## Chapter 26: Comparing Counts

### Test for Goodness of Fit:

To analyze categorical data, we construct two-way tables and examine the counts or percents of the explanatory and response variables.

We want to compare the observed counts to the expected counts. The null hypothesis is that there is no difference between the observed and expected counts. The alternative hypothesis is that there is a difference between the observed and expected counts.

$X^2 = \sum \frac{(O-E)^2}{E}$ is called the chi-square test statistic. It measures how well the observed counts fit the expected counts, assuming that the null hypothesis is true.

The distribution of the *chi-square statistic* is called the *chi-square distribution,* $\chi^2$. This distribution is a density curve.

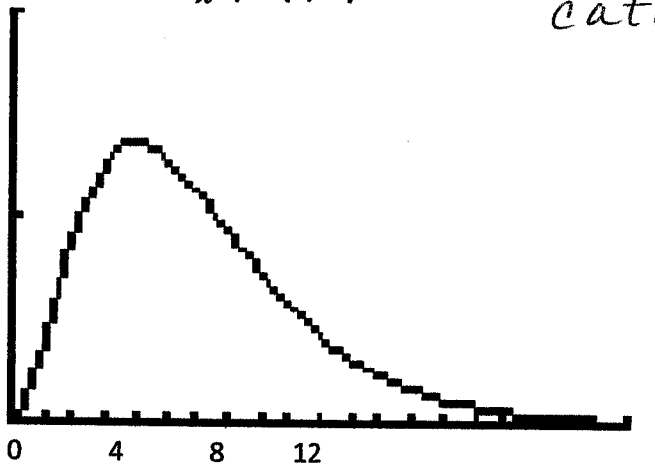
- The total area under the curve is 1.
- The curve begins at 0 on the horizontal axis and is skewed right.
- As the degrees of freedom increase, the shape of the curve becomes more symmetric

We can find the probability of obtaining a $\chi^2$ value at least as extreme as the one we observe in our sample (assuming the null hypothesis is true). This gives the p-value for what is called the **"Goodness of Fit Test."**

The chi-square distribution has $K-1$ degrees of freedom:

K is the number of categories (rows)

TI-83⁺: $\chi^2$pdf (X, df)



**CONDITIONS**: The Goodness of Fit Test may be used when:

- we are working with Counts
- all counts are at least 1 and most (at least 80%) of the expected counts are at least 5.

*Following the Goodness of Fit Test, check to see which component made the greatest contribution to the chi-square statistic to see where the biggest changes occurred.

**Inference for Two Way Tables:**

To compare two proportions, we use a 2-Proportion Z Test. If we want to compare three or more proportions, we need a new procedure.

The first step in the overall test for comparing several proportions is to arrange the data in a two-way table.

Think of the counts as elements of a matrix with r rows and c columns. This is called an r x c **table** with (r)(c) cells.

Our null hypothesis is that there is no difference among the proportions. The alternative hypothesis is that there is some difference among the proportions.

We will use the chi-square test to measure how far the observed values are from the expected values.

To calculate the expected counts, multiply the row total by the column total, and divide by the table total:

$$expected\ count = \frac{row\ total \times column\ total}{table\ total}$$

The chi-square statistic is the sum over all r x c cells in the table:

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

The degrees of freedom is $(r-1)(c-1)$.

The P-value is the area to the right of the $X^2$ statistic under the chi-square density curve.

**In Summary:**

1. Use a chi-square Goodness of Fit test when you want to show that the distribution of counts in one categorical variable matches the distribution predicted by a model.

   (one categorical variable, one-way table)

EX: $H_0$: The distribution of M&M colors is the same as the advertised distribution.

2. Use a chi-square test for homogeneity when you want to show that the distribution of counts for two or more groups is the same.

   (one categorical variable, two-way table)

EX: $H_0$: The distribution of X is the same as the distribution of Y.

A chi-square test for homogeneity is used to answer the question, "Are the groups homogeneous?"

This type of test is used with more than one population. **(SPLIT, THEN SAMPLE)**

3. Use a chi-square test for independence when you want to show that two categorical variables are independent for one group of individuals.

   (two categorical variables, two-way table)

EX: $H_0$: X and Y are independent.

A chi-square test for independence is used to answer the question, "Are the variables independent?"

This type of test is used with one population. **(SAMPLE, THEN SPLIT)**