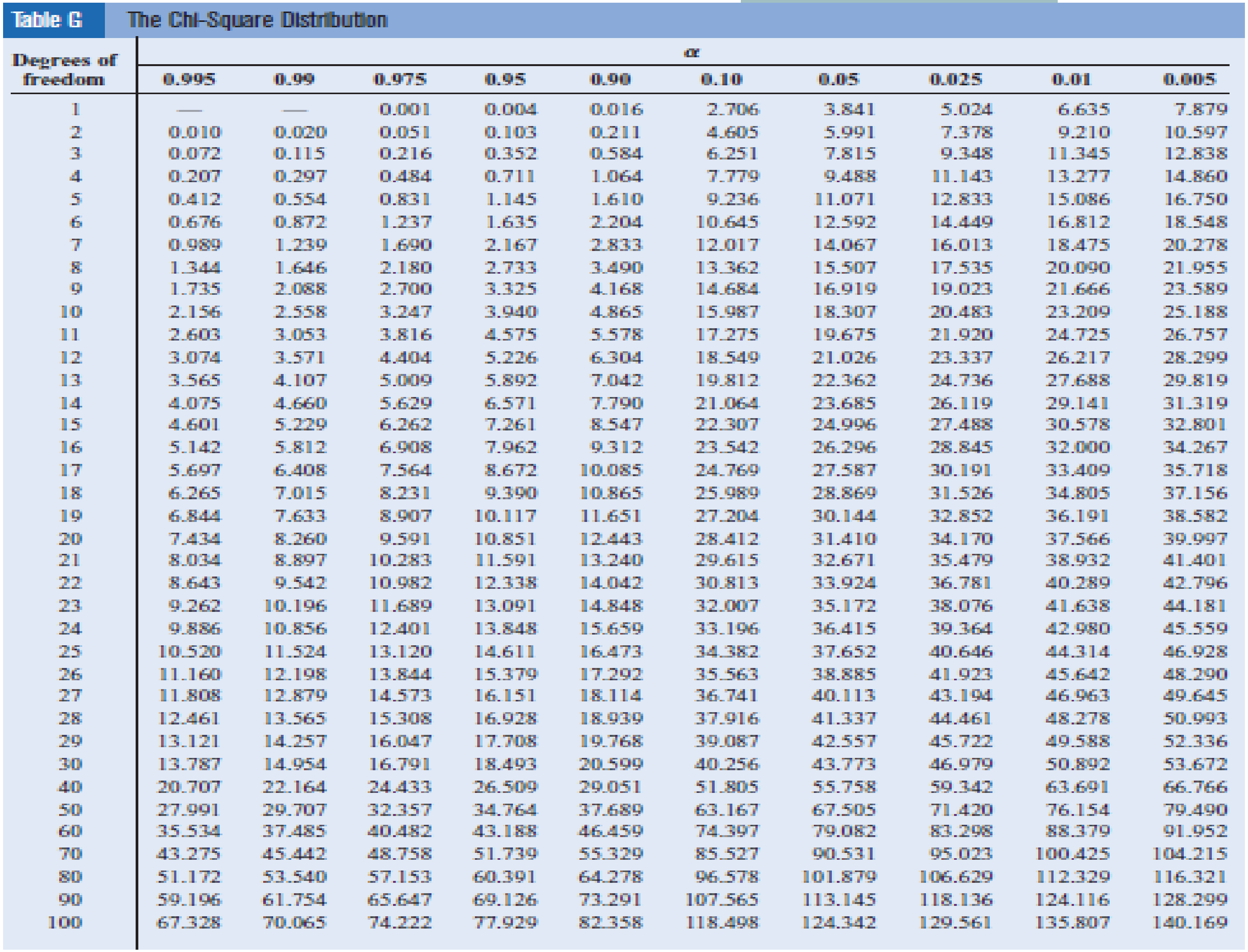


Table G The Chi-Square Distribution

Degrees of freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	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	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.262	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Goodness of Fit-Application-1

Suppose as a market analyst you wished to see whether consumers have any preference among five flavors of a new fruit soda. A sample of 100 people provided these data. Is there enough evidence to reject the claim that there is no preference in the selection of fruit soda flavors? Let α=0.05.

Soda	Cherry	S Berry	Orange	Lime	Grape
Freq	32	28	16	14	10

Goodness of Fit-Solution

If there were no preference, you would expect each flavor to be selected with equal frequency. In this case, the equal frequency is (100/5)=20. That is, approximately 20 people would select each flavor.

Since the frequencies for each flavor were obtained from a sample, these actual frequencies are called the observed frequencies.

The frequencies obtained by calculation (as if there were no preference) are called the expected frequencies. A completed table for the test is shown..

