11.1: Chi-Square Tests for Independence

Learning Objectives

- To understand what chi-square distributions are.
- To understand how to use a chi-square test to judge whether two factors are independent.

Chi-Square Distributions

As you know, there is a whole family of t-distributions, each one specified by a parameter called the degrees of freedom, denoted df. Similarly, all the chi-square distributions form a family, and each of its members is also specified by a parameter df, the number of degrees of freedom. Chi is a Greek letter denoted by the symbol χ and chi-square is often denoted by χ^2.

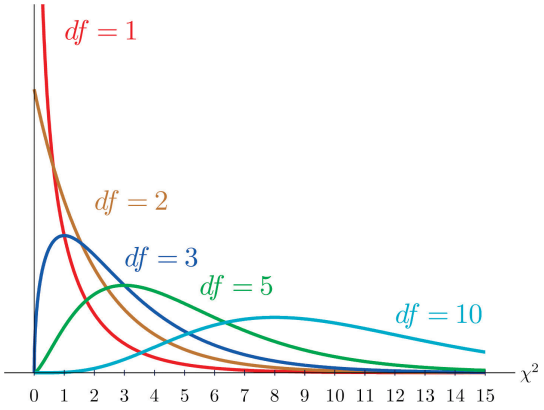


Figure 11.1.1: Many χ Distributions

Figure 11.1.1 shows several χ-square distributions for different degrees of freedom. A chi-square random variable is a random variable that assumes only positive values and follows a χ-square distribution.

Definition: critical value

The value of the chi-square random variable χ^2 with $df = k$ that cuts off a right tail of area c is denoted χ^2_c and is called a critical value (Figure 11.1.2).

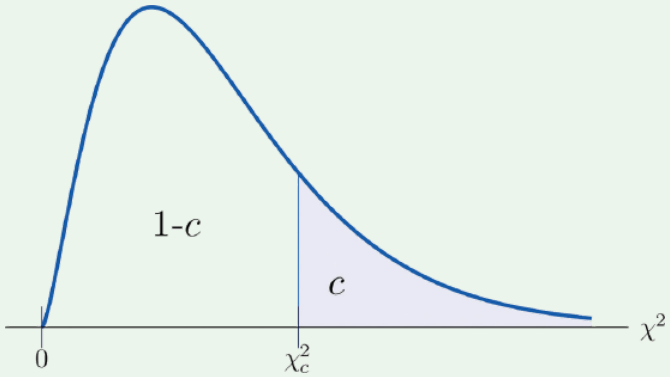
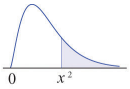


Figure 11.1.2: χ^2_c Illustrated

Figure 11.1.3 below gives values of χ^2_c for various values of c and under several chi-square distributions with various degrees of freedom.



Critical Values of Chi-Square Distributions

	x^2 Right-Tail Area									
df	0.995	0.99	0.975	0.95	0.90	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	10.597
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	11.070	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	28.300
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	14.041	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	43.195	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	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.458	15.655	17.539	19.281	21.434	41.422	44.985	48.232	52.191	55.003
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.815	17.074	19.047	20.867	23.110	43.745	47.400	50.725	54.776	57.648
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.192	18.509	20.569	22.465	24.797	46.059	49.802	53.203	57.342	60.275
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.586	19.96	22.106	24.075	26.492	48.363	52.192	55.668	59.893	62.883
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.996	21.426	23.654	25.695	28.196	50.660	54.572	58.120	62.428	65.476
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
41	21.421	22.906	25.215	27.326	29.907	52.949	56.942	60.561	64.950	68.053
42	22.138	23.650	25.999	28.144	30.765	54.090	58.124	61.777	66.206	69.336
43	22.859	24.398	26.785	28.965	31.625	55.230	59.304	62.990	67.459	70.616
44	23.584	25.148	27.575	29.787	32.487	56.369	60.481	64.201	68.710	71.893
45	24.311	25.901	28.366	30.612	33.350	57.505	61.656	65.410	69.957	73.166
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Figure 11.1.3: Critical Values of Chi-Square Distributions

Tests for Independence

Hypotheses tests encountered earlier in the book had to do with how the numerical values of two population parameters compared. In this subsection we will investigate hypotheses that have to do with whether or not two random variables take their values independently, or whether the value of one has a relation to the value of the other. Thus the hypotheses will be expressed in words, not mathematical symbols. We build the discussion around the following example.

