4. Goodness of Fit Test - Discrete Distributions

The Goodness of Fit Hypothesis Test for Discrete Distributions

1/1 point (graded)

Let $X_1, \ldots, X_n$ be iid samples from a discrete distribution $\mathbf{P}_{\mathbf{p}}$ for some unknown $\mathbf{p} \in \Delta_K$. Let $\mathbf{p}^0 \in \Delta_K$ define a fixed pmf.

Which of the following represent valid goodness of fit tests to know whether there is statistical evidence that $X_1, \ldots, X_n$ could have been generated by the pmf $\mathbf{p}^0$ as opposed to any other pmf? (Choose all that apply.)

- [x] $H_0 : \mathbf{p} = \mathbf{p}^0, H_1 : \mathbf{p} \neq \mathbf{p}^0$
- [ ] $H_0 : \|\mathbf{p}\|_2 = \|\mathbf{p}^0\|_2, H_1 : \|\mathbf{p}\|_2 \neq \|\mathbf{p}^0\|_2$

✔

Solution:

The first choice is a valid goodness of test while the second choice is not. Our aim in a goodness of fit test is to know whether there is statistical evidence that the data was generated by only our candidate distribution (against all other possible distributions). The first choice clearly achieves this aim.

The second choice is not a valid goodness of fit test. The failure to reject the null hypothesis does not necessarily imply that there is statistical evidence to say $\mathbf{p}^0$ is the only distribution that could have generated the observed data (with some probability). There are many vectors $\mathbf{p}$ that satisfy $\|\mathbf{p}\|_2 = \|\mathbf{p}^0\|_2$ so the failure to reject means that we have statistical evidence that many possible candidate distributions could have

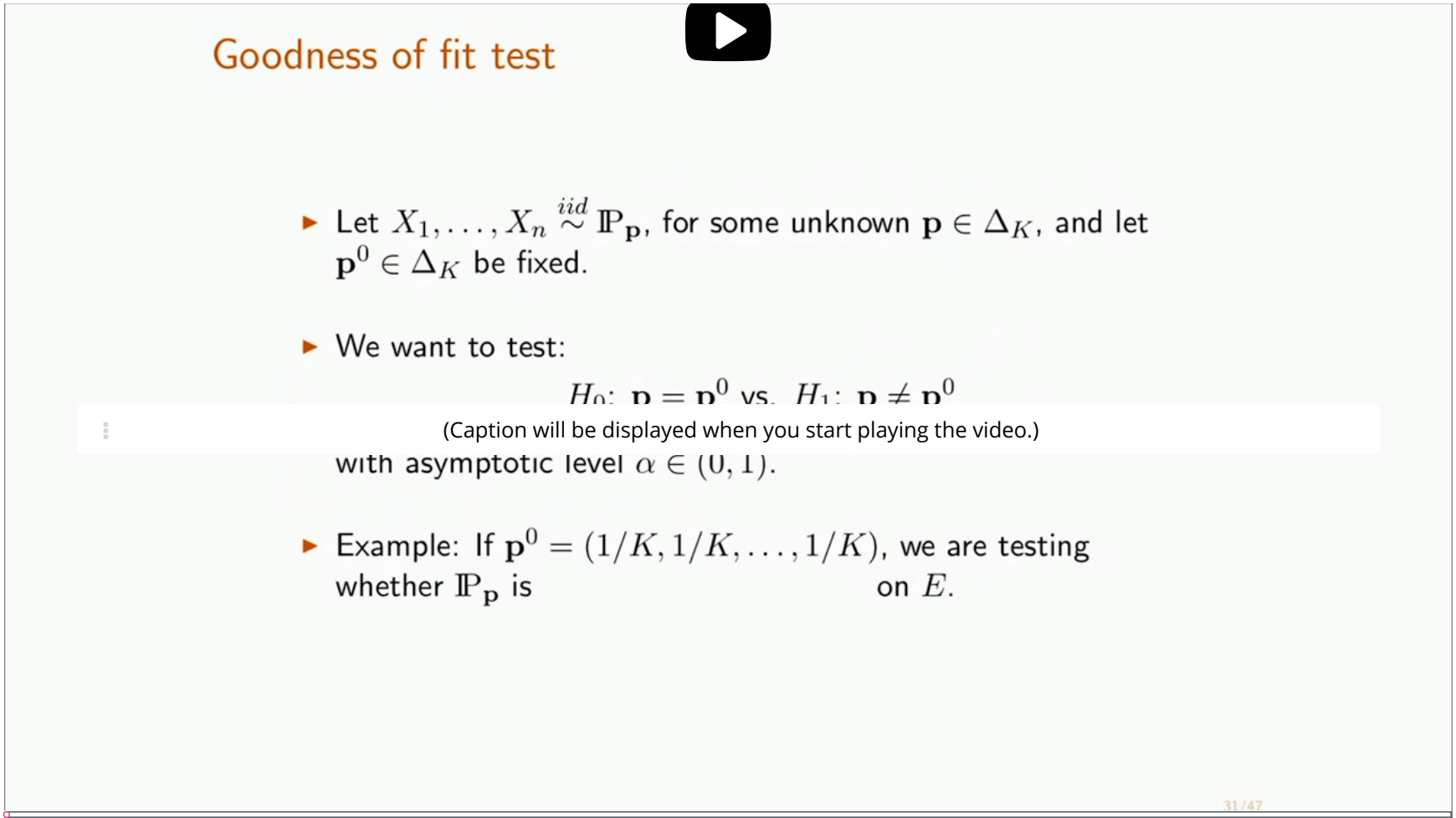
generated the data.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Video and Lecture Note: Throughout this lecture (including in the video below) we see the terms "multinomial distribution" and "multinomial likelihood" being used in places where the more appropriate terms "categorical distribution" and "categorical likelihood", respectively, should be used. The note following the below video introduces both multinomial and categorical distributions and clarifies that the categorical distribution is a special case of the multinomial distribution.