Frequency	Cherry	S Berry	Orange	Lime	Grape
Observed	32	28	16	14	10
Expected	20	20	20	20	20

Goodness of Fit-Solution

Step 1 State the hypotheses and identify the claim.

H_0: Consumers show no preference for flavors (claim).

H_1: Consumers show a preference.

Step 2 Find the critical value. The degrees of freedom are 5-1=4, and α=0.05.

Hence, the critical value from Chi-square-Table value is 9.488.

Step 3 Compute the test value by subtracting the expected value from the corresponding observed value, squaring the result and dividing by the expected value, and finding the sum. The expected value for each category is 20, as shown previously.

Goodness of Fit-Solution

$$\chi^2 = \sum \frac{(Oi - Ei)^2}{Ei}$$

Frequency	Cherr y	S Berry	Orange	Lime	Grape	Total
Observed	32	28	16	14	10	100
Expected	20	20	20	20	20	100

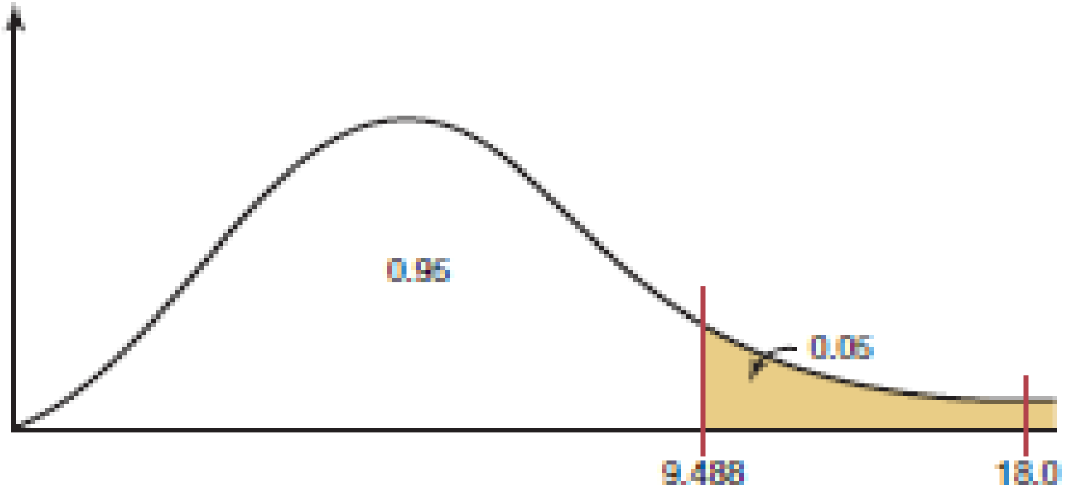
$$\chi^2 = \sum \frac{(Oi - Ei)^2}{Ei} = \frac{(32-20)^2}{20} + \frac{(28-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(14-20)^2}{20} + \frac{(10-20)^2}{20}$$

$$\chi^2 = \sum \frac{(Oi - Ei)^2}{Ei} = 18.00$$

Goodness of Fit- Solution

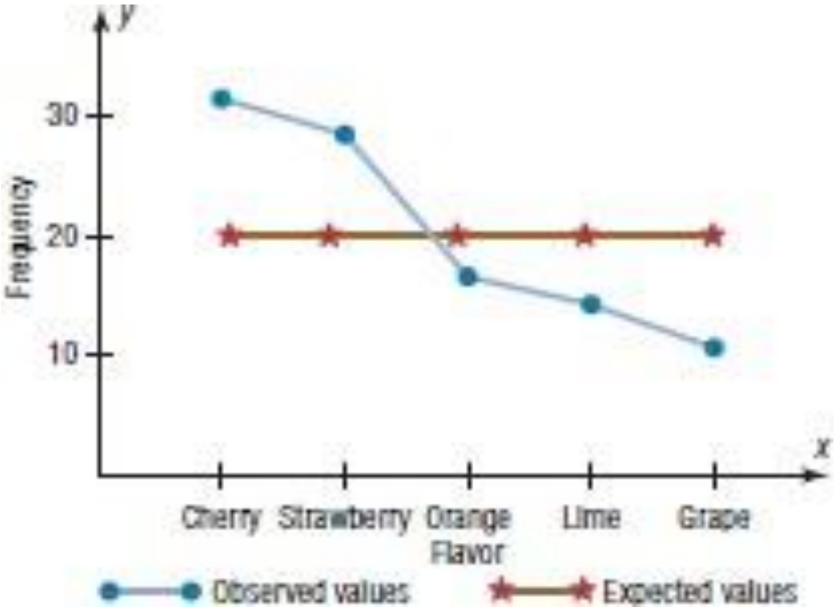
Step 4 Make the decision. The decision is to reject the null hypothesis, since 9.488<18.00, as shown in Table and Figure.

