Tufts UNIVERSITY

Bruce M. Boghosian

Goodness of Fit Tests

Introduction and the Multinomial Distribution

Bruce M. Boghosian

Tufts UNIVERSITY | School of Arts and Sciences

Department of Mathematics

Tufts University

Outline

Introduction and motivation

- Most of the methods that we have looked at thus far require a priori knowledge of the probability distribution of the data that we collect.
- We are able to do interval estimation if we have a priori knowledge of the distribution.
- We are able to do hypothesis testing if we have a priori knowledge of the distribution.
- But what if we don't have a priori knowledge of the distribution?

Confirming a posited form for the distribution

- If we are using a test or estimation strategy for normally distributed data, is there at least some method to confirm that this is a reasonable assumption for our data?
- Such a method is called a *goodness-of-fit* test.
- General strategy for such tests
  - Observed data are grouped into k classes.
  - Each class's expected occupancy is calculated on the basis of the presumed model.
  - Ocupancies compared with observed occupancies.
  - If comparison indicates substantial difference, conclude that presumed $p_X(k)$ or $f_X(x)$ was incorrect.

Variants of goodness-of-fit tests

- GoF tests for distribution with specified parameters
- GoF tests for distribution with unknown parameters
- GoF tests to determine whether or not presumed independent data is indeed independent.

GoF tests and the multinomial distribution

- GoF tests are based on statistic that is asymptotically χ^2.
- That statistic derives from the *multinomial distribution*, so we study that first.
- It is a generalization of the *binomial distribution*.

The binomial distribution (yet again)

- Given n independent Bernoulli trials, each with success probability p,
- Probability that total number of successes is X,

$$p_X(k) = \text{Prob}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

for $k = 0, \ldots, n$

- This is because $\binom{n}{k}$ is the number of ways that k successes can occur in n trials.
- A clear generalization is an experiment in which any one of t outcomes can occur for each trial.

The multinomial distribution

Bruce M. Boghosian

- Given n independent trials, possible outcomes $r_1, \ldots, r_t$
- Each outcome has respective probability $p_1, \ldots, p_t$.
- These probabilities must sum to unity, $\sum_{j=1}^{t} p_j = 1$
- Suppose that outcomes are

$$X_i = \text{number of times outcome } r_i \text{ appears}$$

- For a given set of n trials, we have

$$X_1 = k_1,\ X_2 = k_2, \ldots,\ X_n = k_n,$$

where $\sum_{j=1}^{t} k_j = n$.

- Question: Can we find an expression for the quantity

$$p_{X_1,X_2,\ldots,X_t}(k_1, k_2, \ldots, k_n) = \text{Prob}\left(X_1 = k_1, X_2 = k_2, \ldots, X_t = k_t\right) ?$$

The multinomial theorem

- **Thm.:** The *multinomial distribution* for r.v.s $(X_1, X_2, \ldots, X_t)$ with parameters n and $(p_1, p_2, \ldots, p_t)$ is

$$\begin{aligned} p_{X_1,X_2,\ldots,X_t}(k_1, k_2, \ldots, k_n) =& \text{Prob}\,(X_1 = k_1, X_2 = k_2, \ldots, X_t = k_t) \\ =& \frac{n!}{k_1!k_2!\cdots k_t!} p_1^{k_1} p_2^{k_2} \cdots p_t^{k_t}. \end{aligned}$$

  where $k = 0, \ldots, n$, and $\sum_{j=1}^t p_j = 1$, and $\sum_{j=1}^t k_j = n$.
- Note this reduces to the binomial theorem when $t = 2$.
- **Pf.:** A single outcome with $k_1, k_2, \ldots, k_t$ has probability $p_1^{k_1} p_2^{k_2} \cdots p_t^{k_t}$. So we need only show that the number of possible outcomes with $k_1, k_2, \ldots, k_t$ is

$$\frac{n!}{k_1!k_2!\cdots k_t!}$$

Multinomial coefficients

- The number of ways of permuting n objects of t types is

$$\frac{n!}{k_1!k_2!\cdots k_t!},$$

where k_1 is the number of objects of type 1, k_2 is the number of objects of type 2, …, k_t is the number of objects of type t. This completes the proof. $\square$

Multinomial coefficients: An example

- **Ex.:** Number of ways to arrange n distinguishable balls is $n!$, since there are n ways to place the first ball, $n-1$ ways to place the second, etc.
- **Ex.:** Number of ways to arrange n balls, k_1 of which are red, k_2 of which are green, and k_3 of which are blue, where $n = k_1 + k_2 + k_3$.
  - Arrange as though they were distinguishable ($n!$ ways).
  - Divide by $k_1!$, number of permutations of the red balls.
  - Divide by $k_2!$, number of permutations of the green balls.
  - Divide by $k_3!$, number of permutations of the blue balls.
- Result is the multinomial coefficient

$$\frac{n!}{k_1!k_2!\cdots k_t!}.$$

Example: A loaded die I

- A die has six faces, $i = 1, \ldots, 6$.
- Suppose p_i is the probability that a roll results in face i.
- Suppose that $p_i = ci$, where c is to be determined.
- Demand

$$1 = \sum_{i=1}^{6} p_i = \sum_{i=1}^{6} ci = c \sum_{i=1}^{6} i = c\frac{6 \cdot 7}{2} = 21c,$$

so that $c = {}^{1}\!/_{21}$ and $p_i = {}^{i}\!/_{21}$.