There is a theory that the gender of a baby in the womb is related to the baby's heart rate: baby girls tend to have higher heart rates. Suppose we wish to test this theory. We examine the heart rate records of 40 babies taken during their mothers' last prenatal checkups before delivery, and to each of these 40 randomly selected records we compute the values of two random measures: 1) gender and 2) heart rate. In this context these two random measures are often called factors. Since the burden of proof is that heart rate and gender are related, not that they are unrelated, the problem of testing the theory on baby gender and heart rate can be formulated as a test of the following hypotheses:

H_0 : Baby gender and baby heart rate are independent

vs.

H_a : Baby gender and baby heart rate are not independent

The factor gender has two natural categories or levels: boy and girl. We divide the second factor, heart rate, into two levels, low and high, by choosing some heart rate, say 145 beats per minute, as the cutoff between them. A heart rate below 145 beats per minute will be considered low and 145 and above considered high. The 40 records give rise to a 2×2 *contingency table*. By adjoining row totals, column totals, and a grand total we obtain the table shown as Table 11.1.1. The four entries in boldface type are counts of observations from the sample of $n = 40$. There were 11 girls with low heart rate, 17 boys with low heart rate, and so on. They form the *core* of the expanded table.

Table 11.1.1: Baby Gender and Heart Rate

		Heart Rate		
		Low	High	Row Total
Gender	Girl	11	7	18
	Boy	17	5	22
Column Total		28	12	Total = 40

In analogy with the fact that the probability of independent events is the product of the probabilities of each event, if heart rate and gender were independent then we would expect the number in each core cell to be close to the product of the row total R and column total C of the row and column containing it, divided by the sample size n. Denoting such an expected number of observations E, these four expected values are:

- 1st row and 1st column: $E = (R \times C)/n = 18 \times 28/40 = 12.6$
- 1st row and 2nd column: $E = (R \times C)/n = 18 \times 12/40 = 5.4$
- 2nd row and 1st column: $E = (R \times C)/n = 22 \times 28/40 = 15.4$
- 2nd row and 2nd column: $E = (R \times C)/n = 22 \times 12/40 = 6.6$

We update Table 11.1.1 by placing each expected value in its corresponding core cell, right under the observed value in the cell. This gives the updated table Table 11.1.2.

Table 11.1.2: Updated Baby Gender and Heart Rate

		Heart Rate		
		Low	High	Row Total
Gender	Girl	$O = 11$ $E = 12.6$	$O = 7$ $E = 5.4$	$R = 18$
	Boy	$O = 17$ $E = 15.4$	$O = 5$ $E = 6.6$	$R = 22$
Column Total		$C = 28$	$C = 12$	$n = 40$

A measure of how much the data deviate from what we would expect to see if the factors really were independent is the sum of the squares of the difference of the numbers in each core cell, or, standardizing by dividing each square by the expected number in the cell, the sum $\sum (O-E)^2/E$. We would reject the null hypothesis that the factors are independent only if this number is large, so the test is right-tailed. In this example the random variable $\sum (O-E)^2/E$ has the chi-square distribution with one degree of freedom. If we had decided at the outset to test at the 10% level of significance, the critical value defining the rejection region would be, reading from Figure 11.1.3, $\chi^2_\alpha = \chi^2_{0.10} = 2.706$, so that the rejection region would be the interval $[2.706, \infty)$. When we compute the value of the standardized test statistic we obtain

$$\sum \frac{(O-E)^2}{E} = \frac{(11-12.6)^2}{12.6} + \frac{(7-5.4)^2}{5.4} + \frac{(17-15.4)^2}{15.4} + \frac{(5-6.6)^2}{6.6} = 1.231$$

Since $1.231 < 2.706$, the decision is not to reject H_0. See Figure 11.1.4 The data do not provide sufficient evidence, at the 10% level of significance, to conclude that heart rate and gender are related.

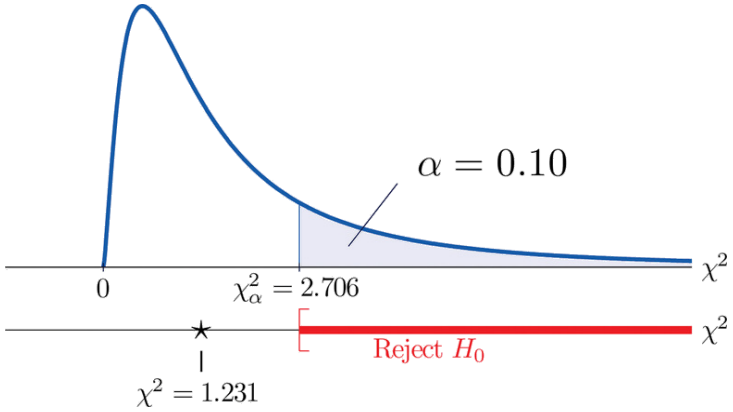


***Figure* 11.1.4**: *Baby Gender Prediction*

With this specific example in mind, now turn to the general situation. In the general setting of testing the independence of two factors, call them Factor 1 and Factor 2, the hypotheses to be tested are

$$H_0 : \text{The two factors are independent}$$
$$vs.$$
$$H_a : \text{The two factors are not independent}$$

