Lecture 8: Chi square tests

One Way Table

Mars, Incorporated makes milk chocolate candies. Here’s what the company’s Consumer Affairs Department says about the color distribution of its M&M’S Milk Chocolate Candies: ***On average, the new mix of colors of M&M’S Milk Chocolate Candies will contain 13 percent of each of browns and reds, 14 percent yellows, 16 percent greens, 20 percent oranges and 24 percent blues.***

The one-way table below summarizes the data from a sample bag of M&M’S Milk Chocolate Candies. In general, one-way tables display the distribution of a categorical variable for the individuals in a sample.

Color	Blue	Orange	Green	Yellow	Red	Brown	**Total**
Count	9	8	12	15	10	6	**60**
	24%	20%	16%	14%	13%	13%	

Color	Blue	Orange	Green	Yellow	Red	Brown	**Total**
Count	9	8	12	15	10	6	**60**

The sample proportion of blue M & M's is $\hat{p} = \frac{9}{60} = 0.15$.

Since the company claims that 24% of all M&M'S Milk Chocolate Candies are blue, we might believe that something fishy is going on. We could use the one-sample *z* test for a proportion from Chapter 8 to test the hypotheses

$$H_0: p = 0.24$$
$$H_a: p \neq 0.24$$

where *p* is the true population proportion of blue M&M'S. We could then perform additional significance tests for each of the remaining colors.

However, performing a one-sample *z* test for each proportion would be pretty inefficient and would lead to the problem of multiple comparisons.

For that, we need a new kind of significance test, called a **chi-square goodness-of-fit test.**

The null hypothesis in a chi-square goodness-of-fit test should state a claim about the distribution of a single categorical variable in the population of interest. In our example, the appropriate null hypothesis is

H_0: The company's stated color distribution forM&M'S Milk Chocolate Candies is correct.

The alternative hypothesis in a chi-square goodness-of-fit test is that the categorical variable does *not* have the specified distribution. In our example, the alternative hypothesis is

H_a: The company's stated color distribution forM&M'S Milk Chocolate Candies is not correct.

Comparing Observed and Expected Counts

We can also write the hypotheses in symbols as

$$H_0\text{: } p_{blue} = 0.24,\ p_{orange} = 0.20,\ p_{green} = 0.16,$$
$$p_{yellow} = 0.14,\ p_{red} = 0.13,\ p_{brown} = 0.13,$$
$$H_a\text{: At least one of the } p_i\text{'s is incorrect}$$

where p_{color} = the true population proportion of M&M'S Milk Chocolate Candies of that color.

The idea of the chi-square goodness-of-fit test is this: we compare the **observed counts** from our sample with the counts that would be expected if H_0 is true. The more the observed counts differ from the **expected counts**, the more evidence we have against the null hypothesis.

In general, the expected counts can be obtained by multiplying the proportion of the population distribution in each category by the sample size.

Example: Computing Expected Counts

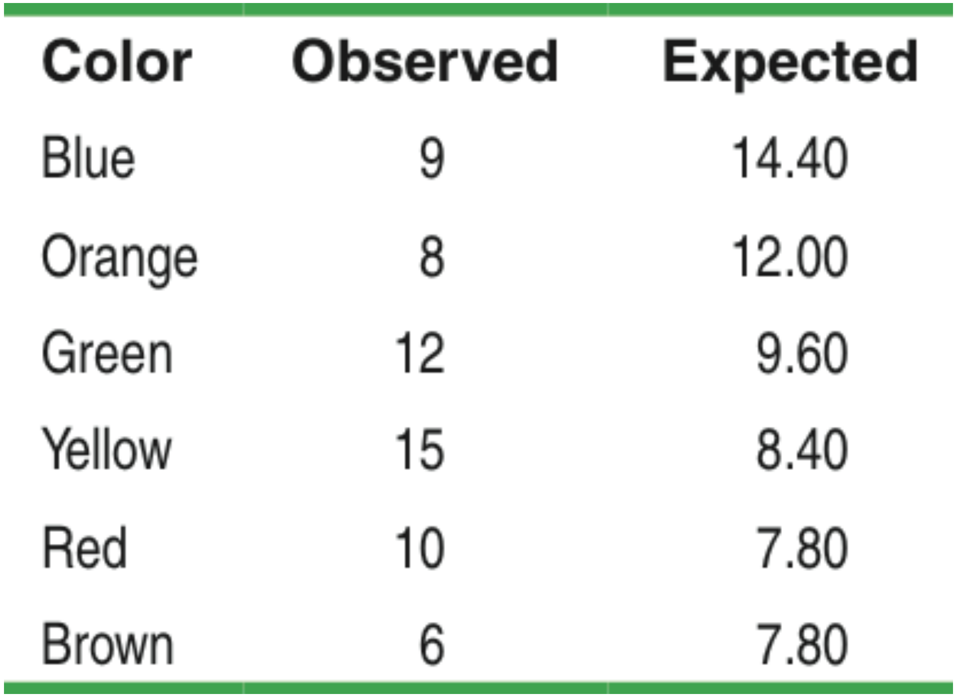
A sample bag of M&M’s milk Chocolate Candies contained 60 candies. Calculate the expected counts for each color.

