Chapter 12:

THE CHI-SQUARE DISTRIBUTION AND THE ANALYSIS OF FREQUENCIES

THE MATHEMATICAL PROPERTIES OF THE CHI-SQUARE DISTRIBUTION

The chi-square distribution may be derived from normal distributions. Suppose that from a normally distributed random variable Y with mean μ and variance σ^2 we randomly and independently select samples of size $n = 1$. Each value selected may be transformed to the standard normal variable z by the familiar formula

$$z_i = \frac{y_i - \mu}{\sigma}$$

Each value of z may be squared to obtain z^2. When we investigate the sampling distribution of z^2, we find that it follows a chi-square distribution with 1 degree of freedom. That is,

$$\chi^2_{(1)} = \left(\frac{y - \mu}{\sigma}\right)^2 = z^2$$

Now suppose that we randomly and independently select samples of size $n = 2$ from the normally distributed population of Y values. Within each sample we may transform each

value of y to the standard normal variable z and square as before. If the resulting values of z^2 for each sample are added, we may designate this sum by

$$\chi^2_{(2)} = \left(\frac{y_1 - \mu}{\sigma}\right)^2 + \left(\frac{y_2 - \mu}{\sigma}\right)^2 = z_1^2 + z_2^2$$

since it follows the chi-square distribution with 2 degrees of freedom, the number of independent squared terms that are added together.

The procedure may be repeated for any sample size n. The sum of the resulting z^2 values in each case will be distributed as chi-square with n degrees of freedom. In general, then,

$$\chi^2_{(n)} = z_1^2 + z_2^2 + \cdots + z_n^2$$

follows the chi-square distribution with n degrees of freedom. The mathematical form of the chi-square distribution is as follows:

$$f(u) = \frac{1}{\left(\frac{k}{2} - 1\right)!} \frac{1}{2^{k/2}} u^{(k/2)-1} e^{-(u/2)}, \quad u > 0$$

Chi-Square Tests

- The Chi-Square Test evaluates the relationship between two variables.
- It is a nonparametric test that is performed on categorical (nominal or ordinal) data.

Types of Chi-Square Tests

As already noted, we make use of the chi-square distribution in this chapter in testing hypotheses where the data available for analysis are in the form of frequencies. These hypothesis testing procedures are discussed under the topics of tests of goodness-of-fit, tests of independence, and tests of homogeneity.

Observed Versus Expected Frequencies

The chi-square statistic is most appropriate for use with categorical variables, such as marital status, whose values are the categories married, single, widowed, and divorced. The quantitative data used in the computation of the test statistic are the frequencies associated with each category of the one or more variables under study. There are two sets of frequencies with which we are concerned, observed frequencies and expected frequencies. The observed frequencies are the number of subjects or objects in our sample that fall into the various categories of the variable of interest.

For example, if we have a sample of 100 hospital patients, we may observe that 50 are married, 30 are single, 15 are widowed, and 5 are divorced. Expected frequencies are the number of subjects or objects in our sample that we would expect to observe if some null hypothesis about the variable is true. For example, our null hypothesis might be that the four categories of marital status are equally represented in the population from which we drew our sample. In that case we would expect our sample to contain 25 married, 25 single, 25 widowed, and 25 divorced patients.

The Chi-Square Test Statistic

The test statistic for the chi-square tests we discuss in this chapter is

$$\chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

In this Equation, O_i is the observed frequency for the ith category of the variable of interest, and E_i is the expected frequency (given that H_0 is true) for the ith category.

