IEOR 165 – Lecture 23
Other Null Hypothesis

---

1 Goodness of Fit Testing

Suppose an item (or person) has one feature. For instance, this feature might be gender (Male or Female). Suppose there are r different possibilities (more formally, the term *category* is used). Then, we can define a null hypothesis

$$H_0 : \mathbb{P}(x = i) = p_i, \forall i \in [r] = \{1, \ldots, r\},$$

where p_i is a fixed and *a priori* known value. Now suppose we have observed n iid items, and let N_i be the number of items that are observed to have the i-th feature. Then test statistic

$$T = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}$$

is approximately described by a $\chi^2(r-1)$, that is a χ^2 distribution with $r-1$ degrees of freedom. And so the approximate p-value is

$$p = \mathbb{P}(\chi^2(r-1) \geq T).$$

2 Testing Independence of Categorical Variables

2.1 Two Categories

Suppose an item (or person) has two features. For instance, one feature might be gender (Male or Female) and the other feature might be educational level (High School, BS, MS, or Graduate/Professional). Suppose that there are r_1 different possibilities (more formally, the term *category* is used) for the first feature x_1, and r_2 different categories for the second feature x_2. For convenience, we can number the first features using the values $[r_1] = \{1, 2, \ldots, r_1\}$ and number the second features using the values $[r_2] = \{1, 2, \ldots, r_2\}$. Furthermore, suppose the model is that the probability that a single item has a first feature that is $i \in [r_1]$ and second feature that is $j \in [r_2]$ is given by

$$\mathbb{P}(x_1 = i, x_2 = j).$$

However, if the probability of the two features are independent, then we can factor this probability as

$$\mathbb{P}(x_1 = i, x_2 = j) = \mathbb{P}(x_1 = i)\mathbb{P}(x_2 = j).$$

This observation forms the basis for Pearson's χ^2-test.

In particular, suppose the null hypothesis is that

$$H_0 : \mathbb{P}(x_1 = i, x_2 = j) = \mathbb{P}(x_1 = i)\mathbb{P}(x_2 = j), \forall i \in [r_1], j \in [r_2].$$

If we have measurements of n items, where the items are iid, then we can define N_{ij} to be the number of items with features i, j. Then we can define an estimate of $\mathbb{P}(x_1 = i)$ by

$$\hat{p}_i = \sum_{j=1}^{r_2} N_{ij}/n$$

and an estimate of $\mathbb{P}(x_2 = j)$ by

$$\hat{q}_j = \sum_{i=1}^{r_1} N_{ij}/n.$$

If we define a test statistic

$$T = \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} \frac{N_{ij}^2}{n\hat{p}_i\hat{q}_j} - n,$$

then it turns out that this is approximately described by a $\chi^2((r_1 - 1)(r_2 - 1))$, that is a χ^2-distribution with $(r_1 - 1)(r_2 - 1)$ degrees of freedom. The approximate p-value is then given by

$$p = \mathbb{P}(\chi^2((r_1 - 1)(r_2 - 1)) \geq T).$$

The intuition for why this test statistic is approximately described by a χ^2-distribution is more involved and beyond the scope of this course. A proof can be found in:L A. Agresti, *Categorical Data Analysis*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons Inc., New York, 1990.

2.2 Multiple Categories

We can test for more general interactions between categorical variables. Suppose we have p categories, and let $[p] = \{1, \ldots, p\}$. Define a simplicial complex Γ to be a set such that:

- $\Gamma \subseteq 2^{[p]}$, where $2^{[p]}$ denotes the power set of $[p]$. Recall that the power set $2^{[p]}$ is the set of all subsets of $[p]$.
- If $F \in \Gamma$ and $S \subset F$, then $S \in \Gamma$.

The elements $F \in \Gamma$ are called the *faces* of Γ, and the inclusion-maximal faces are the *facets* of Γ. Then we can define a null hypothesis

$$H_0 : \mathbb{P}(x_1 = i_1, \ldots, x_p = i_p) = \frac{1}{Z(\theta)} \prod_{F \in \mathsf{facets}(\Gamma)} \theta^{(F)}_{i_F},$$

where $Z(\theta)$ is a normalizing constant so that this distribution sums to one, and $\theta^{(F)}$ are sets of parameters indexed by the values of $i_F = \{i_{f_1}, i_{f_2}, \ldots\}$, where $F = \{f_1, f_2, \ldots\}$. This probability

distribution model is known as a hierarchical log-linear model.

This null hypothesis may be difficult to interpret, and so it is useful to consider some special cases. If $\Gamma = [1][2]$ (meaning there are two facets which are singleton), then this corresponds to a null hypothesis in which the first and second categories are independent. If $\Gamma = [12][13][23]$, then this corresponds to a null hypothesis in which there are no three-way interactions between the categories. Even more general null hypothesis are possible. For instance, we may have $\Gamma = [12][23][345]$.

It turns out that we can use Monte Carlo algorithms to compute the p-value for this null hypothesis test. We can also use an approximation that $-2\log(\text{likelihood})$ is approximately described by a χ^2-distribution, where the degrees of freedom depends on the topology of the simplicial complex. Fortunately, there is software that can aid with this analysis.

3 Kolmogorov-Smirnov Test

3.1 One-Sample Test

Suppose we have a null hypothesis

$$H_0 : X_i \sim F(u),$$

where the X_i are iid and $F(u)$ is a known distribution. This test is used to analyze whether the data samples $X_1, X_2, \ldots, X_n$ are drawn from the distribution $F(u)$. The idea is to begin with the sample distribution function

$$\hat{F}(u) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq u),$$

and define the test statistic

$$T = \sup_u |\hat{F}(u) - F(u)|.$$

Recall that $\sup_u g(u)$ means the least upper bound of $g(u)$, and it is similar to $\max_u g(u)$ which is the maximum of $g(u)$. The difference is that if a maximum exists, then it is equal to the supremum; however, a supremum always exists when we look at subsets of the real line, whereas a maximum may not exist.

It turns out that $\sqrt{n} \cdot T$ can be approximated by the Kolmogorov distribution, which has a complicated definition. However, the key point is that the Kolmogorov distribution does not depend upon the $F(u)$ distribution from the null hypothesis. And once again, we can use software to compute p-values for this hypothesis test. One-sided tests are also possible, and are defined in a similar way.

3.2 Two-Sample Test

Suppose we have a null hypothesis

$$H_0 : X_i, Y_i \sim F(u),$$

where the X_i, Y_i are iid and $F(u)$ is an unknown distribution. This test is used to analyze whether the data samples X_i, for $i = 1, \ldots, n_x$, and Y_i for $i = 1, \ldots, n_y$ are drawn from the distribution $F(u)$. The idea is to begin with the two sample distribution functions

$$\hat{F}_x(u) = \frac{1}{n_x} \sum_{i=1}^{n_x} 1(X_i \leq u)$$

$$\hat{F}_y(u) = \frac{1}{n_y} \sum_{i=1}^{n_y} 1(Y_i \leq u)$$

and define the test statistic

$$T = \sup_u |\hat{F}_x(u) - \hat{F}_y(u)|.$$

It turns out that this test statistic can be used to compute a p-value. And once again, we can use software to compute p-values for this hypothesis test. One-sided tests are also possible, and are defined in a similar way. The approximate distributions are more complex, and references can be found in the "R" language documentation for the function "ks.test stats".