- **Question:** The die is tossed twelve times. What is the probability that each face will appear exactly twice?

Example: A loaded die II

- The outcome we want is $X_i = k_i = 2$ for $i = 1, \ldots, 6$.
- The result is then

$$
\begin{aligned}
p_{X_1,X_2,\ldots,X_6}(2,2,\ldots,2) =& \text{Prob}\left(X_1 = 2, X_2 = 2, \ldots, X_6 = 2\right) \\
=& \frac{12!}{2!\,2!\,\cdots\,2!}\left(\frac{1}{21}\right)^2\left(\frac{2}{21}\right)^2\cdots\left(\frac{6}{21}\right)^2 \\
=& 0.00052746\ldots
\end{aligned}
$$

- Note that if die not loaded, above result would have been

$$
\frac{12!}{2!\,2!\,\cdots\,2!}\left(\frac{1}{6}\right)^{12} = 0.00343829\ldots
$$

Example: Sampling a known distribution I

Bruce M. Boghosian

- Five observations are drawn from

$$f_Y(y) = 6y(1-y)$$

- What is probability that
  - one of the observations lies in $[0, 0.25)$, and
  - none of the observations lies in $[0.25, 0.50)$, and
  - three of the observations lie in $[0.50, 0.75)$, and
  - one of the observations lies in $[0.75, 1.00]$?
- Bin the observations into one of four ranges. Range r_1 is $[0, 0.25)$, range r_2 is $[0.25, 0.50)$, range r_3 is $[0.50, 0.75)$, and range r_4 is $[0.75, 1.00]$.

Example: Sampling a known distribution II

Bruce M. Boghosian

- We have the corresponding probabilities

$$p_1 = \int_0^{1/4} dy\ 6y(1-y) = {}^5\!/_{32}$$

$$p_2 = \int_{1/4}^{1/2} dy\ 6y(1-y) = {}^{11}\!/_{32}$$

$$p_3 = \int_{1/2}^{3/4} dy\ 6y(1-y) = {}^{11}\!/_{32}$$

$$p_4 = \int_{3/4}^{1} dy\ 6y(1-y) = {}^5\!/_{32}$$

- The desired result is

$$\begin{aligned} p_{X_1,X_2,X_3,X_4}(1,0,3,1) =& \text{Prob}\left(X_1 = 1, X_2 = 0, X_3 = 3, X_4 = 1\right) \\ =& \frac{5!}{1!\,0!\,3!\,1!}\left(\frac{5}{32}\right)^1\left(\frac{11}{32}\right)^0\left(\frac{11}{32}\right)^3\left(\frac{5}{32}\right)^1 \\ =& 0.0198334\ldots \end{aligned}$$

A mutlinomial / binomial relationship

- **Def.:** Given a multivariate distribution of $(X_1, X_2, \ldots, X_t)$, the individual distribution of X_i obtained by summing over all the other values of the random variables, X_j with $j \neq i$, is called the *marginal distribution* of X_i,

$$p_{X_i}(k_i) = \sum_{j \neq i}^{t} \sum_{k_j=0}^{n} p_{X_1, X_2, \ldots, X_t}(k_1, k_2, \ldots, k_n).$$

- **Thm.:** Let $(X_1, X_2, \ldots, X_t)$ be a multinomial random variable with parameters n and $(p_1, p_2, \ldots, p_t)$. Then the marginal pdf for any single one of these random variables, X_i, is binomial with parameters n and (p_i).

A mutlinomial / binomial relationship

- **Pf.:** Dichotomize the outcomes into "result r_i" or "not result r_i". The probability of result r_i is p_i, and the probability of "not result r_i" is

$$1 - p_i = \sum_{j \neq i}^{t} p_j.$$

- Then X_i is the number of successes out of n trials, where success is defined as obtaining result r_i, which happens with probability p_i. This is, however, precisely the definition of a binomial random variable with parameters n and p_i. $\square$

A mutlinomial / binomial relationship

Bruce M. Boghosian

- In fact, the above proof generalizes.
- The marginal distribution of any subset of multinomial random variables is multinomial.
- Proof can be obtained by the same reasoning and using mathematical induction on t, using the binomial case $(t = 2)$ as the anchor case.

This theorem can be used to obtain variances

- **Ex.:** From past experience, professor expects exam grades to be normally distributed with $\mu = 80$, $\sigma = 5$. Letter grades assigned, with B given for 80 to 89. What is mean and variance of number of students receiving a B?
- Answer for mean

$$\begin{aligned} p_B =& \text{Prob}\,(80 \le Y \le 89) \\ =& \text{Prob}\left(\frac{80-80}{5} \le \frac{Y-80}{5} \le \frac{89-80}{5}\right) \\ =& \text{Prob}\,(0 \le Z \le 0.8) \\ =& 0.4772 \end{aligned}$$

- Answer for variance

$$\text{Var}(X_B) = np_B\,(1 - p_B) = 12.47$$

Summary

- We have motivated and described GoF tests.
- We have related GoF tests to the multinomial distribution.
- We have derived the mulinomial distribution and given its probability distribution.
- We have related its algebraic form to the combinatoric problem of permutations of n objects, each of t categories.
- We have shown the marginal distribution of a multinomial r.v. is binomial.
- We have given several examples.