Step 5 Summarize the results. There is enough evidence to reject the claim that consumers show no preference for the flavors.



Goodness of Fit- Solution

To get some idea of why this test is called the goodness-of-fit test, examine graphs of the observed values and expected values. From the graphs, you can see whether the observed values and expected values are close together or far apart.



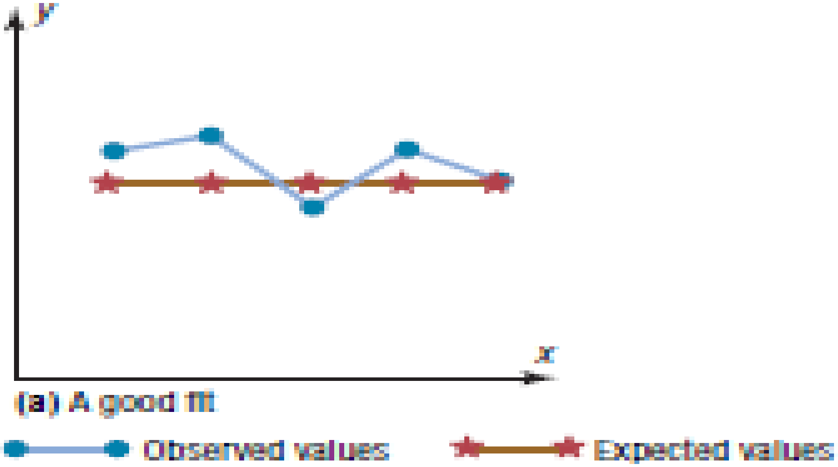
NOTE: When there is perfect agreement between the observed and the expected values, Chi-square=0. Also, Chi-value can never be negative.

Finally, the test is right-tailed because
"H_0: Good fit" and
"H_1: Not a good fit" mean that Chi-square will be small in the first case and large in the second case.

Goodness of Fit- Solution

When the observed values and expected values are close together, the chi-square test value will be small. Then the decision will be to not reject the null hypothesis—hence, there is “a good fit.” See Figure (a).

When the observed values and the expected values are far apart, the chi-square test value will be large. Then the null hypothesis will be rejected—hence, there is “not a good fit.” See Figure(b).



(a) A good fit

(b) Not a good fit

Goodness of Fit-Application-2

An Association surveyed retired senior executives who had returned to work. They found that after returning to work, 38% were employed by another organization, 32% were self-employed, 23% were either freelancing or consulting, and 7% had formed their own companies. To see if these percentages are consistent with those of County residents, a local researcher surveyed 300 retired executives who had returned to work and found that 122 were working for another company, 85 were self-employed, 76 were either freelancing or consulting, and 17 had formed their own companies. At α=0.10, test the claim that the percentages are the same for those people in Allegheny County.

Goodness of Fit-Solution

Step 1 State the hypotheses and identify the claim.

H_0: The retired executives who returned to work are distributed as follows: 38% are employed by another organization, 32% are self-employed, 23% are either freelancing or consulting, and 7% have formed their own companies (claim).

H_1: The distribution is not the same as stated in the null hypothesis.

Step 2 Find the critical value. Since α= 0.10 and the degrees of freedom are 4-1=3, the critical value is 6.251.

Step 3 Compute the test value. The expected values are computed as follows: 0.38*300=114; 0.32*300=96; 0.23*300=69; and 0.07*300=21

Frequency	Another Org.	Self-Emp.	Consulting	Own
Observed	122	85	76	17
Expected	114	96	69	21

Goodness of Fit-Solution

$$\chi^2 = \sum \frac{(Oi - Ei)^2}{Ei}$$

Frequency	Another Org.	Self-Emp.	Consulting	Own
Observed	122	85	76	17
Expected	114	96	69	21

