6.9: Chi-square distribution

As noted earlier, the **normal deviate** or **Z score** can be viewed as randomly sampled from the **standard normal distribution**. The chi-square distribution describes the probability distribution of the squared standardized normal deviates with **degrees of freedom**, df, equal to the number of samples taken. (The number of independent pieces of information needed to calculate the estimate, see Ch. 8.) We will use the chi-square distribution to test statistical significance of categorical variables in goodness of fit tests and contingency table problems.

The equation of the chi-square is

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(f_i - \hat{f}_i\right)^2}{\hat{f}_i}$$

where k is the number of groups or categories, from 1 to k, and f_i is the **observed frequency** and $\hat{f}_i$ is the **expected frequency** for the k^{th} category. We call the result of this calculation the chi-square test statistic. We evaluate how often that value or greater of a test statistic will occur by applying the chi-square distribution function. Graphs to show chi-square distribution for degrees of freedom equal to 1 through 5, 10, and 20 (Fig. $6.9.1$).

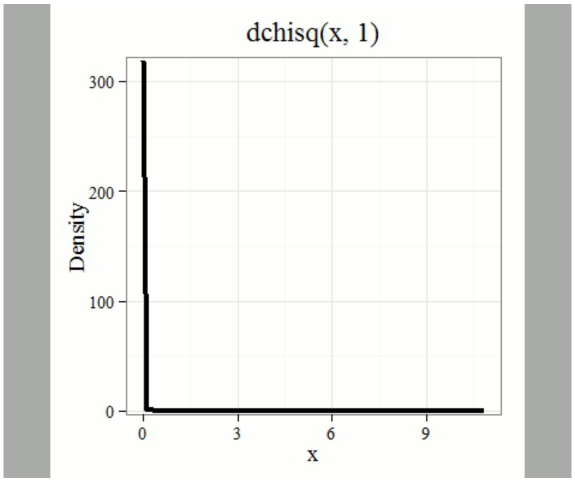

Figure $6.9.1$: Animated GIF of plots of chi-square distribution over a range of degrees of freedom.

Note that the distribution is asymmetric, particularly at low **degrees of freedom**. Thus, tests using the chi-square are one-tailed (Fig. $6.9.2$).

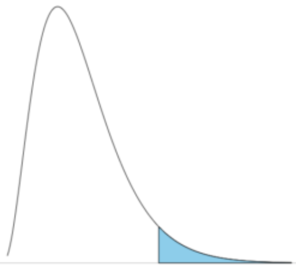

Figure $6.9.2$: The test of the chi-square is typically one-tailed. In this case, the highlighted area shows the probability of values greater than the critical value.

By convention in the Null Hypothesis Significance Testing protocol (NHST), we compare the test statistic to a **critical value**. The critical value is defined as the value of the test statistic — the cutoff boundary between **statistical significance** and insignificance — that occurs at the Type I error rate, which is typically set to 5%. The interpretation of the result is as follows: after calculating a test statistic, we can judge significance of the results relative to the null hypothesis expectation. If our test statistics is greater than the critical value, then the **p-value** of our results are less than 5% (R will report an exact p-value for the test statistic). You are not expected to be able to follow this logic just yet — rather, we teach it now as a sort of mechanical understanding to develop in the NHST tradition. The justification for this approach to testing of statistical significance is developed in Chapter 8. A portion of the critical values of the chi-square distribution are shown in Figure $6.9.3$.

α(1)	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
DF/1	1.323	2.706	3.841	5.024	6.635	7.879	9.141	10.828	12.116
2	2.773	4.605	5.991	7.378	9.210	10.597	11.983	13.816	15.202
3	4.108	6.251	7.815	9.348	11.345	12.838	14.320	16.266	17.730
4	5.385	7.779	9.488	11.143	13.277	14.860	16.424	18.467	19.997
5	6.626	9.236	11.070	12.833	15.086	16.750	18.386	20.515	22.105
6	7.841	10.645	12.592	14.449	16.812	18.548	20.249	22.458	24.103
7	9.037	12.017	14.067	16.013	18.475	20.278	22.040	24.322	26.018
8	10.219	13.362	15.507	17.535	20.090	21.955	23.774	26.124	27.868
9	11.389	14.684	16.919	19.023	21.666	23.589	25.462	27.877	29.666
10	12.549	15.987	18.307	20.483	23.209	25.188	27.112	29.588	31.420

Figure 6.9.3: Portion of the table of some critical values of chi-square distribution, one tailed (right-tailed or "upper" portion of distribution).

