statistical description of data

Numerical Methods for IT

introduction

The analysis of data inevitably involves some trafficking with the field of statistics.

Nếu một thống kê rơi vào một phần hợp lý của phân phối, chúng ta không được mắc sai lầm khi kết luận rằng giả thuyết vô hiệu đã được xác minh hoặc chứng minh. Đó là lời nguyền của thống kê, rằng nó không bao giờ có thể chứng minh mọi thứ, chỉ có thể bác bỏ chúng.

Moments of a distribution: mean, median, mode

When a set of values has a sufficiently strong central tendency then it may be useful to characterize the set by a few numbers that related to its moments, the sums of integer powers of the values.

. Best know is the mean of values: $x_1, \dots, x_N \colon < x > \equiv \bar{x} = \frac{1}{N}\Sigma_{j=1}^{N} x_j$.

. For values drawn from a probability distribution with very broad tail, the mean may converge poorly, or not at all, as the number of sampled point is increased.

. The median of distribution function $p(x)$ is the value x_{med} for which larger or smaller values of x are equally probable: $\int_{-\infty}^{x} p(x)dx = \frac{1}{2} = \int_{x_{med}}^{\infty} p(x)dx$. The median of a distribution is estimated from a sample of values by finding the value x_i which has equal numbers of values above it and below it: $x_{med} = \begin{cases} x_{(N-1)/2}, & N\colon odd \\ \frac{1}{2}(x_{N/2-1} + x_{N/2}), & N\colon even \end{cases}$.

. The mode of the distribution $p(x)$ is the values x where it takes on a maximum value. The mode is useful primarily when where is a single, sharp maximum, in which case it estimates the central value. Occasionally, a distribution will be bimodal, with two relative maxima; the one may wish to know the two modes individually.

Moments of a distribution: variance

Having characterized a distribution's value, one conventionally next characterizes its "width" or "variability" around the value. Again, more than one measure is avaiable. Most common is the variance: $Var(x_1, \dots, x_N) = \frac{1}{N-1} \Sigma_{j=1}^{N} \left(x_j - x\right)^2$.

Square root of variance, the standard deviation: $\sigma(x_1, \dots, x_N) = \sqrt{Var(x_1, \dots, x_N)}$.

Why $\frac{1}{N-1}$ in the variance formular? It is a long story!
Here we will be content to note that the N - 1 should be changed to N if you're ever in the situation of measuring the variance the variance of a distribution whose mean is know priori rather than being estimated from the data.
A more robust estimator of the width is the average deviation or mean absolute deviation:
$ADev(x_1, \dots, x_N) = \frac{1}{N} \Sigma_{j=1}^{N} |x_j - \bar{x}|$.
One often substitutes the sample median x_{med} for $\bar{x}$. For any fixed sample, the median in fact minimizes the mean absolute deviation.

Moments of a distribution: skewness, kurtosis

Skewness or third moment characterizes the degree of asymmetry of a distribution around its mean.
While the mean, standard deviation, and average deviation are dimensional quantities, that is, have a same units as the measured quantities, the skewness is conventionally defined in such away as to make it nondimensional. It is a pure number that characterizes only the shape of the distribution.

The usual definition: $Skew(x_1, \ldots, x_N) = \frac{1}{N} \Sigma_{j=1}^{N} \left(\frac{x-\bar{x}}{\sigma}\right)^3$. Skew > 0: asymmetric tail extending out toward more positive x_j; otherwise, toward more negative x_j.
The Kurtosis is also nondimensional quantities. It measures the relative peaked-ness or flat-nes of a distribution.

The conventionally definition:$Kurt(x_1, \ldots, x_N) = \frac{1}{N} \Sigma_{j=1}^{N} \left(\frac{x-\bar{x}}{\sigma}\right)^4 - 3$, where tẻm -3 makes the value zero for a normal distribution.

2 distributions have the same mean/variance

Not uncommonly we want to know whether two distributions have the same mean?

Ex: a 1st set of measured values may have been gathered before some event, a 2nd set after it. We want to know whether the event, a “treatment” or change in a control parameter””, made a difference?

Our 1st thought is to ask "how many standard deviations" one sample mean from the other? We will be meeting distinct concepts of *strength* and *significance*.

A quantity that measures the significance of a difference of mean is not the number of standard deviations that they are apart, but the number of so-called standard-errors that they are apart. The standard error of of a set of values measures accuracy with which the sample estimates the population mean.

t-test for the significantly different means

When the 2 distributions have the same, but possibly different means, the t-test compared as follows:

(1) Estimate the standard error of the difference of the means:

$S_D = \sqrt{\frac{\Sigma_{i \in A}(x_i - \bar{x}_A)^2 + \Sigma_{i \in B}(x_i - \bar{x}_B)^2}{N_A + N_B - 2}\left(\frac{1}{N_A} + \frac{1}{N_B}\right)}$, where each sum is over the points in one sample; and N_A, N_B are the numbers of points in the 1st and the 2nd samples.

(2) Compute t by $t = \frac{\bar{x}_A - \bar{x}_B}{S_D}$.

(3) Evaluate the p-value or significance of this value of t with $N_A + N_B - 2$ degree of freedom.

The p-value is a number between 0 and 1. It is the probability that $|t|$ could be this large or larger just by chance, for distributions with equal means. Therefor, a small numerical value of the p-value (0.01 or 0.001) means that the observed difference is “very significant”.

t-test for the significantly different variances

To find out whether the two datasets have variances that are different. The relevant statistic for unequal-variance t-test:

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{Var(x_A)}{N_A} + \frac{Var(x_B)}{N_B}}}.$$

Chi-square test

Suppose that N_i is the number of events observed in the ith bin, and that n_i is the number expected according to some know distribution. Note that N_i's are integers, while n_i's may not be. Then the chi-square statistic is: $\chi^2 = \frac{\Sigma_i (N_i - n_i)^2}{n_i}$, where the sum is over all bin.
A large value of χ^2indicates that the null hypothesis (that the N_i's are dran from the population represented by the n_i's.

In the case of comparing two binned datasets. Let R_i be the number of events in bin I for the first dataset and S_i the number of events in the the same bin I for the second dataset. Then the chi-square statistic is $\chi^2 = \Sigma_i (R_i - S_i)^2/(R_i + S_i)$.