Categorical Data and Nonparametric Methods

March 4, 2017

1 Concepts of nonparametric inference

Nonparametric statistics fall into two categories: (a) procedures that do not involve or depend on parametric assumptions, though the underlying population distribution may belong to a particular parametric family; and (b) methods that do not require that the data belong to a particular parametric family of distributions. Distribution procedures such as the KS test, rank statistics, the sign and Wilcoxon signed rank tests are in the first category. Contingency tables and the variety of density estimation methods (histograms, kernel smoothing, nearest neighbor and nonparametric regressions) are in the second category. In the latter group of cases, the structure of the relationship between variables is treated nonparametrically, while there may be parametric assumptions about the distribution of model residuals. The term semi-parametric is sometimes used for procedures which combine parametric modeling with principles of nonparametrics.

Some nonparametric procedures are analogous to parametric procedures but operate on the ranks, or numbered position in the ordered sequence of data points, rather than the

measured values of the data points. Rank tests are often the most powerful available for classificatory variables which are ordered but not with meaningful numerical values. Note, however, that ranks cannot be reliably defined for multivariate datasets. Bayesian non-parametrics can be viewed as an oxymoron as Bayesian inference requires a mathematical probability model for the data in terms of well-defined parameters.

2 Tests of Goodness-of-Fit

Problems in which the possible distributions of the observations are not restricted to a specific parametric family are called nonparametric problems, and the statistical methods that are applicable in such problems are called nonparametric methods.

2.1 The χ^2 Test

Suppose that a large population consists of items of k different types, and let p_i denote the probability that an item selected at random will be of type $i (i = 1, \cdots, k)$. $p_i \geqslant 0$ for $i = 1, \cdots, k$ and $\sum_{i=1}^{k} p_i = 1$. Let $p_1^0, \cdots, p_k^0$ be specific numbers such that $p_i^0 > 0$ for $i = 1, \cdots, k$ and $\sum_{i=1}^{k} p_i^0 = 1$, and suppose that the following hypotheses are to be tested:

$$H_0 : p_i = p_i^0 \quad \text{for } i = 1, \cdots, k \ , \tag{1}$$

$$H_1 : p_i \neq p_i^0 \quad \text{for at least one value of } i \ . \tag{2}$$

We shall assume that a random sample of size n is to be taken from the given population. That is, n independent observations are to be taken, and there is probability p_i that each observation will be of type $i (i = 1, \cdots, k)$. For $i = 1, \cdots, k$, let N_i denote the number of observations in the random sample that are of type i. Thus, $N_1, \cdots, N_k$ are

nonnegative integers such that $\sum_{i=1}^{k} N_i = n$. $(N_1, \cdots, N_k)$ has the multinomial distribution with parameters n and $p = (p_1, \cdots, p_k)$. When the null hypothesis H_0 is true, the expected number of observations of type i is $np_i^0 (i = 1, \cdots, k)$. The difference between the actual number of observations N_i and the expected number np_i^0 will tend to be smaller when H_0 is true than when H_0 is not true. It seems reasonable, therefore, to base a test of the above hypotheses on values of the differences $N_i - np_i^0$ for $i = 1, \cdots, k$ and reject H_0 when the magnitudes of these differences are relatively large.

Theorem 2.1: χ^2 Statistic

The following statistic

$$Q = \sum_{i=1}^{k} \frac{(N_i - np_i^0)^2}{np_i^0} \tag{3}$$

has the property that if H_0 is true and the sample size $n \to \infty$, then Q converges in distribution to the χ^2 distribution with $k - 1$ degrees of freedom.

If H_0 is true and the sample size n is large, the distribution of Q will be approximately the χ^2 distribution with $k - 1$ degrees of freedom.

3 Goodness-of-Fit for Composite Hypotheses

4 Contingency Tables

5 Tests of Homogeneity

6 Simpsons Paradox

7 Kolmogorov-Smirnov Tests

The Kolmogorov-Smirnov test (KS-test) tries to determine if two datasets differ significantly. The KS-test has the advantage of making no assumption about the distribution of data. (Technically speaking it is non-parametric and distribution free.) However this generality comes at some cost: other tests (for example Student's t-test) may be more sensitive if the data meet the requirements of the test.

In a typical experiment, data collected in one situation (let's call this the control group) is compared to data collected in a different situation (let's call this the treatment group) with the aim of seeing if the first situation produces different results from the second situation. If the outcomes for the treatment situation are "the same" as outcomes in the control situation, we assume that treatment causes no effect. Rarely are the outcomes of the two groups identical, so the question arises: How different must the outcomes be? Statistics aim to assign numbers to the test results; P-values report if the numbers differ significantly. Reject the null hypothesis if P is "small".