Assuming that the color distribution stated by Mars, Inc., is true, 24% of all M&M’s milk Chocolate Candies produced are blue.

For random samples of 60 candies, the average number of blue M&M’s should be (0.24)(60) = 14.40. This is our expected count of blue M&M’s.

Using this same method, we can find the expected counts for the other color categories:

Orange:	**(0.20)(60) = 12.00**
Green:	**(0.16)(60) = 9.60**
Yellow:	**(0.14)(60) = 8.40**
Red:	**(0.13)(60) = 7.80**
Brown:	**(0.13)(60) = 7.80**



Color	Observed	Expected
Blue	9	14.40
Orange	8	12.00
Green	12	9.60
Yellow	15	8.40
Red	10	7.80
Brown	6	7.80

To see if the data give convincing evidence against the null hypothesis, we compare the observed counts from our sample with the expected counts assuming H_0 is true. If the observed counts are far from the expected counts, that's the evidence we were seeking.

The statistic we use to make the comparison is the **chi-square statistic.**

Definition:

The **chi-square statistic** is a measure of how far the observed counts are from the expected counts. The formula for the statistic is

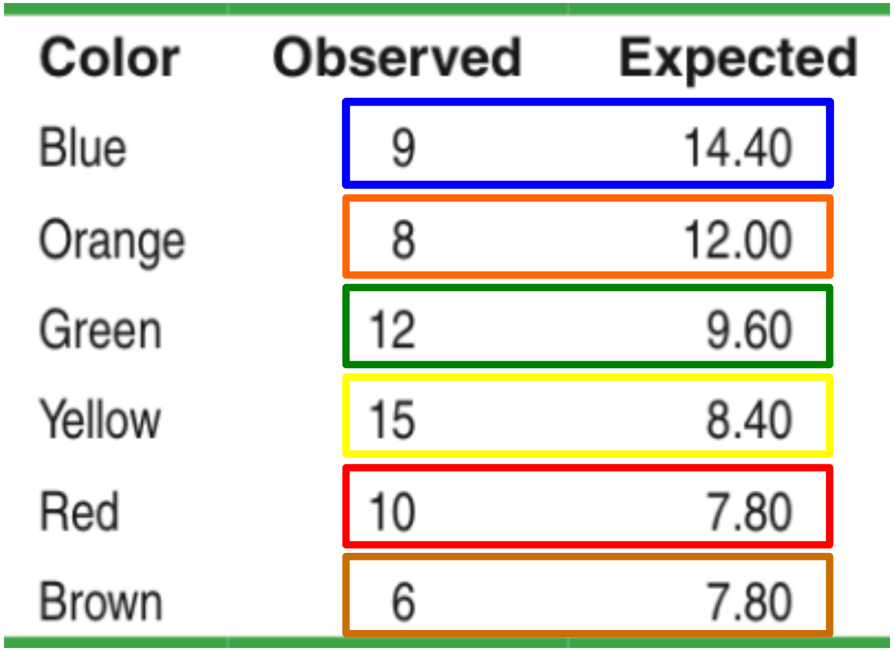
$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

where the sum is over all possible values of the categorical variable.

Example: Return of the M&M's

The table shows the observed and expected counts for our sample of 60 M&M's Milk Chocolate Candies. Calculate the chi-square statistic.

