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 χ^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.

  H_0 : _____ is the same as _____

  H_a : _____ is different from _____

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

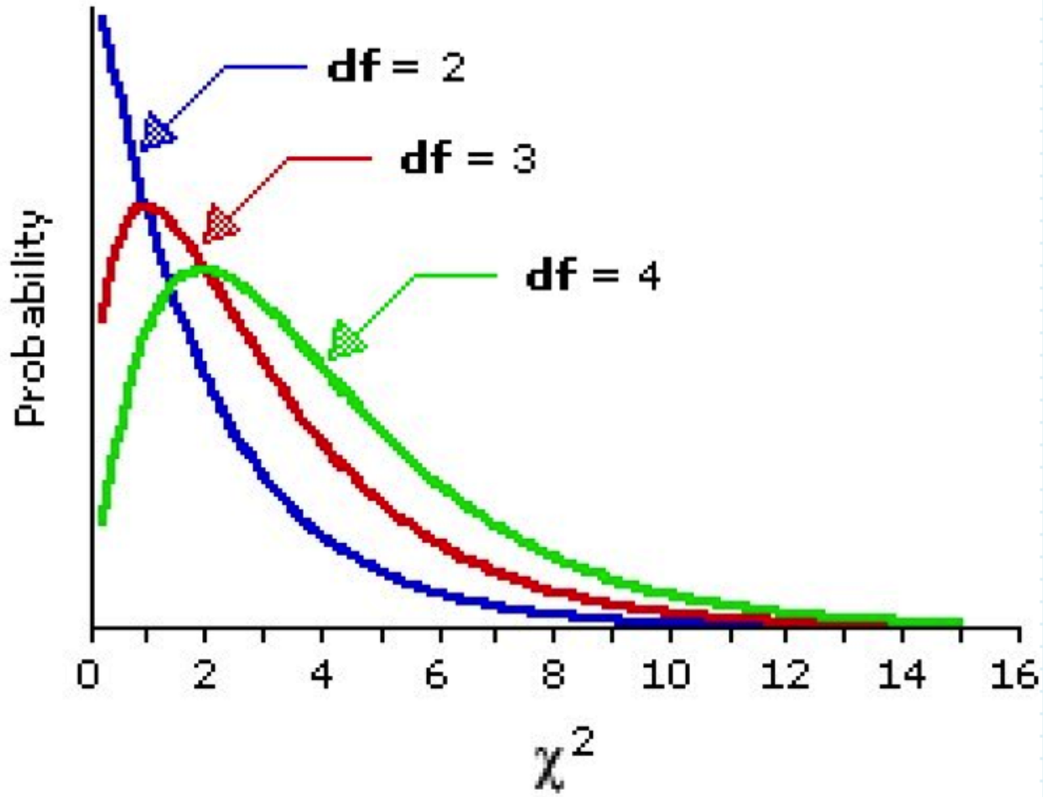
Observed Counts (O)	Expected Counts (E)	$\frac{(O-E)^2}{E}$

- The sum of the third column is called the Chi-square test statistic.

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

- We can use a table, the calculator, or R-Studio to find *p*-values for χ^2 with $n - 1$ degrees of freedom.

- 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 χ^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 distributions:
  - TI-83/84 calculator uses the command χ^2 **cdf** found under the DISTR menu.
  - R-Studio command is: 1 – pchisq(*test statistic, df*)

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
O	15 lb	11 lb	13 lb	11 lb

$n = 4$

$df = 3$

$\alpha = .01$

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

H_0: the data distribution of nuts is same as population

H_a: at least one proportion is different

	O	E	$\frac{(O-E)^2}{E}$
Cashews	15	.4(50) = 20	1.25
Brazil	11	.15(50) = 7.5	1.63
Almonds	13	.2(50) = 10	.9
Peanuts	11	.25(50) = 12.5	.18

Sum = $3.96 = \chi^2$

p value $\chi^2 = 3.96$ df $= 4-1 = 3$
$= .2658 > \alpha$

χ^2cdf(test stat, 10^6, df)
1- pchisq(teststat, df)

Fail to reject H_0