divergence tests of goodness of fit

goodness of fit tests of hypothetical multivariate discrete distributions
(as suggested by association graphs)

$p =$ general model based on empirical distribution with estimated likelihood function $L(p)$

$q =$ data follows a specified probability model with estimated likelihood function $L(q)$

☑ log likelihood ratio test statistic with d degrees of freedom (for large n)

$$2 \log \frac{L(p)}{L(q)} = 2nD(p,q) \underset{approx}{\sim} \chi^2 \left(d \right)$$

where

$D(p,q)$ is the information divergence (expected log likelihood ratio) with

$d = d(p) - d(q)$ degrees of freedom (numbers of parameters estimated to get p and q)

☑ critical region with approximately 95% confidence level (for large n)

$$\chi^2(d) \geq d + 2\sqrt{2d} = d + \sqrt{8d}$$

divergence tests of goodness of fit

testing uniform distribution

of random variable X with n observations on r_X outcomes

p = model based on empirical distribution $p(x) = n(x)/n$ (the relative frequencies) with $d(p) = r_X - 1$

$q = X$ is uniformly distributed on r_X outcomes with $d(q) = 0$

☑ log likelihood ratio test statistic

$$\begin{aligned}\chi^2(r_X - 1) &= 2nD(p, q) \\ &= 2n[\log r_X - H(X)]\end{aligned}$$

where $H(X)$ is the empirical entropy of X

☑ uniformity is rejected if

$$\chi^2(r_X - 1) \geq r_X - 1 + \sqrt{8(r_X - 1)}$$

or if $H(X)$ deviates from its maximum values $\log(r_X)$ by more than $[r_X - 1 + \sqrt{8(r_X - 1)}]/2n$