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$



Color	Observed	Expected
Blue	9	14.40
Orange	8	12.00
Green	12	9.60
Yellow	15	8.40
Red	10	7.80
Brown	6	7.80

$$\chi^2 = \boxed{\frac{(9-14.40)^2}{14.40}} + \boxed{\frac{(8-12.00)^2}{12.00}} + \boxed{\frac{(12-9.60)^2}{9.60}}$$

$$+ \boxed{\frac{(15-8.40)^2}{8.40}} + \boxed{\frac{(10-7.80)^2}{7.80}} + \boxed{\frac{(6-7.80)^2}{7.80}}$$

$$\chi^2 = 2.025 + 1.333 + 0.600 + 5.186 + 0.621 + 0.415$$
$$= 10.180$$

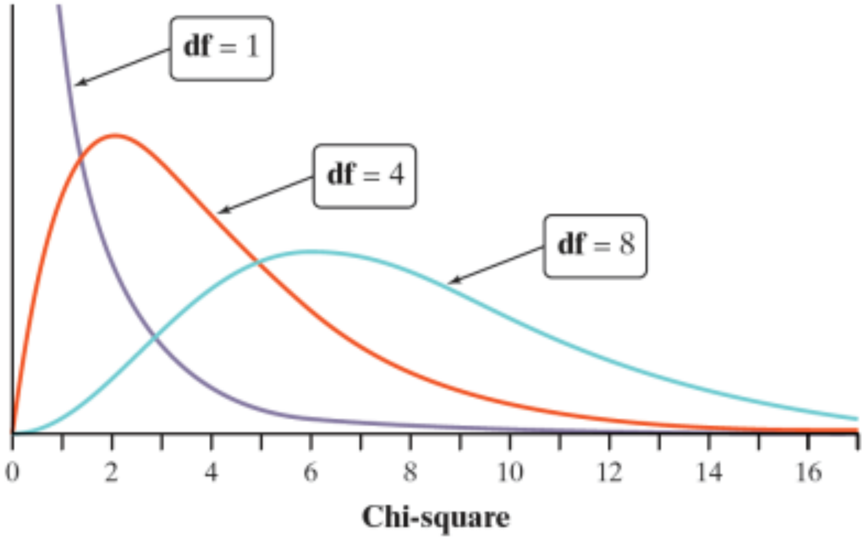
The Chi-Square Distributions and *P*-Values

The sampling distribution of the chi - square statistic is not a Normal distribution. It is a right - skewed distribution that allows only positive values because χ^2 can never be negative.

When the expected counts are all at least 5, the sampling distribution of the χ^2 statistic is close to a **chi - square distribution** with degrees of freedom (df) equal to the number of categories minus 1.

The Chi-Square Distributions

The chi-square distributions are a family of distributions that take only positive values and are skewed to the right. A particular chi-square distribution is specified by giving its degrees of freedom. The chi-square goodness-of-fit test uses the chi-square distribution with degrees of freedom = the number of categories - 1.



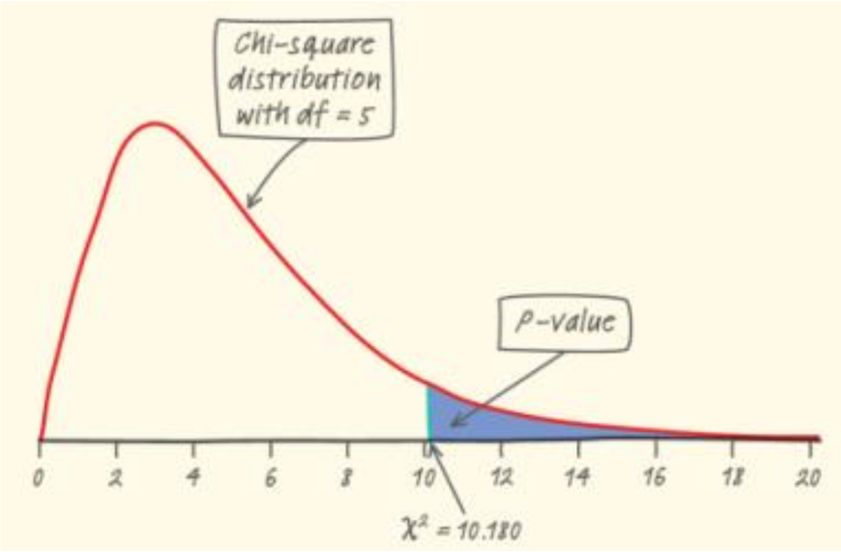
- When conducting a goodness of fit test to evaluate how well the observed data follow an expected distribution, the degrees of freedom are calculated as the number of categories (k) minus 1.

$$df = k - 1$$

- In this example, $k = 6$; Therefore

$$df = 6 - 1 = 5$$

- The p-value for a chi-square test is defined as the tail area above the calculated test statistic.