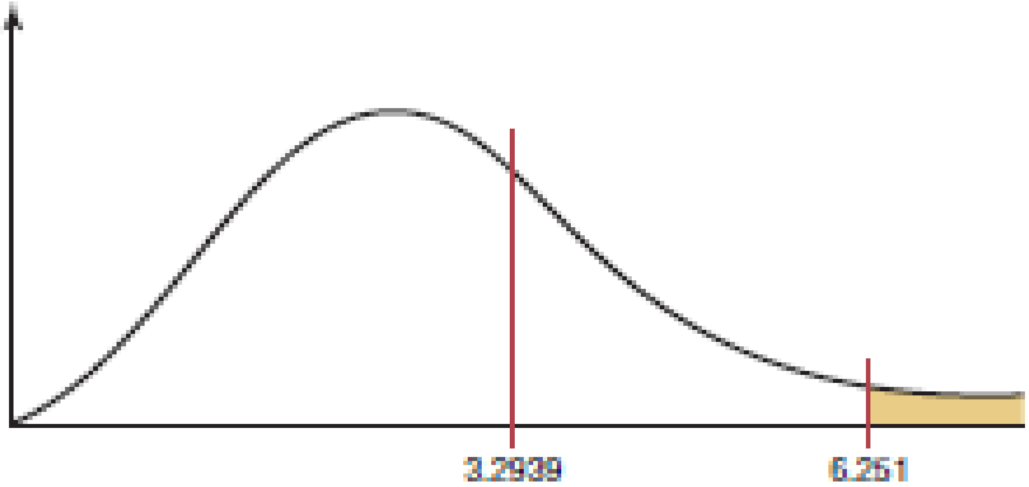
$$\chi^2 = \sum \frac{(Oi - Ei)^2}{Ei} = \frac{(122-114)^2}{114} + \frac{(85-96)^2}{96} + \frac{(66-69)^2}{69} + \frac{(17-21)^2}{21}$$

$$\chi^2 = \sum \frac{(Oi - Ei)^2}{Ei} = 3.2939$$

Goodness of Fit- Solution

Step 4 Make the decision. The decision is not to reject the null hypothesis, since 3.2939<6.251, as shown in Table and Figure.

Step 5 Summarize the results. There is not enough evidence to reject the claim. It can be concluded that the percentages are not significantly different from those given in the null hypothesis.



Goodness of Fit-Application-3

A researcher read that firearm-related deaths for people aged 1 to 18 were distributed as follows: 74% were accidental, 16% were homicides, and 10% were suicides. In her district, there were 68 accidental deaths, 27 homicides, and 5 suicides during the past year. At α=0.10, test the claim that the percentages are equal. **(Use Tab=4.605)**

Goodness of Fit-Application-4

A researcher wishes to see if the five ways (drinking decaffeinated beverages, taking a nap, going for a walk, eating a sugary snack, other) people use to combat midday. drowsiness are equally distributed among office workers. A sample of 60 office workers is selected, and the following data are obtained. At α=0.10 can it be concluded that there is no preference? Why would the results be of interest to an employer? **(Use Tab=7.779)**

Method	Beverage	Nap	Walk	Snack	Other
Number	21	16	10	8	5

Goodness of Fit-Application-5

In a recent study, the following percentages of Country retail car sales based on size were reported: 30.6% small, 45% midsize, 7.3% large, and 17.1% luxury.

A recent survey of sales in a particular county indicated that of 100 cars sold, 25 were small, 50 were midsize, 10 were large, and 15 were luxury cars. At the α=0.05 level of significance, can it be concluded that the proportions differ from those stated in the report? **(Use Tab=7.815)**

The Chi-square Test-Test of Normality

- The **chi-square goodness-of-fit test** can be used to test a variable to see if it is normally distributed. The null hypotheses are
- **H0**: The variable is normally distributed.
- **H1:** The variable is not normally distributed.
- It involves finding the expected frequencies for each class of a frequency distribution by using the standard normal distribution. Then the actual frequencies (i.e., observed frequencies) are compared to the expected frequencies, using the chi-square goodness-of-fit test.

The Chi-square Test-Test of Normality

- If the observed frequencies are close in value to the expected frequencies, the chi-square test value will be small, and the null hypothesis cannot be rejected. In this case, it can be concluded that the variable is approximately normally distributed.
- On the other hand, if there is a large difference between the observed frequencies and the expected frequencies, the chi-square test value will be larger, and the null hypothesis can be rejected. In this case, it can be concluded that the variable is not normally distributed.

Tests Using Contingency Tables

When data can be tabulated in table form in terms of frequencies, several types of hypotheses can be tested by using the chi-square test. Two such tests are the independence of variables test and the homogeneity of proportions test.

The test of independence of variables is used to determine whether two variables are independent of or related to each other when a single sample is selected.

The test of homogeneity of proportions is used to determine whether the proportions for a variable are equal when several samples are selected from different populations.

Both tests use the chi-square distribution and a contingency table, and the test value is found in the same way. The independence test will be explained first.

Chi-square test for Independence: Procedure

Step 1 State the hypotheses and identify the claim.
Step 2 Find the critical value in the right tail. Use Chi-square Table.
Step 3 Compute the test value. To compute the test value, first find the expected values. For each cell of the contingency table, use the formula to get the expected value.

$$Expected-value=\frac{(row-total)(column-total)}{Grang-total}$$

To find the test value, use the formula

$$\chi^2=\sum\frac{(Oi-Ei)^2}{Ei}$$

Step 4 Make the decision.
Step 5 Summarize the results.

