Chi Square Goodness-of-fit tests

Goodness-of-fit tests are devised for checking the hypothesis
H_0: "sample was taken from distribution (family of distributions) F" versus
H_1: "sample was *not* taken from F".
The most popular goodness-of-fit tests are Chi-square and Kolmogorov-Smirnov:

- **Chi-square test.**
  This test should be used in two cases. First, for samples with the discrete values. Second, for samples with continuous values when parameter(s) of the distribution F is (are) estimated from the sample. (Eg. the sample average is used to estimate the parameter of an exponential distribution.)

  In order to perform the test we should assign our data to intervals (as in the case of a histogram). Assume there are n intervals. Let $\{\nu_i\}_{i=1}^n$ denote *sample frequencies* (histogram values). Let p_i denote probabilities that a random variable with distribution F takes value in the interval number i. Then Np_i are *theoretical frequencies*. It is recommended to assign intervals in such a way that $Np_i > 5$ for all i.

  The chi-square statistics is equal to

$$\chi^2 = \sum_{i=1}^{n} \frac{(\nu_i - Np_i)^2}{Np_i}$$

  If hypothesis H_0 is true, χ^2 is distributed according to $\chi^2(n-1-p)$ distribution, where p is the number of parameters that we estimate from data. (Eg. $p = 1$ if we estimate the rate of exponential distribution and $p = 2$ if we estimate m and σ of the normal distribution.)

  Use the Excel function CHIDIST to calculate p-values. We should reject H_0 if the p-value is smaller than *significance level* α ($\alpha = 0.05$ is conventional).

  Figure 4 and Table 1 below illustrate the results of a chi-square test for
  H_0 : "service times are exponentially distributed".
  (Compare with Figure 1: we plot another histogram for the same data.) Since $n = 6$ and $p = 1$ (average of the exponential distribution was estimated from data), the distribution $\chi^2(4)$ should be used.

Figure 4

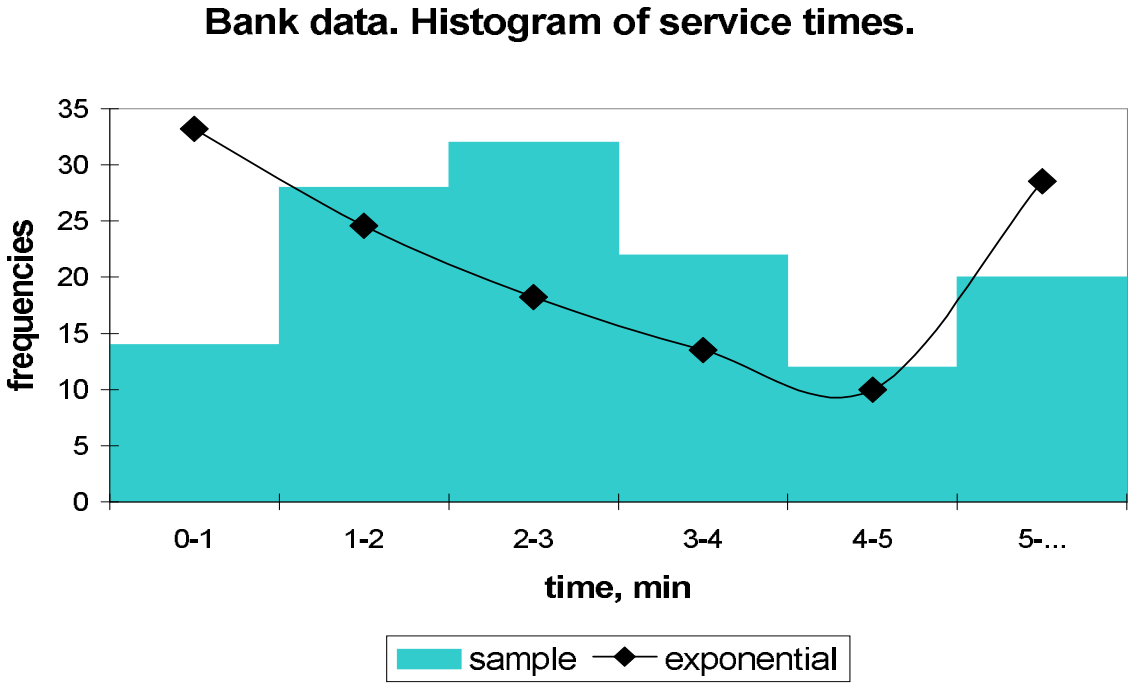

Table 1

Interval	Theoretical	Frequency	
0-1	33.18	14	11.09
1-2	24.58	28	0.48
2-3	18.21	32	10.45
3-4	13.49	22	5.37
4-5	9.99	12	0.40
5-...	28.55	20	2.56
Chi-square			30.35
p-value			1.16E-06

The hypothesis H_0 is rejected. Note that CV is close to 1, but the distribution pattern is clearly non-exponential.

- **Kolmogorov-Smirnov test.** This test works better for continuous distributions with apriori specified parameters (no parameters are estimated from data). The test statistics is based on the difference between the empirical cdf and the theoretical distribution function.