As in the example each factor is divided into a number of categories or levels. These could arise naturally, as in the boy-girl division of gender, or somewhat arbitrarily, as in the high-low division of heart rate. Suppose Factor 1 has I levels and Factor 2 has J levels. Then the information from a random sample gives rise to a general $I \times J$ contingency table, which with row totals, column totals, and a grand total would appear as shown in Table 11.1.3. Each cell may be labeled by a pair of indices (i, j). O_{ij} stands for the observed count of observations in the cell in row i and column j, R_i for the i^{th} row total and C_j for the j^{th} column total. To simplify the notation we will drop the indices so Table 11.1.3becomes Table 11.1.4. Nevertheless it is important to keep in mind that the Os, the Rs and the Cs, though denoted by the same symbols, are in fact different numbers.

Table 11.1.3: General Contingency Table

		Factor 2 Levels					
		1	$\cdots$	j	$\cdots$	J	Row Total
Factor 1 Levels	1	O_{11}	$\cdots$	O_{1j}	$\cdots$	O_{1J}	R_1
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	i	O_{i1}	$\cdots$	O_{ij}	$\cdots$	O_{iJ}	R_i
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	I	O_{I1}	$\cdots$	O_{Ij}	$\cdots$	O_{IJ}	R_I
Column Total		C_1	$\cdots$	C_j	$\cdots$	C_J	n

Table 11.1.4: Simplified General Contingency Table

		Factor 2 Levels					
		1	$\cdots$	j	$\cdots$	J	Row Total
Factor 1 Levels	1	O	$\cdots$	O	$\cdots$	O	R
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	i	O	$\cdots$	O	$\cdots$	O	R
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

		Factor 2 Levels					
		1	$\cdots$	j	$\cdots$	J	Row Total
	I	O	$\cdots$	O	$\cdots$	O	R
Column Total		C	$\cdots$	C	$\cdots$	C	n

As in the example, for each core cell in the table we compute what would be the expected number E of observations if the two factors were independent. E is computed for each core cell (each cell with an O in it) of Table 11.1.4 by the rule applied in the example:

$$E = R \times Cn$$

where R is the row total and C is the column total corresponding to the cell, and n is the sample size

After the expected number is computed for every cell, Table 11.1.4 is updated to form Table 11.1.5 by inserting the computed value of E into each core cell.

Table 11.1.5: Updated General Contingency Table

		Factor 2 Levels					
		1	$\cdots$	j	$\cdots$	J	Row Total
Factor 1 Levels	1	O E	$\cdots$	O E	$\cdots$	O E	R
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	i	O E	$\cdots$	O E	$\cdots$	O E	R
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	I	O E	$\cdots$	O E	$\cdots$	O E	R
Column Total		C	$\cdots$	C	$\cdots$	C	n

Here is the test statistic for the general hypothesis based on Table 11.1.5, together with the conditions that it follow a chi-square distribution.

Test Statistic for Testing the Independence of Two Factors

$$\chi^2 = \sum (O - E)^2 E$$

where the sum is over all core cells of the table.

If

1. the two study factors are independent, and
2. the observed count O of each cell in Table 11.1.5 is at least 5,

then χ^2 approximately follows a chi-square distribution with $df = (I-1) \times (J-1)$ degrees of freedom.

The same five-step procedures, either the critical value approach or the p-value approach, that were introduced in Section 8.1 and Section 8.3 are used to perform the test, which is always right-tailed.

Example 11.1.1

A researcher wishes to investigate whether students' scores on a college entrance examination (CEE) have any indicative power for future college performance as measured by GPA. In other words, he wishes to investigate whether the factors CEE and GPA are independent or not. He randomly selects $n = 100$ students in a college and notes each student's score on the entrance examination and his grade point average at the end of the sophomore year. He divides entrance exam scores into two levels and grade point averages into three levels. Sorting the data according to these divisions, he forms the contingency table shown as Table 11.1.6, in which the row and column totals have already been computed.

Table 11.1.6: CEE *versus* GPA *Contingency Table*

		GPA			
		< 2.7	2.7 to 3.2	> 3.2	Row Total
CEE	< 1800	35	12	5	52
	≥ 1800	6	24	18	48
Column Total		41	36	23	Total = 100

Test, at the 1% level of significance, whether these data provide sufficient evidence to conclude that CEE scores indicate future performance levels of incoming college freshmen as measured by GPA.