P-value=$P\left(\chi^2_{df=5} > 10.180\right)$

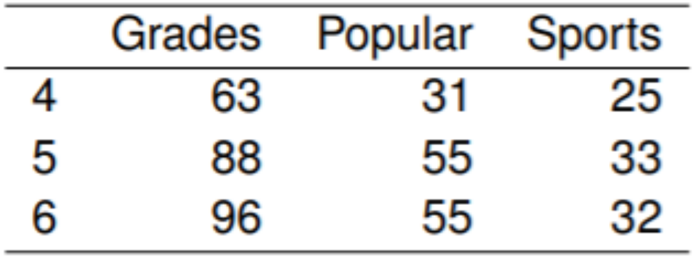
Example 2: When Were You Born?

Are births evenly distributed across the days of the week? The one-way table below shows the distribution of births across the days of the week in a random sample of 140 births from local records in a large city. Do these data give significant evidence that local births are not equally likely on all days of the week?

Day	Sun	Mon	Tue	Wed	Thu	Fri	Sat
Births	13	23	24	20	27	18	15

Two Way Table (Test for Independence)

In the dataset popular, students in grades 4-6 were asked whether good grades, athletic ability, or popularity was most important to them. A two-way table separating the students by grade and by choice of most important factor is shown below. Do these data provide evidence to suggest that goals vary by grade?



	Grades	Popular	Sports
4	63	31	25
5	88	55	33
6	96	55	32

Hypotheses

H_0: Grade and goals are independent. Goals do not vary by grade.

H_a: Grade and goals are dependent. Goals vary by grade.

Expected counts in two-way tables

Expected counts in two-way tables

$$\text{Expected Count}_{\textit{row } i,\textit{col } j} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{table total}}$$

	Grades	Popular	Sports	Total
4	63	31	25	119
5	88	55	33	176
6	96	55	32	183
Total	247	141	90	478

$$E_{\textit{row } 1,\textit{col } 1} = \frac{119 \times 247}{478} = 61 \qquad E_{\textit{row } 1,\textit{col } 2} = \frac{119 \times 141}{478} = 35$$

Expected counts in two-way tables

Clicker question

What is the expected count for the highlighted cell?

	Grades	Popular	Sports	Total
4	63	31	25	119
5	88	*55*	33	176
6	96	55	32	183
Total	247	141	90	478

(a) $\frac{176 \times 141}{478}$

(b) $\frac{119 \times 141}{478}$

(c) $\frac{176 \times 247}{478}$

(d) $\frac{176 \times 478}{478}$

Calculating the test statistic in two-way tables

- The calculation of the test statistic is exactly the same as before.

$$\chi^2 = \sum \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

{ let's think for a second about why this works... }

- Then calculate the degrees of freedom

$$df = (R - 1) \times (C - 1)$$

where R is the number of rows, and C is the number of columns.

Note: *We calculate df differently for one way and two-way tables.*

Calculating the test statistic in two-way tables

Expected counts are shown in (blue) next to the observed counts.

	Grades	Popular	Sports	Total
4	63 (61)	31 (35)	25 (23)	119
5	88 (91)	55 (52)	33 (33)	176
6	96 (95)	55 (54)	32 (34)	183
Total	247	141	90	478

$$\chi^2 = \sum \frac{(63-61)^2}{61} + \frac{(31-35)^2}{35} + \cdots + \frac{(32-34)^2}{34} = 1.3121$$