The Decision Rule

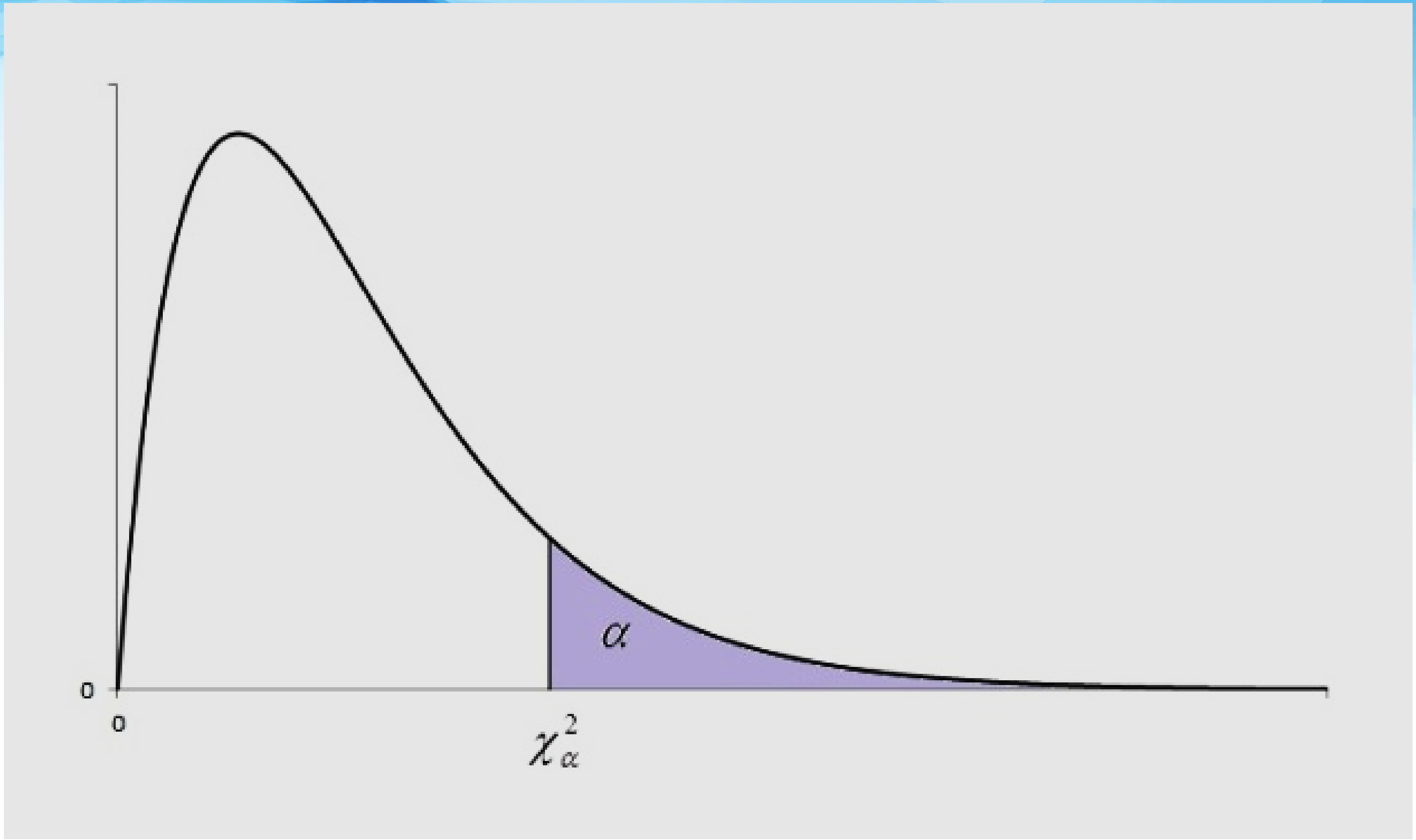
The Decision Rule The quantity $\sum[(O_i - E_i)^2/E_i]$ will be small if the observed and expected frequencies are close together and will be large if the differences are large.

The computed value of X^2 is compared with the tabulated value of χ^2 with $k - r$ degrees of freedom. The decision rule, then, is: Reject H_0 if X^2 is greater than or equal to the tabulated χ^2 for the chosen value of α.

Small Expected Frequencies Frequently in applications of the chi-square test the expected frequency for one or more categories will be small, perhaps much less than 1. In the literature the point is frequently made that the approximation of X^2 to χ^2 is not strictly valid when some of the expected frequencies are small. There is disagreement among writers, however, over what size expected frequencies are allowable before making some adjustment or abandoning χ^2 in favor of some alternative test. Some writers, especially the earlier ones, suggest lower limits of 10, whereas others suggest that all expected frequencies should be no less than 5. Cochran (4,5), suggests that for goodness-of-fit tests of unimodal distributions (such as the normal), the minimum expected frequency can be as low as 1. If, in practice, one encounters one or more expected frequencies less than 1, adjacent categories may be combined to achieve the suggested minimum. Combining reduces the number of categories and, therefore, the number of degrees of freedom. Cochran's suggestions appear to have been followed extensively by practitioners in recent years.

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value									
	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	9.260	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


0
0
α
χ^2_α

TESTS OF GOODNESS-OF-FIT

Say you wish to test that babies are born more or less often on certain days of the week. Being born on a Saturday has no measurement associated with it, but it is distinct from being born on a Tuesday.

Example: Are babies born with different proportions on certain days? Test at 5% level of significance where the observed data are presented below:

Days	Mon	Tues	Wed	Thur	Fri	Sat	Sun	Total
Births	110	124	104	94	112	72	84	700

$H_0: All\ proportions\ are\ equal\ \left(P_1 = P_2 = \cdots = P_7 = \frac{1}{7}\right)$

$H_1: At\ least\ one\ proportion\ differs$

$$DF = (k-1) = (7-1) = 6$$

$$\chi^2_{0.05,6} = 12.592$$

Now, calculate the test statistic:

$$\chi^2 = \frac{\sum (O_i - E_i)^2}{E_i}, \qquad E_i = \frac{f_c \times f_r}{n}$$

For example: How many births are expected on Monday

$$E_{Mon} = \frac{1}{7} \times 700 = 100$$

Observed								
Days	Mon	Tues	Wed	Thu	Fri	Sat	Sun	Total
Births	110	124	104	94	112	72	84	700

Expected								
Days	Mon	Tues	Wed	Thu	Fri	Sat	Sun	Total
Births	100	100	100	100	100	100	100	700

$$\chi^2 = \frac{\sum(O_i - E_i)^2}{E_i} = \frac{(110-100)^2}{100} + \frac{(124-100)^2}{100} + \cdots + \frac{(84-100)^2}{100} = 19.12$$

Results:

As $\chi^2 > 12.592 \longrightarrow reject\ H_0$

Interpretation:

There is sufficient evidence at 5% level of significance to suggest

the proportion of births on each day differs from $\frac{1}{7}$

TESTS OF INDEPENDENCE

We have already tested for relationships between quantitative variables such as (height VS weight). Linear correlation and regression assess relationships between quantitative variables, but what if we want to test for a relationship between gender and favorite color. Let us see the following example.

