9-Goodness-of-Fit Tests

ENSY SILVER[1]

Wednesday 9th September, 2020

[1]Thanks to my family, my friend and freedom.

1 Introduction

In general, any procedure that seeks to determine whether a set of data could reasonably have originated from some given probability distribution, or class of probability distributions, is called a **goodness-of-fit** test. The principle behind the particular *goodness-of-fit* test we will look at is very straightforward: First the observed data are grouped, more or less arbitrarily, into k classes; then each class' s "expected" occupancy is calculated on the basis of the presumed model. If it should happen that the set of observed and expected frequencies shows considerably more disagreement than sampling variability would predict, our conclusion will be that the supposed $p_X(k)$ or $f_Y(y)$ was incorrect.

2 The multinomial distribution

To test the distribution of grouped data, we first need to introduce the multinomial distribution.

Theorem 2.1. *Let X_i denote the number of times that the outcome r_i occurs, $i = 1, 2, \cdots, t$, in a series of n independent trials, where $p_i = P(r_i)$. Then the vector $(X_1, X_2, \cdots, X_t)$ has a multinomial distribution and*

$$p_{X_1,X_2,\cdots,X_t}(k_1, k_2, \cdots, k_t) = \frac{n!}{k_1!k_2!\cdots k_t!} p_1^{k_1} p_2^{k_2} \cdots p_t^{k_t}$$

where $0 \leq k_i \leq n$ and $\sum_{i=1}^{t} k_i = n$.

2.1 Relation between multinomial and binomial distribution

Theorem 2.2. *Suppose the vector $(X_1, X_2, \cdots, X_t)$ is a multinomial random variable with parameters n, $p_1, p2, \cdots, p_t$. Then the marginal distribution of X_i, $i = 1, 2, \cdots, t$, is the binomial pdf with parameters n and p_i .*

3 Goodness-of-Fit tests: the procedure

Suppose we have known a distribution with its parameters, and we also have a set of data. In order to test whether the data fits the distribution, we partition the data into several groups, then the outcome from the distribution has a probability p_i falling into the group r_i, and we convert the problem to measuring the difference between the data and the expected multinomial distribution.

4 Goodness-of-Fit tests: all parameters known

Theorem 4.1. *Let $r_1, r_2, \cdots, r_t$ be the set of possible outcomes (or ranges of outcomes) associated with each of n independent trials, where $P(r_i) = p_i$, $i = 1, 2, \cdots, t$. Let X_i = number of times r_i occurs, $i = 1, 2, \cdots, t$. Then*

1. *The random variable*
$$D = \sum_{i=1}^{t} \frac{(X_i - np_i)^2}{np_i}$$
has approximately a χ^2 distribution with $t-1$ degrees of freedom. For the approximation to be adequate, the t classes should be so defined so that $np_i \geq 5$, for all i.

2. *Let $k_1, k_2, \cdots, k_t$ be the observed frequencies for the outcomes $r_1, r_2, \cdots, r_t$, respectively, and let $n\tilde{p}_1, n\tilde{p}_2, \cdots, n\tilde{p}_t$ be the corresponding expected frequencies based on the null hypothesis. At the α level of significance, $H_0 : f_Y(y) = f_{expected}(y)$ is rejected if*
$$d = \sum_{i=1}^{t} \frac{(k_i - n\tilde{p}_i)^2}{n\tilde{p}_i} \geq \chi^2_{1-\alpha,t-1}$$
where $n\tilde{p}_i \geq 5$ for all i.

4.1 An exception

Sometimes, researchers falsify their data, making the data too good to be true. In this case, we test the data and reject the null hypothesis if $d \leq \chi^2_{1-\alpha,t-1}$.

5 Goodness-of-Fit tests: parameters unknown

Similar to theorem 4.1, we have the theorem in the case with unknown parameters.

Theorem 5.1. *Suppose that a random sample of n observations is taken from $f_Y(y)$ [or $p_X(k)$], a pdf having s unknown parameters. Let $r_1, r_2, \cdots, r_t$ be a set of mutually exclusive ranges (or outcomes) associated with each of the n observations. Let $\hat{p}_i$ = estimated probability of r_i, $i = 1, 2, \cdots, t$. Let X_i denote the number of times that r_i occurs, $i = 1, 2, \cdots, t$.*

1. *The random variable*
$$D = \sum_{i=1}^{t} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i}$$
has approximately a χ^2 distribution with $t-1-s$ degrees of freedom. For the approximation to be adequate, the t classes should be so defined so that $np_i \geq 5$, for all i.

2. *Let $k_1, k_2, \cdots, k_t$ be the observed frequencies for the outcomes $r_1, r_2, \cdots, r_t$, respectively, and let $n\hat{p}_1, n\hat{p}_2, \cdots, n\hat{p}_t$ be the corresponding expected frequencies*

based on the null hypothesis. At the α level of significance, $H_0 : f_Y(y) = f_{expected}(y)$ is rejected if

$$d = \sum_{i=1}^{t} \frac{(k_i - n\hat{p_i})^2}{n\hat{p_i}} \geq \chi^2_{1-\alpha,t-1-s}$$

where $n\hat{p_i} \geq 5$ for all i.

6 Contingency tables

Theorem 6.1. *Suppose that n observations are taken on a sample space partitioned by the events $A_1, A_2, \cdots, A_r$ and also by the events $B_1.b_2, \cdots, B_c$. Let $p_i = P(A_i), q_j = P(B_j)$, and $p_{ij} = P(A_i \cap B_j)$, $i = 1, 2, \cdots, r$; $j = 1, 2, \cdots, c$. Let X_{ij} denote the number of observations belonging to the intersection $A_i \cap B_j$. Then*

1. *The random variable*

$$D = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(X_{ij} - np_{ij})^2}{np_{ij}}$$

has approximately a χ^2 distribution with $rc-1$ degrees of freedom (provided $np_{ij} \geq 5$ for all i and j) .

2. *to test H_0: the A_i' s are independent of the B_j' s, calculate the test statistic*

$$d = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(X_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}}$$

The null hypothesis should be rejected at the α level of significance if

$$d \geq \chi^2_{1-\alpha,(r-1)(c-1)}$$

7 Degrees of freedom

In general, the number of degrees of freedom associated with a goodness-of-fit statistic is given by the formula

degrees of freedom = number of classes − 1 − number of estimated parameters