Test for Independence -Application-1

Suppose a new postoperative procedure is administered to a number of patients in a large hospital. The researcher can ask the question, Do the doctors feel differently about this procedure from the nurses, or do they feel basically the same way?

Note that the question is not whether they prefer the procedure but whether there is a difference of opinion between the two groups.

To answer this question, a researcher selects a sample of nurses and doctors and tabulates the data in table form, as shown. Let α=0.05.

Groups	Prefer new Procedure	Prefer old Procedure	No Preference
Nurses	100	80	20
Doctors	50	120	30

Test for Independence -Solution

As the survey indicates, 100 nurses prefer the new procedure, 80 prefer the old procedure, and 20 have no preference; 50 doctors prefer the new procedure, 120 like the old procedure, and 30 have no preference.

Since the main question is whether there is a difference in opinion, the null hypothesis is stated as follows:

H_0: The opinion about the procedure is independent of the profession.

The alternative hypothesis is stated as follows:

H_1: The opinion about the procedure is dependent on the profession.

If the null hypothesis is not rejected, the test means that both professions feel basically the same way about the procedure and the differences are due to chance. If the null hypothesis is rejected, the test means that one group feels differently about the procedure from the other.

Test for Independence -Solution

Remember that rejection does not mean that one group favors the procedure and the other does not. Perhaps both groups favor it or both dislike it, but in different proportions.

To test the null hypothesis by using the chi-square independence test, you must compute the expected frequencies, assuming that the null hypothesis is true. These frequencies are computed by using the observed frequencies given in the table.

When data are arranged in table form for the chi-square independence test, the table is called a **contingency table**.

The table is made up of R rows and C columns. The table here has two rows and three columns.

Test for Independence -Solution

A contingency table is designated as an R X C (rows by columns) table. In this case, R=2 and C=3; hence, this table is a 2 X 3 contingency table. Each block in the table is called a cell and is designated by its row and column position. For example, the cell with a frequency of 80 is designated as O_{12}, or row 1, column 2. The cells are shown below.

Groups	Column-1	Column-2	Column-3	Total
Row-1	(100) O_{11}	(80)O_{12}	(20)O_{13}	**(200) R_1**
Row-2	(50) O_{21}	(120)O_{22}	(30)O_{23}	**(200) R_2**
Total	**(150) C_1**	**(200)C_2**	**(50)C_3**	**N=400**

Test for Independence -Solution

The degrees of freedom for any contingency table are (rows-1) times (columns-1) times; that is, d.f. =(R-1)(C-1). In this case, (2-1)(3-1)=(1)(2)=2.

The reason for this formula for d.f. is that all the expected values except one are free to vary in each row and in each column.

Using the previous table, you can compute the expected frequencies for each block (or cell), as shown next.

Groups	Prefer new Procedure	Prefer old Procedure	No Preference	Total
Nurses	100	80	20	200 (R1)
Doctors	50	120	30	200 (R2)
Total	150 (C1)	200 (C2)	50 (C3)	N=400

Test for Independence -Solution

For each cell, multiply the corresponding row sum by the column sum and divide by the grand total, to get the expected value:

$$E11 = \frac{(R1)(C1\ )}{N} = \frac{(200)(150\ )}{400} = 75$$

Groups	Prefer new Procedure	Prefer old Procedure	No Preference	Total
Nurses	E11=75	E12=100	E13=25	200 (R1)
Doctors	E21=75	E22=100	E23=25	200 (R2)
Total	150 (C1)	200 (C2)	50 (C3)	N=400

Test for Independence -Solution

The expected values can now be placed in the corresponding cells along with the observed values, as shown.

Groups	Prefer new Procedure	Prefer old Procedure	No Preference	Total

Nurses	O11=100
	E11=75
		O12=80
		E12=100
			O13=20
			E13=25
				200 (R1)

Doctors	O11=50
	E21=75
		O22=120
		E22=100
			O23=30
			E23=25
				200 (R2)

Total	150 (C1)	200 (C2)	50 (C3)	N=400

Test for Independence-Solution

$$\chi^2 = \sum \frac{(Oi - Ei)^2}{Ei}$$

Groups	Prefer new Procedure	Prefer old Procedure	No Preference	Total

Nurses	O11=100
	E11=75
		O12=80
		E12=100
			O13=20
			E13=25
				200 (R1)

Doctors	O11=50
	E21=75
		O22=120
		E22=100
			O23=30
			E23=25
				200 (R2)