Kolmogorov-Smirnov Tests can be used to not only test the null hypothesis that a random sample came from a particular continuous distribution against the alternative hypothesis that the sample did not come from that distribution, but also test the null hypothesis that two independent samples came from the same distribution against the alternative hypothesis that they came from two different distributions.

7.1 The Sample Distribution Function

Construct an estimator of the distribution of the random sample that does not rely on the assumption that the distribution was normal. Suppose that the random variables $X_1, \cdots, X_n$ form a random sample from some continuous distribution, and let $x_1, \cdots, x_n$ denote the observed values of $X_1, \cdots, X_n$. Since the observations come from a continuous distribution, there is probability 0 that any two of the observed values $x_1, \cdots, x_n$ will be equal. Therefore, we shall assume for simplicity that all n values are different. We shall consider now a function $F_n(x)$, which is constructed from the values $x_1, \cdots, x_n$ and will serve as an estimate of the c.d.f. from which the sample was drawn.

Let $x_1, \cdots, x_n$ be the observed values of a random sample $X_1, \cdots, X_n$. For each number $x(\infty < x < \infty)$, define the value $F_n(x)$ as the proportion of observed values in the sample that are less than or equal to x. In other words, if exactly k of the observed values in the sample are less than or equal to x, then $F_n(x) = k/n$. The function $F_n(x)$ defined in this way is called the sample distribution function, or simply the sample c.d.f. Sometimes $F_n(x)$ is called the empirical c.d.f.

Theorem 7.1: Glivenko-Cantelli Lemma

Let F_n be the sample c.d.f. from an i.i.d. sample $X_1, \cdots, X_n$ from the c.d.f. F. Define

$$D_n = \sup_{-\infty<x<\infty} |F_n(x) - F(x)| \ . \tag{4}$$

Then $D_n \xrightarrow{P} 0$.

Before the values of $X_1, \cdots, X_n$ have been observed, the value of D_n is a random variable. When the sample size n is large, the sample c.d.f. $F_n(x)$ is quite likely to be close to the c.d.f. $F(x)$ over the entire real line. In this sense, when the c.d.f. $F(x)$ is unknown, the sample c.d.f. $F_n(x)$ can be considered to be an estimator of $F(x)$.

7.2 The Kolmogorov-Smirnov Test of a Simple Hypothesis

test the simple null hypothesis that the unknown c.d.f. $F(x)$ is actually a particular continuous c.d.f. $F^\star(x)$ against the general alternative that the actual c.d.f. is not $F^\star(x)$, i.e. test the following hypotheses:

$$H_0 : F(x) = F^\star(x) \quad \text{for} \quad -\infty < x < \infty \ , \tag{5}$$

$$H_1 : \text{The hypothesis } H_0 \text{ is not true.} \tag{6}$$

This problem is a nonparametric problem because the unknown distribution from which the random sample is taken might be any continuous distribution. The χ^2 test of goodness-of-fit can be used to test above hypotheses. That test, however, requires grouping the observations into a finite number of intervals in an arbitrary manner.

8 Robust Estimation

9 Sign and Rank Tests

10 Univariate problems

10.1 KolmogorovSmirnov and other e.d.f. tests

The empirical distribution function (e.d.f.) is the simplest and most direct nonparametric estimator of the underlying cumulative distribution function (c.d.f.) for the underlying population. The univariate dataset $X_1, X_2, \cdots, X_n$ is assumed to be drawn as independently and identically distributed (i.i.d.) samples from a common distribution function F. The e.d.f. $\hat{F}_n$ is defined to be

$$\hat{F}_n = \frac{1}{n} \sum_{i=1}^{n} I[X_i \leqslant x] \ , \tag{7}$$

for all real numbers x. The e.d.f. thus ranges from 0.0 to 1.0 with step heights of $1/n$ located at the values X_i. For each x, $F_n(x)$ follows the binomial distribution which is asymptotically normal. The mean and variance of $\hat{F}_n(x)$ are

$$\begin{aligned} E[\hat{F}_n(x)] &= F(x) \\ Var[\hat{F}_n(x)] &= \frac{F(x)[1 - F(x)]}{n} \ . \end{aligned}$$

For a chosen significance level α, the $100(1-\alpha)\%$ asymptotic confidence interval for $\hat{F}_n(x)$ is

$$\hat{F}_n(x) \pm z_{1-\alpha/2} \sqrt{\hat{F}_n(x)[1 - \hat{F}_n(x)]/n} \tag{8}$$

where z are the quantiles of the Gaussian distribution.

The e.d.f. is an unbiased and consistent estimator of the population distribution function, and is the generalized maximum likelihood estimator of the population c.d.f. The e.d.f. $\hat{F}_n$ uniquely, completely and accurately embodies all information in the measured X_i without any ancillary choices.