See Appendix for a complete chi-square table.

Example

Professor Hermon Bumpus of Brown University in Providence, Rhode Island, received 136 House Sparrows (*Passer domesticus*) after a severe winter storm 1 February 1898. The birds were collected from the ground; 72 of the birds survived, 64 did not (Table 6.9.1). Bumpus made several measures of morphology on the birds and the data set has served as a classical example of Natural Selection (Chicago Field Museum). We'll look at this data set when we introduce Linear Regression.

Table 6.9.1. Survival statistics of Bumpus House sparrows.

	Yes	No
Female	21	28
Male	51	36

Was there a survival difference between male and female House Sparrows? This is a classic contingency table analysis, something we will at length in Chapter 9. For now, we report the Chi-square test statistic for this test was 3.1264 and the test had one degree of freedom. What is the critical value of the chi-square distribution at 5% and one degree of freedom?. Typically we would simply use R to look this up

```
qchisq(c(0.05), df=1, lower.tail=FALSE)
```

But we can also get this from the table of critical values (Fig. 6.9.4). Simply select the row based on the degrees of freedom for the test then scan to the column with the appropriate significance level, again, typically 5% (0.05).

α(1)	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
DF/1	1.323	2.706	3.841	5.024	6.635	7.879	9.141	10.828	12.116
2	2.773	4.605	5.991	7.378	9.210	10.597	11.983	13.816	15.202
3	4.108	6.251	7.815	9.348	11.345	12.838	14.320	16.266	17.730
4	5.385	7.779	9.488	11.143	13.277	14.860	16.424	18.467	19.997
5	6.626	9.236	11.070	12.833	15.086	16.750	18.386	20.515	22.105
6	7.841	10.645	12.592	14.449	16.812	18.548	20.249	22.458	24.103
7	9.037	12.017	14.067	16.013	18.475	20.278	22.040	24.322	26.018
8	10.219	13.362	15.507	17.535	20.090	21.955	23.774	26.124	27.868
9	11.389	14.684	16.919	19.023	21.666	23.589	25.462	27.877	29.666
10	12.549	15.987	18.307	20.483	23.209	25.188	27.112	29.588	31.420

Figure 6.9.4: Portion of the chi-square distribution which shows how to find critical value of the chi-square distribution.

For 1 degree of freedom at 5% significance, the critical value is 3.841. Back to our hypothesis: Did male and female survival differ in the Bumpus data set? Following the NHST logic, if the test statistic value (e.g., 3.1264) is greater than the critical value (3.841), then we would reject the null hypothesis. For this example, we would conclude no statistical difference between male and female survival because the test statistic was smaller than the critical value. How likely are these results due to chance? That's where the p-value comes in. Our test statistic value falls between 5% and 10% ($2.706 < 3.1264 < 3.841$). In order to get the actual p-value of our test statistic we would need to use R.

R code

Given a chi-square test statistic, you can use R to calculate the probability of that value against the null hypothesis. At the R prompt, enter:

```
pchisq(c(3.1264), df=1, lower.tail=FALSE)
```

The R output is

```
[1] 0.07703368
```

Because we are using R Commander, simply select the command by following the menu options.

Rcmdr: Distributions → Continuous distributions → Chi-squared distribution → Chi-squared probabilities …

Enter the chi-square value and degrees of freedom (Fig. 6.9.5).

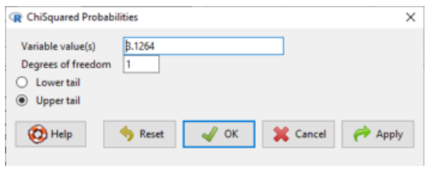

Figure 6.9.5: Screenshot of input box in Rcmdr for Chi-square probability values.

Questions

1. What happens to the shape of the chi-square distribution as degrees of freedom are increased from 1 to 5 to 20 to 100?

Be able to answer the following questions using the Chi-square table or using Rcmdr:

2. For probability $\alpha = 5\%$, what is the critical value of the chi-square distribution (upper tail)?
3. The value of the chi-square test statistic is given as 12. With 3 degrees of freedom, what is the approximate probability of this value or greater from the chi-square distribution?