Total	150 (C1)	200 (C2)	50 (C3)	N=400

$$\chi^2 = \sum \frac{(Oi - Ei)^2}{Ei} = \frac{(100-75)^2}{75} + \frac{(80-100)^2}{100} + \frac{(20-25)^2}{25} +$$

$$\frac{(50-75)^2}{75} + \frac{(120-100)^2}{100} + \frac{(30-25)^2}{25}$$

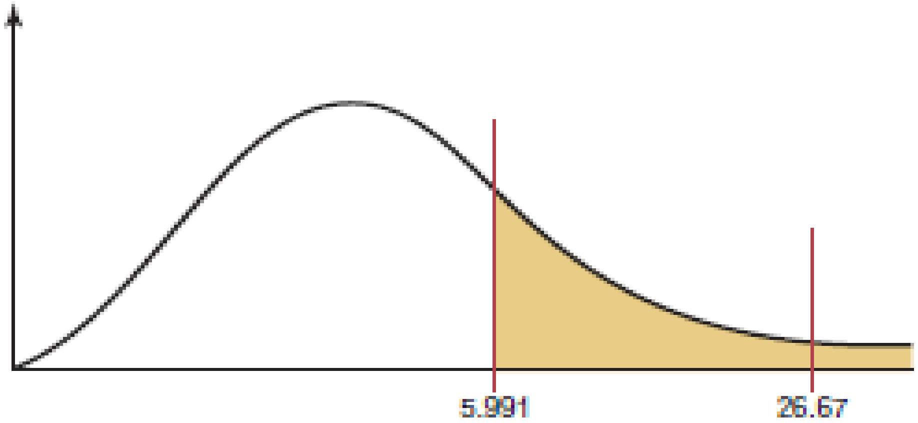
$$\chi^2 = \sum \frac{(Oi - Ei)^2}{Ei} = 26.67$$

Test for Independence-Solution

The final steps are to make the decision and summarize the results.

This test is always a right-tailed test, and the degrees of freedom are (R-1)(C -1)=(2-1)(3-1)=2. If α=0.05, the critical value from Chi-square Table is 5.991.

Hence, the decision is to reject the null hypothesis, since 26.67>5.991. See Figure



The conclusion is that there is enough evidence to support the claim that opinion is related to (dependent on) profession—that is, that the doctors and nurses differ in their opinions about the procedure.

Test for Independance-Application-2

A study is being conducted to determine whether the age of the customer is related to the type of movie he or she rents. A sample of renters gives the data shown here. At α= 0.10, is the type of movie selected related to the customer's age?.

	Type of movie		
Age	**Documentary**	**Comedy**	**Mystery**
12-20	14	9	8
21-29	15	14	9
30-38	9	21	39
39-47	7	22	17
48 and above	6	38	12

Test for Independance-Application-3

An instructor wishes to see if the way people obtain information is independent of their educational back-ground. A survey of 400 high school and college graduates yielded this information. At α= 0.05, test the claim that the way people obtain information is independent of their educational background.

Groups	Television	News Paper	Other sources
High School	159	90	51
College	27	42	31

Test for Independance-Application-4

The table below shows the number of students (in thousands) participating in various programs at both two-year and four-year institutions. At α= 0.05, can it be concluded that there is a relationship between program of study and type of institution?.

Groups	Two-year	Four-year
Agriculture and related sciences	36	52
Criminal justice	210	231
Foreign languages and literature	28	59
Mathematics and statistics	28	63

Test for Independance-Application-5

A book publisher wishes to determine whether there is a difference in the type of book selected by males and females for recreational reading. A random sample provides the data given here. At α= 0.05, test the claim that the type of book selected is independent of the gender of the individual.

	Type of book		
Gender	**Mystery**	**Romance**	**Self-help**
Male	243	201	191
Female	135	149	201

Chi-square test for Homogeneity

- The second chi-square test that uses a contingency table is called the homogeneity of proportions test.
- In this situation, samples are selected from several different populations, and the researcher is interested in determining whether the proportions of elements that have a common characteristic are the same for each population.
- The sample sizes are specified in advance, making either the row totals or column totals in the contingency table known before the samples are selected.

Chi-square test for Homogeneity: Procedure

Step 1 State the hypotheses and identify the claim.
Step 2 Find the critical value in the right tail. Use Chi-square Table.
Step 3 Compute the test value. To compute the test value, first find the expected values. For each cell of the contingency table, use the formula to get the expected value.

$$Expected-value = \frac{(row-total)(column-total)}{Grang-total}$$

To find the test value, use the formula

$$\chi^2 = \sum \frac{(Oi-Ei)^2}{Ei}$$

Step 4 Make the decision.
Step 5 Summarize the results.