Several statistics have been developed to assist inference on the consistency of an observed e.d.f. with a model specified in advance. We want to test the null hypothesis that $F(x) = F_0(x)$ for all x, against the alternative that $F(x) \neq F_0(x)$ for some x, where F_0 is a distribution specified independently of the dataset under study. One-sample KolmogorovSmirnov (KS) test measures the maximum distance between the e.d.f and the model,

$$M_{\mathrm{KS}} = \sqrt{n} \max_x |\hat{F}_n(x) - F_0(x)| \ . \tag{9}$$

A large value of the supremum M_{KS} allows rejection of the null hypothesis that $F = F_0$ at a chosen significance level. The distribution of M_{KS} is independent of the shape of F as long as it is a continuous function. For large n and a chosen significance level α (e.g. $\alpha = 0.05$), the cumulative distribution of the one-sample KS statistic is approximately

$$P_{\mathrm{KS}}(M_{\mathrm{KS}} > x) \simeq 2 \sum_{r=1}^{\infty} (-1)^{r-1} e^{-2r^2x^2} \tag{10}$$

with critical value

$$M_{\mathrm{KS}}^{\mathrm{crit}} > \left(-\frac{1}{2} \ln \left(\frac{\alpha}{2} \right) \right)^{1/2} \ . \tag{11}$$

These results are based on advanced weak convergence theory. For small samples, tables of $M_{\mathrm{KS}}^{\mathrm{crit}}$ or bootstrap simulations must be used.

The KS test is not distribution-free - so the widely tabulated critical values of the KS

statistic are not valid - if the model parameters were estimated from the same dataset being tested. The critical values are only correct if the model parameters (except for normalization which is removed in the construction of the dataset e.d.f. and model c.d.f.) are known in advance of the dataset under consideration.

The distribution of the KS statistic is also not distribution-free when the dataset has two or more dimensions. The reason is that a unique ordering of points needed to construct the e.d.f. cannot be defined in multivariate space. A two-dimensional KS-type statistic can be constructed and has been used fairly often in astronomy. But the distribution of this statistic is not knowable in advance and is not distribution-free; probabilities would have to be calculated for each situation using bootstrap or similar resampling techniques.

The KS test is sensitive to global differences between two e.d.f.s or one e.d.f. $\hat{F}_n$ and the model c.d.f. F_0 producing different mean values. But the test is less efficient in uncovering small-scale differences near the tails of the distribution.

The Cramér-von Mises (CvM) statistic, $T_{\rm CvM}$, measures the sum of the squared differences between $\hat{F}_n$ and F_0,

$$\begin{aligned} T_{{\rm CvM},n} &= n\int_{-\infty}^{\infty}[\hat{F}_n(x)-F_0(x)]^2 {\rm d}F_0(x) \\ &= \frac{1}{12n}+\sum_{i=1}^{n}\left(\frac{2i-1}{2n}-F(X_{(i)})\right)^2 \end{aligned} \tag{12}$$

It captures both global and local differences between the data and model, and thus often performs better than the KS test. Two-sample KS and CvM tests are also commonly used, where the e.d.f. of one sample is compared to the e.d.f. of another sample rather than the c.d.f of a model distribution where $X_{(i)}$ is the i-th entry when $X_1, X_2, \cdots, X_n$ are placed in increasing order.

Anderson-Darling (AD) statistic, $A^2_{\rm AD}$, a weighted variant of the CvM statistic,

$$A^2_{\rm AD} = n \sum_{i=1}^{n} \frac{[i/n - F_0(X_i)]^2}{F_0(X_i)(1 - F_0(X_i))} \ . \tag{13}$$

The AndersonDarling test is more effective than other e.d.f. tests, and is particularly better than the KolmogorovSmirnov test, under many circumstances.

10.2 Robust statistics of location

11 Non-parametric tests: single samples

'Non-parametric tests implies that no distribution is assumed'. Various tests exploit different things, but a common method is to use counting probabilities. In the chi-square test, the number of items in bin i is N_i, and we expect E_i. For smallish numbers, Poisson statistics tell us that the variance is also E_i. So $(N_i - E_i)^2/E_i$ should be roughly a squared Gaussian variable, of unit variance. The runs test is just using the assumption that each successive observation is equally likely to be 'up' or 'down', so a Binomial distribution applies. The assumptions underlying non-parametric tests are weaker, and so more general, than for parametric tests.

These make fewer assumptions about the data. If indeed the underlying distribution is unknown, there is no alternative.

If the sample size is small, probably we must use a non-parametric test.

The non-parametric tests can cope with data in non-numerical form, e.g. ranks, classifications. There may be no parametric equivalent.

Non-parametric tests can treat samples of observations from several different populations.