Solution

We perform the test using the critical value approach, following the usual five-step method outlined at the end of Section 8.1.

- **Step 1**. The hypotheses are

$$H_0 : \text{CEE and GPA are independent factors}$$
$$vs.$$
$$H_a : \text{CEE and GPA are not independent factors}$$

- **Step 2**. The distribution is chi-square.
- **Step 3**. To compute the value of the test statistic we must first computed the expected number for each of the six core cells (the ones whose entries are boldface):
  - 1st row and 1st column: $E = (R \times C)/n = 41 \times 52/100 = 21.32$
  - 1st row and 2nd column: $E = (R \times C)/n = 36 \times 52/100 = 18.72$
  - 1st row and 3rd column: $E = (R \times C)/n = 23 \times 52/100 = 11.96$
  - 2nd row and 1st column: $E = (R \times C)/n = 41 \times 48/100 = 19.68$
  - 2nd row and 2nd column: $E = (R \times C)/n = 36 \times 48/100 = 17.28$
  - 2nd row and 3rd column: $E = (R \times C)/n = 23 \times 48/100 = 11.04$

Table 11.1.6 is updated to Table 11.1.6.

Table 11.1.7: Updated CEE versus GPA Contingency Table

		GPA			
		< 2.7	2.7 to 3.2	> 3.2	Row Total
CEE	< 1800	$O = 35$ $E = 21.32$	$O = 12$ $E = 18.72$	$O = 5$ $E = 11.96$	$R = 52$
	≥ 1800	$O = 6$ $E = 19.68$	$O = 24$ $E = 17.28$	$O = 18$ $E = 11.04$	$R = 48$
Column Total		$C = 41$	$C = 36$	$C = 23$	$n = 100$

The test statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$
$$= \frac{(35-21.32)^2}{21.32} + \frac{(12-18.72)^2}{18.72} + \frac{(5-11.96)^2}{11.96} + \frac{(6-19.68)^2}{19.68} + \frac{(24-17.28)^2}{17.28} + \frac{(18-11.04)^2}{11.04}$$
$$= 31.75$$

- **Step 4**. Since the CEE factor has two levels and the GPA factor has three, $I = 2$ and $J = 3$. Thus the test statistic follows the chi-square distribution with $df = (2-1) \times (3-1) = 2$ degrees of freedom.

  Since the test is right-tailed, the critical value is $\chi^2_{0.01}$. Reading from Figure 7.1.6 "Critical Values of Chi-Square Distributions", $\chi^2_{0.01} = 9.210$, so the rejection region is $[9.210, \infty)$.

- **Step 5**. Since $31.75 > 9.21$ the decision is to reject the null hypothesis. See Figure 11.1.5. The data provide sufficient evidence, at the 1% level of significance, to conclude that CEE score and GPA are not independent: the entrance exam score has predictive power.

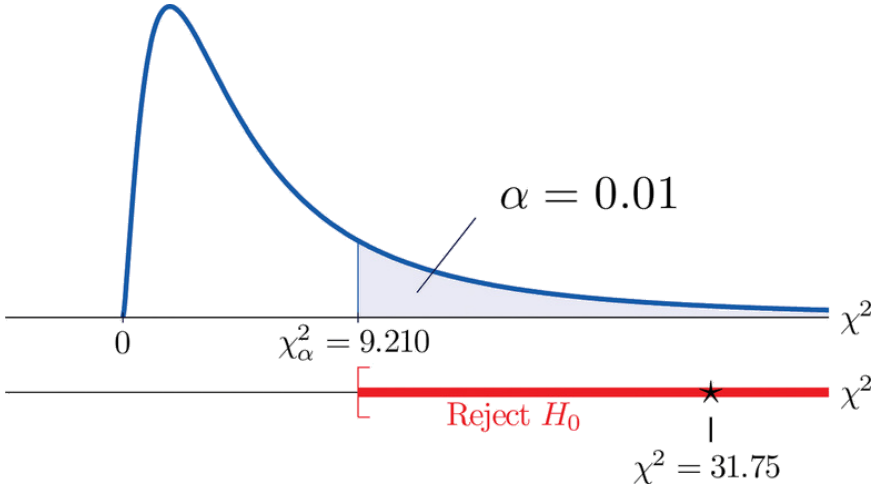


Figure 11.1.5: "Example 11.1.1"

Key Takeaway

- Critical values of a chi-square distribution with degrees of freedom df are found in Figure 7.1.6.
- A chi-square test can be used to evaluate the hypothesis that two random variables or factors are independent.