## Section 8.5
## Goodness of Fit Test

Suppose we want to make an inference about a group of data (instead of just one or two). Or maybe we want to test counts of categorical data. **Chi-square** (or $\chi^2$) testing allows us to make such inferences.

There are several types of Chi-square tests but in this section we will focus on the goodness-of-fit test. **Goodness-of-fit** test is used to test how well one sample proportions of categories "match-up" with the known population proportions stated in the null hypothesis statement. The Chi-square goodness-of-fit test extends inference on proportions to more than two proportions by enabling us to determine if a particular population distribution has changed from a specified form.

The null and alternative hypotheses do not lend themselves to symbols, so we will define them with words. These are different from what we're used to seeing.

$H_0$: _____ is the same as _____

$H_a$: _____ is different from _____

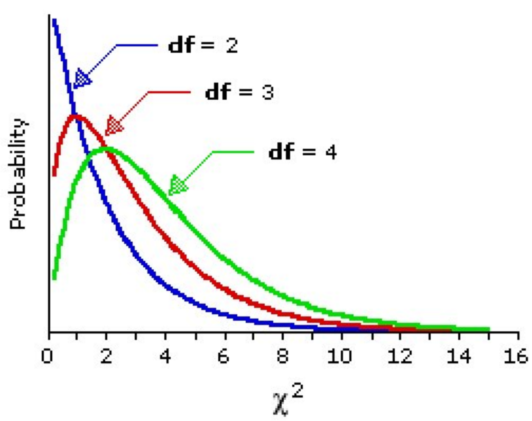
For each problem you will make a table with the following headings:

| Observed Counts (O) | Expected Counts (E) | $\frac{(O-E)^2}{E}$ |
|---|---|---|

The sum of the third column is called the Chi-square test statistic. $\Sigma$ represents the sum.

$$\chi^2 = \Sigma \frac{(\text{observed - expected})^2}{\text{expected}}$$

Chi-square distributions have only positive values and are skewed right. As the degrees of freedom increase it becomes more normal. The total area under the $\chi^2$ curve is 1.



**The assumptions for a Chi-square goodness-of-fit test are:**

1. The sample must be an SRS from the populations of interest.
2. The population size is at least ten times the size of the sample.
3. All expected counts must be at least 5.

**To find probabilities for $\chi^2$ distributions:**
R command is: 1 – pchisq(*test statistic*, *df*)

Note: degrees of freedom = df = (number of categories – 1)

Example: The Mixed-Up Nut Company advertises that their nut mix contains (by weight) 40% cashews, 15% Brazil nuts, 20% almonds and only 25% peanuts. The truth-in-advertising investigators took a random sample (of size 50 lbs) of the nut mix and found the distribution to be as follows:

| *Cashews* | *Brazil Nuts* | *Almonds* | *Peanuts* |
|---|---|---|---|
| 15 lb | 11 lb | 13 lb | 11 lb |

At the 1% level of significance, is the claim made by Mixed-Up Nuts true?

The following is how you'll normally set up your null and alternate hypothesis.
$H_0$: The data distribution of nuts **is the same as** the population.
$H_a$: The data distribution of nuts **is different from** the population.

Let's first find the *expected*:

| *Cashews* | *Brazil Nuts* | *Almonds* | *Peanuts* |
|---|---|---|---|
| | | | |

Next, let's find the test statistic:

$$\chi^2 = \Sigma \frac{(\text{observed} - \text{expected})^2}{\text{expected}} =$$

Before we find the p-value, what is df = (number of categories – 1)?

Now find the p-value: