4. Goodness of Fit Test - Discrete Distributions

The Goodness of Fit Hypothesis Test for Discrete Distributions

0/1得分 (计入成绩)

Let $X_1, \ldots, X_n$ be iid samples from a discrete distribution $\mathbf{P_p}$ 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.)

□ $H_0 : \mathbf{p} = \mathbf{p}^0, H_1 : \mathbf{p} \neq \mathbf{p}^0$ □

☑ $H_0 : \|\mathbf{p}\|_2 = \|\mathbf{p}^0\|_2, H_1 : \|\mathbf{p}\|_2 \neq \|\mathbf{p}^0\|_2$

□

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.

提交 你已经尝试了2次（总共可以尝试2次）

□ 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

So now, what are we going to do?

We have data X1 to Xn.

They are IID according to this distribution, so IID

with a certain PMF boldface p.

And this boldface p is unknown.

And I'm going to try to test if I have a very

specific distribution p0.

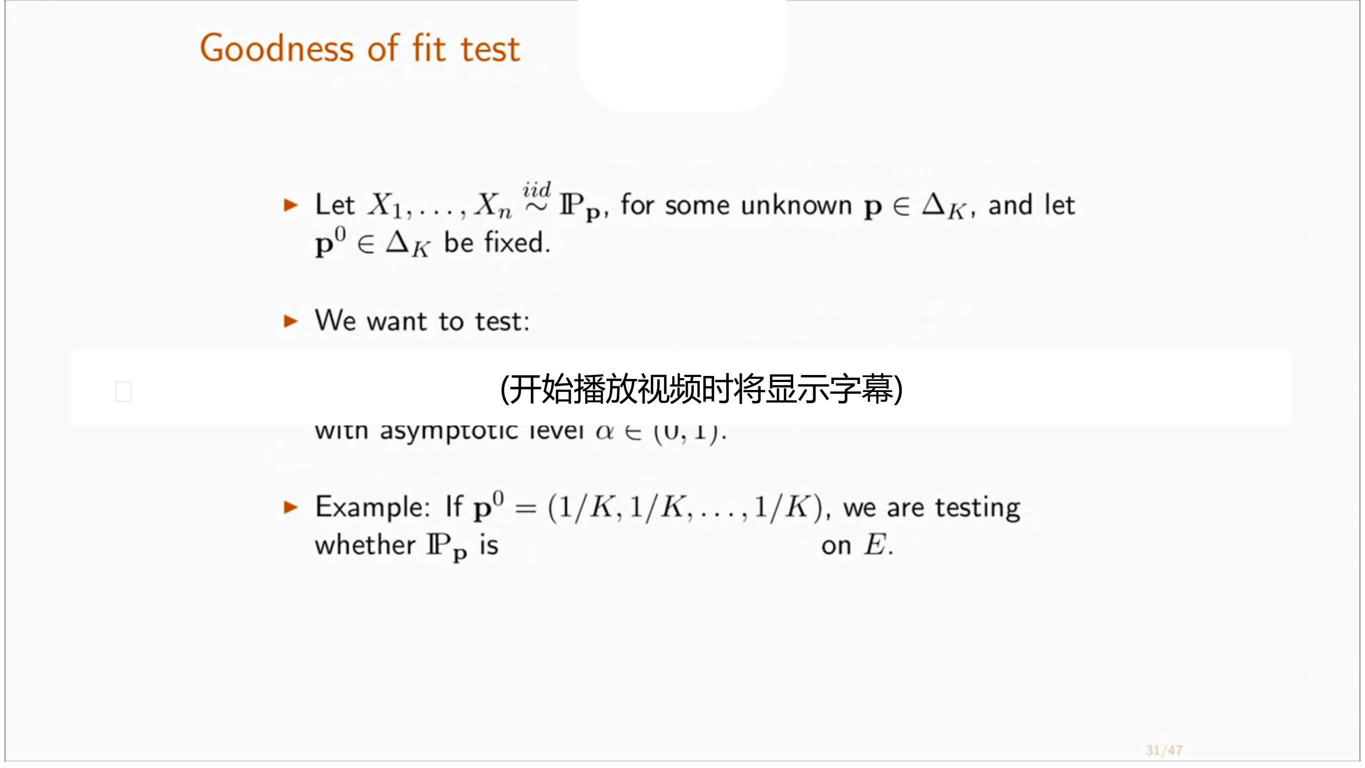
So how do I find this p0?

Well, let's go back to a couple examples.

Here, p0 is this guy.

So p0 here is--

sorry, p up 0.



Goodness of fit test

- Let $X_1, \dots, X_n \overset{iid}{\sim} \mathbb{P}_{\mathbf{p}}$, for some unknown $\mathbf{p} \in \Delta_K$, and let $\mathbf{p}^0 \in \Delta_K$ be fixed.
- We want to test:

(开始播放视频时将显示字幕)

with asymptotic level $\alpha \in (0, 1)$.

- Example: If $\mathbf{p}^0 = (1/K, 1/K, \dots, 1/K)$, we are testing whether $\mathbb{P}_{\mathbf{p}}$ is on E.

31/47

So this is what?

Well, it's 0.72, 0.07, 0.12, and 0., for K is equal to 4.

So this is p0

视频

下载视频文件

字幕

下载 SubRip (.srt) file

下载 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 $n', p_1, \dots, p_K$ 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 $i = 1, \dots, K$, and $\displaystyle\sum_{i=1}^{K} p_i = 1$.

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 $i = 1, \dots, K$. Note that $\sum_{i=1}^{K} N^{(i)} = n'$. The **multinomial pmf** for all $\mathbf{n}$ such that $\sum_{i=1}^{K} n^{(i)} = n'$, $n^{(i)} \geq 0, i = 1, \dots, K$, and $n^{(i)} \in \mathbb{Z}, i = 1, \dots, K$ is given by

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

Categorial (Generalized Bernoulli) Distribution and its Likelihood

The multinomial distribution, when specialized to $\boxed{n' = 1}$ 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

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

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

$$\boxed{P\left(X = a_j\right) = \prod_{i=1}^{K} p_i^{\mathbf{1}(a_i = a_j)} = p_j, \quad j = 1, \dots, K.}$$

Let $\mathbf{P}_\mathbf{p}$ 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 $\left(\{a_1, \ldots, a_K\}, \{\mathbf{P}_\mathbf{p}\}_{\mathbf{p} \in \Delta_K}\right)$.

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

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

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

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

讨论

隐藏讨论

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

添加帖子

□ **所有讨论帖**

Changes in Lecture 15

□
□
□

由 sudarsanvsr_mit (员工) 于5天 以前发布此讨论帖

□ 已固定

Hello everyone,

We have made a few changes within Lecture 15 that I thought I should make a post to draw the attention of everyone who already finished watching L15 and attempted all the problems.

If you have not started with L15, you may skip this post.

1. First, under Vertical 4, we have now added a note to say what a multinomial distribution is and what a categorical distribution is. The categorical distribution, whose pmf we use, is a special case of the multinomial distribution. This is not, strictly speaking, a big technical change and it is merely a clarification of what terminology to use.

2. Because of the above, all references in text to "multinomial distribution" and "multinomial likelihood" have been changed to "categorical distribution" and "categorical likelihood" because the latter pair of terms is more appropriate.

3. Finally, Problem 1 in Vertical 2 has been modified to reflect what we intended to ask in the first place.

Thank you and happy learning!

--Course Staff

此帖对所有人可见。

添加回复

1个回复

SergK (社区助教)
5天 以前

□
□

I would not say that categorical distribution is a special case of multinomial. Even if $n' = 1$ the distributions are different, because multinomial r.v. is a vector while categorical r.v. is a scalar. Say you roll a die once and it lands 3; then the categorical r.v. takes value 3 while multinomial r.v. takes value

$$\begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

From the point of view of what the pmf captures, it is a special case. What we mean here is that the multinomial distribution models the probability of counts of the K possible outcomes in n' trials, while the categorical distribution models the probability of counts of the K possible outcomes in $n' = 1$ trials. Agreed, the random variable (or vector) used to represent the two is different. I could have equally used the same random vector when $n' = 1$ but a random variable with a sample space $\{a_1, \ldots, a_K\}$ is more succinct and does the same job.

sudarsanvsr_mit (员工) 在5天 以前前发表

It becomes a special case if the representation is changed to a "one-hot" vector, which has, as SergK shows, a one in the component of the vector of length K corresponding to the category k of the original categorical variable. Then the progression to a multinomial distribution is smooth -- as each new one-hot vector observation comes in, just add it into the previous vector of counts. This one-hot vector is a standard representation in other fields (and even in data science), so isn't obscure. (The name "one-hot" comes from electronics, where one might represent a number in, say, binary, or represent a number k out of K as a non-ground voltage on wire k out of a set of K wires, with the rest at ground -- the wire with the non-ground voltage is "hot"...as in, "Yee..ouch! Yep, that one's hot...")

There is even a function in R that will convert your categorical data into a one-hot representation. ...Which you have to do if you want to do (e.g.) linear regression, because the labels for your categories probably have no numerical meaning -- a category labeled "2" isn't twice as much as a category labeled "1", and worse if the categories have names (e.g "male" and "female"). If you have a column in your table that represents a category and has multiple labels as values, the function will produce a set of columns that are binary, one for each category label, each representing whether or not the category row contained that label. Only one of these will be set to true, and the rest false. Now it can be considered as a trivial form of discrete probability -- it adds up to one.

As we've seen, so long as every row actually had a category label, one of these one-hot columns is redundant, because of the constraint that one of them has to be on (equivalent to a probability summing to one). So, you can tell the R function to regard one category as the default, and omit its column. (E.g. if you had a category with "male" and "female", you could add two columns, one that had a 1 if "male", 0 otherwise, and the other with 1 for "female". But now you've got two columns that are just Boolean inverses of each other. Just as with a Bernoulli representation, it's standard to omit one of the columns.) If you don't do this, but leave the redundant column in, Bad Things happen when you then attempt linear regression...

The R name for these binary columns that replace a category column are "dummy variables", and the function I use for conversion is dummyVars from the caret machine learning package. This being R, there are lots of dummy variable functions, and everyone has their own favorite. If you tell R that your categorical column is a factor data type (as you should), then many R numerical tools will secretly unpack your factor column into dummy variables before starting their work.

So...although the two representations -- a single variable holding a category label versus a one-hot vector of variables with a constraint -- may be able to represent the same things, they are very different in usage in the Real World. Bottom line: You can use the one-hot / dummy variable representation in numerical methods, but not raw column labels. If you have column labels, you're pretty much limited to testing for equality -- you can't say one category label is "greater than" another with a numerical test. If you're ok with just decision trees for your classifiers, then go ahead, use category labels. Otherwise, convert those category columns!

ptressel 在4天 以前前发表

@ptressel: Great points. The "special case" equivalency is only mathematical in nature, as you point out. Mathematically, using a random variable to represent a random vector when the two are capturing the same details of the underlying probability space and are capturing the same probabilities means that the random variable version is equivalent to the random vector version. We could write a $K \times K$ matrix with all-zeros along the off-diagonal elements and with a 1 to denote the count of i^{th} element at position (i, i) and that is still an equivalent representation.

The bottom line is, the understanding of the categorical random variable is much easier when we view it as a random variable and not as some complicated vector.

sudarsanvsr_mit (员工) 在4天 以前前发表

That's for sure, sudarsanvsr_mit!!! Covering "dummy variables" in data science courses always provokes some confusion. Especially knowing when to omit the redundant column...

ptressel 在4天 以前前发表

添加评论

显示所有的回复

添加一条回复：

预览

提交