**Test of Goodness of Fit**

It is often the case that information is collected on categorical variables. When wanting to compare proportions of categories for univariate data, we use a Goodness of Fit Test. A Goodness of Fit Test is used to see if the distribution of proportions for a single variable follows a given distribution or not.

Setup of a Goodness of Fit Test:

 **Hypotheses:**

H0: p1 = hypothesized proportion for category 1

 p2 = hypothesized proportion for category 2

 :

 pk = hypothesized proportion for category k

\*\*make sure to define these in context\*\*

Ha: At least one of the proportions differs from hypothesized value.

 **Assumptions**:

1. Observed cell counts are based on a random sample.
2. The sample size is large. The sample size is large enough if every expected cell count is at least 5.

**Calculations:**

**List expected values**: expected cell count = n \* (hypothesized proportion for category)

**Test Statistic(you don’t have to show work):**



**Degrees of Freedom: df = # categories – 1**.

 **P-value:**

The P-value is the area to the right of  under the df= k – 1 chi-square curve.

You can find these values in Table C of your reference packet. You can also use the cdf( value, upper bound, df) function on your calculator.

**On the calculator**: On certain calculators you can do a Goodness of Fit test by going to

 STAT🡪TESTS🡪D: GOF-TEST. Enter observed counts into one

 list and expected counts into another.

**Test of Homogeneity**

To compare **two or more populations** or treatments on the basis of a categorical variable to see whether the category proportions in one population are equal to the same category proportions in other populations, we conduct a  Test of Homogeneity.

 **Hypotheses:**

H0: The true category proportions in one population are equal to the same category

 proportions in all populations. (obv. In **context!)**

Ha: The true category proportions are NOT the same for all the populations.

 **Assumptions**:

1. The data consists of independently chosen random samples (or subjects were assigned at

 random to treatment groups.)

1. The sample size is large. The sample size is large enough if every expected cell count is at least 5.

 **Calculations:**

**Write down expected values**

Expected cell count = (row total)(column total) / grand total < -- **they are also stored in Matrix B after you run the test.**

**Test Statistic:**



**df = (# of rows – 1)(# of columns – 1).**

**P-value**

**On the calculator:** Go to STAT🡪TESTS🡪C: -TEST. Enter observed counts in a matrix and

 expected counts will be calculated and placed in another matrix for you.

**Test of Independence**

In a test of independence, we compare **two categorical variables from the same population** to one another to see if they are independent. The null hypothesis will always assume they are independent. The alternative is that they are not independent (they are dependent). The set up for the test is basically the same as for a test of homogeneity, except what we are actually testing for:

 **Hypotheses:**

H0: The two variables are independent.

Ha: The two variables are not independent.

 **Assumptions**:

1. Observed cell counts are based on a random sample.
2. The sample size is large. The sample size is large enough if every expected cell count is atleast 5.

 **Calculations:**

**Write down expected values**

Expected cell count = (row total)(column total) / grand total < -- **they are also stored in Matrix B after you run the test.**

**Test Statistic:**



**df = (# of rows – 1)(# of columns – 1).**

**P-value**

**On the calculator:** Go to STAT🡪TESTS🡪C: -TEST. Enter observed counts in a matrix and

 expected counts in another matrix.

**Test of homogeneity vs independence**

The two tests are very similar and the difference is a matter of design (similar to a 2-sample t vs a paried t)

The "test of homogeneity" is a way of determining whether two or more populations share the same distribution of a single categorical variable. In the test of homogeneity, the data are collected by randomly sampling from each population **separately**.

The "test of independence" is a way of determining whether two categorical variables are associated with one another in a single population. In the test of independence, observational units are collected at random from **a** population and two categorical variables are observed for each unit.