11.1 Chi-square test

Consider observational data which can be binned, and a model/ hypothesis which predicts the population of each bin. The chi-square statistic describes the goodness-of-fit of the data to the model. If the observed numbers in each of k bins are O_i, and the expected values from the model are E_i, then this statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \ . \tag{14}$$

The null hypothesis H_0 is that the number of objects falling in each category is E_i; the chi-square procedure tests whether the O_i are sufficiently close to E_i to be likely to have occurred under H_0. The sampling distribution under H_0 of the statistic χ^2 follows the chi-square distribution with $\nu = (k-1)$ degrees of freedom. One degree of freedom is lost because of the constraint that $\sum_i O_i = \sum_i E_i$. The chi-square distribution is given by

$$f(x) = \frac{2^{-\nu/2}}{\Gamma[\nu/2]} x^{\nu/2-1} \mathrm{e}^{-x/2} \ , \tag{15}$$

for $x \geqslant 0$, the distribution function of the random variable $Y^2 = Z_1^2 + Z_2^2 + \cdots + Z_\nu^2$ where the Z_i are independent random variables of the standard Normal distribution. If χ^2 exceeds these values, H_0 is rejected at that level of significance.

The premise of the chi-square test is that the deviations from E_i are due to statistical fluctuations from limited numbers of observations per bin, i.e. 'noise' or Poisson statistics, and the chi-square distribution simply gives the probability that the chance deviations from E_i are as large as the observations O_i imply. We need enough data per bin to ensure that each term in the chi-square summation is approximately Gaussian.

The mean of the chi-square distribution equals the number of degrees of freedom, while the variance equals twice the number of degrees of freedom. If χ^2 should come out (for

more than four bins) as $\sim$ (number of bins -1) then accept H_0. But if χ^2 exceeds twice (number of bins -1), probably H_0 will be rejected.

The data must be binned to apply the test, and the bin populations must reach a certain size because it is obvious that instability results as $E_i \to 0$. And > 80 per cent of the bins must have $E_i > 5$. However, the binning of data in general, and certainly the binning of bins, results in loss of efficiency and information, resolution in particular. The chi-square test cannot tell direction, i.e. it is a two-tailed test; it can only tell whether the differences between sample and prediction exceed those which can be reasonably expected on the basis of statistical fluctuations due to the finite sample size.

11.2 Kolmogorov-Smirnov one-sample test

Calculate $S_{\rm e}(x)$, the predicted cumulative (integral) frequency distribution under H_0.

Consider the sample of N observations, and compute $S_{\rm o}(x)$, the observed cumulative distribution, the sum of all observations to each x divided by the sum of all N observations.

Find

$$D = \max|S_{\rm e}(x) - S_{\rm o}(x)| \tag{16}$$

Consult the known sampling distribution for D under H_0 to determine the fate of H_0. If D exceeds a critical value at the appropriate N, then H_0 is rejected at that level of significance.

As for the chi-square test, the sampling distribution indicates whether a divergence of the observed magnitude is 'reasonable' if the difference between observations and prediction is due solely to statistical fluctuations.

The Kolmogorov-Smirnov test treats the individual observations separately, and no infor-

mation is lost because of grouping. It works for small samples; for very small samples it is the only alternative. For intermediate sample sizes it is more powerful. The Kolmogorov-Smirnov test is non-directional or two-tailed, as is the chi-square test. However, a method of finding probabilities for the one-tailed test does exist.

To apply the Kolmogorov-Smirnov test, the distributions must be continuous functions of the variable. The chi-square test is applicable to data which can be simply binned, grouped, categorized - there is no need for measurement on a numerical scale.

12 Non-parametric tests: two independent samples

Suppose we have two samples, we want to know whether they could have been drawn from the same population, or from different populations, and if the latter, whether they differ in some predicted direction.

12.1 Chi-square two-sample (or k-sample) test

Each sample is binned in the same r bins (a $k \times r$ contingency table). H_0 is that the k samples are from the same population. Then compute

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \ . \tag{17}$$

The E_{ij} are the expectation values, computed from

$$E_{ij} = \frac{\sum_{j=1}^{k} O_{ij} \cdot \sum_{i=1}^{r} O_{ij}}{\sum_{i=1}^{r} \sum_{j=1}^{k} O_{ij}} \ . \tag{18}$$

Under H_0 this is distributed as χ^2, with $(r-1)(k-1)$ degrees of freedom.

If there are only 2×2 cells, the total (N) must exceed 30; if not, use the Fisher exact probability test.

12.2 Kolmogorov-Smirnov two-sample test

The maximum deviation between the cumulative distributions of two samples with m and n members. H_0 is that the two samples are from the same population, and H_1 can be that they differ (two-tailed test), or that they differ in a specific direction (one-tailed test). Exchange the cumulative distributions S_e and S_o for S_m and S_n corresponding to the two samples.

For large samples, one-tailed test, compute

$$\chi^2 = 4D^2 \frac{mn}{m+n} \tag{19}$$