Test for Homogeneity -Application-1

A researcher selected 100 passengers from each of 3 airlines and asked them if the airline had lost their luggage on their last flight. The data are shown in the table. At α=0.05, test the claim that the proportion of passengers from each airline who lost luggage on the flight is the same for each airline.

Groups	Airline-1	Airline-2	Airline-3
Yes	10	7	4
No	90	93	96

Test for Homogeneity -Solution

H_0: $P_1=P_2=P_3$

H_1: At least one mean differs from the other.

Test for Homogeneity -Solution

A contingency table is designated as an R X C (rows by columns) table. In this case, R=2 and C=3; hence, this table is a 2 X 3 contingency table. Each block in the table is called a cell and is designated by its row and column position. For example, the cell with a frequency of 07 is designated as O_{12}, or row 1, column 2. The cells are shown below.

Groups	Column-1	Column-2	Column-3	Total
Row-1	(10) O_{11}	(7)O_{12}	(4)O_{13}	**(21) R_1**
Row-2	(90) O_{21}	(93)O_{22}	(96)O_{23}	**(279) R_2**
Total	**(100) C_1**	**(100)C_2**	**(100)C_3**	**N=300**

Test for Homogeneity -Solution

The degrees of freedom for any contingency table are (rows-1) times (columns-1) times; that is, d.f. =(R-1)(C-1). In this case, (2-1)(3-1)=(1)(2)=2.

The reason for this formula for d.f. is that all the expected values except one are free to vary in each row and in each column.

Using the previous table, you can compute the expected frequencies for each block (or cell), as shown next.

Groups	Airline-1	Airline-2	Airline-3	Total
Yes	10	07	04	21 (R1)
No	90	93	96	279 (R2)
Total	100 (C1)	100 (C2)	100 (C3)	N=300

Test for Homogeneity -Solution

For each cell, multiply the corresponding row sum by the column sum and divide by the grand total, to get the expected value:

$$E11 = \frac{(R1)(C1)}{N} = \frac{(21)(100)}{300} = 7$$

Groups	Airline-1	Airline-2	Airline-3	Total
Yes	E11=7	E12=7	E13=7	21 (R1)
No	E21=93	E22=93	E23=93	279 (R2)
Total	100 (C1)	100 (C2)	100 (C3)	N=300

Test for Homogeneity -Solution

$$\chi^2 = \sum \frac{(Oi - Ei)^2}{Ei}$$

Groups	Airline-1	Airline-2	Airline-3	Total

Yes	O11=10
	E11=7
		O12=7
		E12=7
			O12=4
			E13=7
				21 (R1)

No	O21=90
	E21=93
		O22=93
		E22=93
			O23=96
			E23=93
				279 (R2)

Total	100 (C1)	100 (C2)	100 (C3)	N=300

$$\chi^2 = \sum \frac{(Oi - Ei)^2}{Ei} = \frac{(10-7)^2}{7} + \frac{(7-7)^2}{7} + \frac{(4-7)^2}{7} +$$

$$\frac{(90-93)^2}{93} + \frac{(93-93)^2}{93} + \frac{(93-93)^2}{93}$$

$$\chi^2 = \sum \frac{(Oi - Ei)^2}{Ei} = 2.765$$

Test for Homogeneity -Solution

The final steps are to make the decision and summarize the results.

This test is always a right-tailed test, and the degrees of freedom are (R-1)(C -1)=(2-1)(3-1)=2. If α=0.05, the critical value from Chi-square Table is 5.991.

Hence, the decision is not to reject the null hypothesis, since 2.765<5.991.

The conclusion There is not enough evidence to reject the claim that the proportions are equal.

Hence it seems that there is no difference in the proportions of the luggage lost by each airline.

Test for Homogeneity -Application-2

According to a recent survey, 59% of Americans aged 8 to 17 would prefer that their mother work outside the home, regardless of what she does now. A school district psychologist decided to select three samples of 60 students each in elementary, middle, and high school to see how the students in her district felt about the issue. At α=0.10, test the claim that the proportions of the students who prefer that their mother have a job are equal.

Groups	Elementary	Middle	High
Prefers mother work	29	38	51
Prefers mother not work	31	22	9

Test for Homogeneity -Application-3

A local college recently made the news by offering foreign language speaking dorm rooms to its students. When questioned at another school, 50 students from each class responded as shown. At α= 0.05, is there sufficient evidence to conclude that the proportions of students favoring foreign language speaking dorms are not the same for each class?

Groups	Freshmen	Sophomores	Seniors	Juniors
Yes	10	15	22	20
No	40	35	28	30

Chi-square Test for 2X2 Contingency Table

Group	Positive	Negative	Total
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	N=a+b+c+d

$$\chi^2 = \frac{N(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} \sim \chi^2(1)df$$

Chi-square test for 2X2 Contingency table: Procedure

Step 1 State the hypotheses and identify the claim.