$$df = (R-1) \times (C-1) = (3-1) \times (3-1) = 2 \times 2 = 4$$

Calculating the p-value

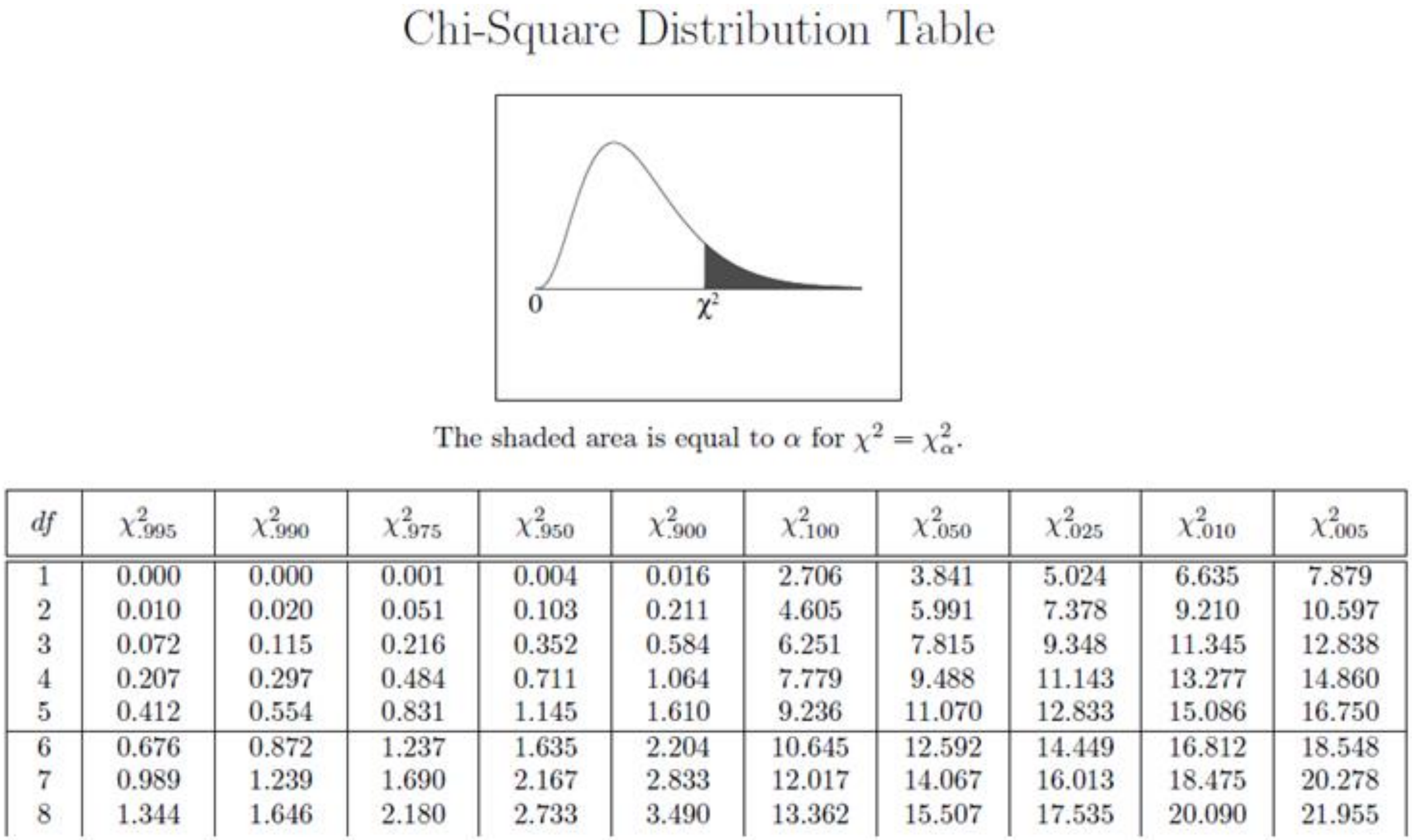
P-value=$P\left(\chi^2_{df=4} > 1.3121\right) =$ X^2cdf(1.3121,9999999,4)

Conclusion at $\alpha = .05$:

Conditions for the chi-square test

- Expected cell count: Adequate expected cell counts. A common rule is 5 or more in all cells of a 2-by-2 table, and 5 or more in 80% of cells in larger tables, but no cells with zero expected count.

- Independence: The observations are always assumed to be independent of each other. This means chi-squared cannot be used to test correlated data (like matched pairs or panel data).



Chi-Square Distribution Table

The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
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