Example:

500 elementary school boys and girls are asked which is their favorite color: blue, green, or pink? Results are shown below

	Blue	Green	Pink	Total
Boys	100	150	20	300
Girls	20	30	180	200
Total	120	180	200	500

Using $\alpha = 0.05$, would you conclude that there is a relationship between gender and favorite color?

$$H_0: For\ the\ population\ of\ elementary\ school\ students, gender\ and\ favorite\ color\ are\ not\ related$$

$$H_1: For\ the\ population\ of\ elementary\ school\ students, gender\ and\ favorite\ color\ are\ related$$

$$DF = (rows - 1)(columns - 1) = (2 - 1)(3 - 1) = 2$$

$$\chi^2_{0.05,2} = 5.99147$$

Now, calculate the test statistic:

$$\chi^2 = \frac{\sum(O_i - E_i)^2}{E_i}, \qquad E_i = \frac{f_c \times f_r}{n}$$

For example: How many boys are expected to have chosen blue as their favorite color

$$(Boys, Blue) = \frac{120 \times 300}{500} = 72$$

Observed	Blue	Green	Pink	Total
Boys	100	150	20	**300**
Girls	20	30	180	**200**
Total	**120**	**180**	**200**	**500**

Expected	Blue	Green	Pink	Total
Boys	72	108	120	**300**
Girls	48	72	80	**200**
Total	**120**	**180**	**200**	**500**

$$\chi^2 = \frac{\sum(O_i - E_i)^2}{E_i} = \frac{(100-72)^2}{72} + \frac{(20-48)^2}{48} + \cdots + \frac{(180-80)^2}{80} = 276.389$$

Results:

If $\chi^2 > 5.99 \longrightarrow reject\ H_0$

Interpretation:

So, in the population, there is a relationship between gender and favorite color

TESTS OF HOMOGENEITY

This test determines if two or more populations have the same distribution of a single categorical variable. The test of homogeneity expands the test for a difference in two population proportions, which is the two-proportion Z-test we learned in Inference for Two Proportions. We use the two-proportion Z-test when the response variable has only two outcome categories and we are comparing two populations. We use the test of homogeneity if the response variable has two or more categories and we wish to compare two or more populations.

Example:

Narcolepsy is a disease involving disturbances of the sleep–wake cycle. Members of the German Migraine and Headache Society studied the relationship between migraine headaches in 96 subjects diagnosed with narcolepsy and 96 healthy controls. The results are shown in the following table. We wish to know if we may conclude, on the basis of these data, that the narcolepsy population and healthy population represented by the samples are not homogeneous with respect to migraine frequency. Use $\alpha = 0.05$

	Reported Migraine Headaches		
	Yes	No	Total
Narcoleptic subjects	21	75	96
Healthy controls	19	77	96
Total	40	152	192

Hypotheses:

H_0: The two populations are homogeneous with respect to migraine frequency.

H_A: The two populations are not homogeneous with respect to migraine frequency.

Test statistic:

$$\chi^2 = \frac{\sum (O_i - E_i)^2}{E_i}$$

Observed	**Reported Migraine Headaches**		
	Yes	No	Total
Narcoleptic subjects	21	75	96
Healthy controls	19	77	96
Total	40	152	192

Expected	**Reported Migraine Headaches**		
	Yes	No	Total
Narcoleptic subjects	20	76	96
Healthy controls	20	76	96
Total	40	152	192

Test statistic:

$$\chi^2 = \frac{\sum(O_i - E_i)^2}{E_i} = \frac{(21-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(75-76)^2}{76} + \frac{(77-76)^2}{76} = 0.126$$

Distribution of test statistic:

$$\chi^2{}_{0.05,(2-1).(2-1)} = \chi^2{}_{0.05,1} = 3.841$$

Decision rule:

Reject H_0 if the computed value of χ^2 is equal to or greater than 3.841.

Statistical decision:

Since .126 is less than the critical value of 3.841, we are unable to reject the null hypothesis.

Conclusion:

We conclude that the two populations may be homogeneous with respect to migraine frequency.

P- value: From the MINITAB output we see that $p = .722$ which is greater than $\alpha = 0.05$, So we accept H_0

Chi-Square Test

Expected counts are printed below observed counts

Rows: Narcolepsy   Columns: Migraine

	No	Yes	All
No	77	19	96
	76.00	20.00	96.00
Yes	75	21	96
	76.00	20.00	96.00
All	152	40	192
	152.00	40.00	192.00

Chi-Square = 0.126, DF = 1, P-Value = 0.722

Application in Excel 2019