Step 2 Find the critical value in the right tail. Use Chi-square Table.

Step 3 Compute the test value. To compute the test value of the contingency table, use the formula to get the test value.

$$\chi^2 = \frac{N(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} \sim \chi^2(1)df$$

Step 4 Make the decision.

Step 5 Summarize the results.

χ^2 Test of Independence Application-I

A researcher randomly selected a sample **286** sexually active individuals and collect information on their HIV status and History of STDs. At $\alpha=$**0.05** level, is there evidence of a **relationship**?

	HIV		
STDs Hx	No	Yes	Total
No	84	32	116
Yes	48	122	170
Total	132	154	286

Chi-square test for 2X2 Contingency table: Direct Solution

Ho: There is no relationship between STDx and HIV.
H1: There is a relationship between STDx and HIV.

The critical value at α=**0.05** level with 1 d.f is **3.841** .
Compute the test value. To compute the test value of the contingency table, use the formula to get the test value.

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \sim \chi^2(1)df$$

$$\chi^2 = \frac{286(84*122-48*32)^2}{(116)(170)(132)(154)} = 54.29 \sim \chi^2(1)df$$

.

Chi-square test for 2X2 Contingency table: Solution

The final steps are to make the decision and summarize the results. This test is always a right-tailed test, and the degrees of freedom is One (1). If α=0.05, the critical value from Chi-square Table is 3.841.

Hence, the decision is to reject the null hypothesis, since 54.29>3.841.

The conclusion There is enough evidence to reject the claim that the are not equal.

Hence it seems that there is evidence of a relationship.

Chi-square test for 2X2 Contingency table: Solution

	HIV				
	No		Yes		
STDs HX	Obs.	Exp.	Obs.	Exp.	Total
No	84	53.5	32	62.5	116
Yes	48	78.5	122	91.5	170
Total	132	132	154	154	286

$$\chi^2 = \sum_{\text{all cells}} \frac{[O_i - E_i]^2}{E_i}$$

$$= \frac{[84 - 53.5]^2}{53.5} + \frac{[32 - 62.5]^2}{62.5} + \cdots + \frac{[122 - 91.5]^2}{91.5} = 54.29$$

Chi-square test for 2X2 Contingency table: Solution

The final steps are to make the decision and summarize the results.

This test is always a right-tailed test, and the degrees of freedom are (R-1)(C -1)=(2-1)(2-1)=1. If α=0.05, the critical value from Chi-square Table is 3.841.

Hence, the decision is to reject the null hypothesis, since 54.29<3.841.

The conclusion There is enough evidence to reject the claim that the are not equal.

Hence it seems that there is evidence of a relationship.

χ^2 Test of Independence Application-II

A researcher randomly selected a sample **400** active individuals and collect information on their mobile status and History. At α=**0.05** level, is there evidence of a **relationship**?

Own Cell Telephone	Male	Female	Total
Yes	60	80	140
No	140	120	260
Total	200	200	400

Yates correction for continuity in a 2x2 contingency table:

In a 2x2 contingency table, the number of d.f. is (2-1)(2-1) = 1. If any one of Expected cell frequency is less than 5, then we use of pooling method for chi-square –test results with `0’‘d.f. (since 1 d.f. is lost in pooling) which is meaningless. In this case we apply a correction due to Yates, which is usually known a Yates correction for continuity.

Chi-square test for 2X2 Contingency table: Yates Correction Procedure

Step 1 State the hypotheses and identify the claim.

Step 2 Find the critical value in the right tail. Use Chi-square Table.

Step 3 Compute the test value. To compute the test value of the contingency table, use the formula to get the test value.

$$\chi^2 = \frac{N\{I(ad-bc)I-(N/2)\}^2}{(a+b)(c+d)(a+c)(b+d)} \sim \chi^2(1)df$$

Step 4 Make the decision.

Step 5 Summarize the results.

Chi-square test : Yates Correction-Application-I

Are the homicide rate and volume of gun sales related for a sample of 25 cities?

	HOMICIDE RATE		
GUN SALES	Low	High	Totals
High	8	5	13
Low	4	8	12
Totals	12	13	N = 25

- The bivariate table showing the relationship between homicide rate (columns) and gun sales (rows). This 2x2 table has 4 cells.

Chi-square test : Yates Correction-Application-II

HIV Infection

Hx of STDs

	yes	no	total
yes	3	7	10
no	5	10	15
total	8	17	

Is HIV Infection related to Hx of STDs in Sub Saharan African Countries? Test at 5% level.

“If you haven't observed it you don't know what you are talking about.”

THANK YOU...