The Goodness of Fit Test: Categorical Likelihoods



Video
Download video file

Transcripts
Download SubRip (.srt) file
Download Text (.txt) file

Multinomial Distribution

The **Multinomial Distribution** with K modalities (or equivalently K possible outcomes in a trial) is a generalization of the binomial distribution. It models the probability of counts of the K possible outcomes of the experiment in n' i.i.d. trials of the experiment.

It is parameterized by the parameters $n', p_1, \ldots, p_K$ where

- n' is the number of i.i.d trials of the experiment;
- p_i is the probability of observing outcome i in any trial, and hence the p_i's satisfy $p_i \geq 0$ for all $i = 1, \ldots, K$, and $\sum_{i=1}^{K} p_i = 1$.

Let $\mathbf{p} \triangleq [p_1 \;\; p_2 \;\; \cdots \;\; p_K]^T$ and note that $\mathbf{p} \in \Delta_K$.

The multinomial distribution can be represented by a random vector $\mathbf{N} \in \mathbb{Z}^K$ to represent the number of instances $N^{(i)}$ of the outcome $i = 1, \ldots, K$. Note that $\sum_{i=1}^{K} N^{(i)} = n'$. The **multinomial pmf** for all $\mathbf{n}$ such that $\sum_{i=1}^{K} n^{(i)} = n'$, $n^{(i)} \geq 0, i = 1, \ldots, K$, and $n^{(i)} \in \mathbb{Z}, i = 1, \ldots, K$ is given by

$$p_{\mathbf{N}}\left(N^{(1)} = n^{(1)}, \ldots, N^{(K)} = n^{(k)}\right) = \frac{n'!}{n^{(1)}!n^{(2)}! \cdots n^{(K)}!} \prod_{i=1}^{K} p_i^{n^{(i)}}.$$

Categorial (Generalized Bernoulli) Distribution and its Likelihood

The multinomial distribution, when specialized to $n' = 1$ for any K gives the **categorical distribution** . When $K = 2$ and the two outcomes are 0 and 1 the categorical distribution is the Bernoulli distribution, and for any $K \geq 2$ the categorical distribution is also known as the **generalized Bernoulli distribution** .

The categorical distribution, therefore, models the probability of counts of the K possible outcomes of a discrete experiment in a single trial. Since the total count is equal to 1 (only one trial), we can use a random variable X to represent the outcome of the trial. This means the sample space of a **categorical random variable** X is

$$E = \{a_1, \ldots, a_K\}.$$

The vector $\mathbf{p}$ is the parameter of a categorical random variable. The pmf of a categorical distribution can be given as

$$P(X = a_j) = \prod_{i=1}^{K} p_i^{\mathbf{1}(a_i = a_j)} = p_j, \;\; j = 1, \ldots, K.$$

Let $\mathbf{P_p}$ denote the distribution of a categorical random variable with sample space $E = \{a_1, \ldots, a_K\}$ and parameter vector $\mathbf{p}$. The **categorical statistical model** can thus be written as the tuple $\left(\{a_1, \ldots, a_K\}, \{\mathbf{P_p}\}_{\mathbf{p} \in \Delta_K}\right)$.

In goodness of fit testing for a discrete distribution, we observe n iid samples $X_1, \ldots, X_n$ of a categorical random variable X and it is our aim to find statistical evidence of whether a certain distribution $\mathbf{p}^0 \in \Delta_K$ could have generated $X_1, \ldots, X_n$.

The **categorical likelihood** of observing a sequence of n iid outcomes $X_1, X_2, \ldots, X_n \sim X$ can be written using the number of occurrences $N_i, i = 1, \ldots, K$, of the K outcomes as

$$L_n(X_1, \ldots, X_n, p_1, \ldots, p_K) = p_1^{N_1} p_2^{N_2} \cdots p_K^{N_K}.$$

The categorical likelihood of the random variable X, when written as a random function, is

$$L(X, p_1, \ldots, p_K) = \prod_{i=1}^{K} p_i^{\mathbf{1}(X = a_i)}.$$

Discussion

Hide Discussion

Topic: Unit 4 Hypothesis testing:Lecture 15: Goodness of Fit Test for Discrete Distributions / 4. Goodness of Fit Test - Discrete Distributions

Add a Post

Show all posts ▼ by recent activity ▼

There are no posts in this topic yet.

✖