Chapter 21

GOODNESS OF FITS TESTS

In point estimation, interval estimation or hypothesis test we always started with a random sample $X_1, X_2, ..., X_n$ of size n from a known distribution. In order to apply the theory to data analysis one has to know the distribution of the sample. Quite often the experimenter (or data analyst) assumes the nature of the sample distribution based on his subjective knowledge.

Goodness of fit tests are performed to validate experimenter opinion about the distribution of the population from where the sample is drawn. The most commonly known and most frequently used goodness of fit tests are the Kolmogorov-Smirnov (KS) test and the Pearson chi-square (χ^2) test. There is a controversy over which test is the most powerful, but the general feeling seems to be that the Kolmogorov-Smirnov test is probably more powerful than the chi-square test in most situations. The KS test measures the distance between distribution functions, while the χ^2 test measures the distance between density functions. Usually, if the population distribution is continuous, then one uses the Kolmogorov-Smirnov where as if the population distribution is discrete, then one performs the Pearson's chi-square goodness of fit test.

21.1. Kolmogorov-Smirnov Test

Let $X_1, X_2, ..., X_n$ be a random sample from a population X. We hypothesized that the distribution of X is $F(x)$. Further, we wish to test our hypothesis. Thus our null hypothesis is

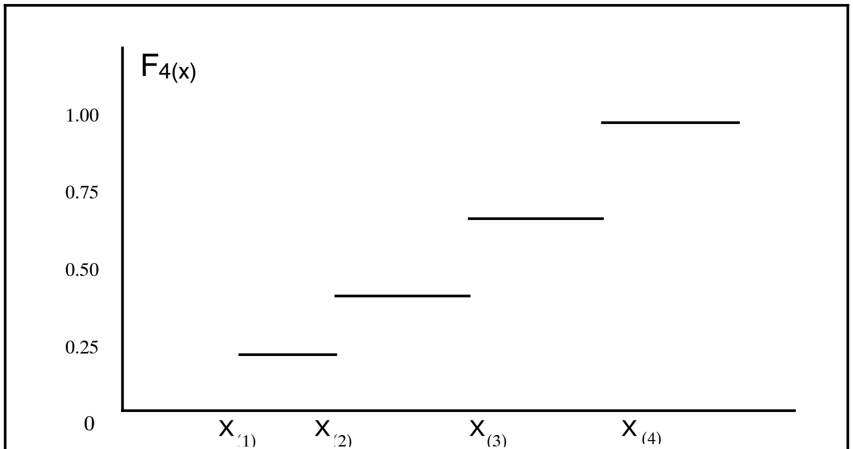
$$H_o : X \sim F(x).$$

We would like to design a test of this null hypothesis against the alternative $H_a : X \not\sim F(x)$.

In order to design a test, first of all we need a statistic which will unbiasedly estimate the unknown distribution $F(x)$ of the population X using the random sample $X_1, X_2, ..., X_n$. Let $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ be the observed values of the ordered statistics $X_{(1)}, X_{(2)}, ..., X_{(n)}$. The empirical distribution of the random sample is defined as

$$F_n(x) = \begin{cases} 0 & \text{if} \quad x < x_{(1)}, \\ \frac{k}{n} & \text{if} \quad x_{(k)} \leq x < x_{(k+1)}, \qquad \text{for } k = 1, 2, ..., n-1, \\ 1 & \text{if} \quad x_{(n)} \leq x. \end{cases}$$

The graph of the empirical distribution function $F_4(x)$ is shown below.



Empirical Distribution Function

For a fixed value of x, the empirical distribution function can be considered as a random variable that takes on the values

$$0, \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}, \frac{n}{n}.$$

First we show that $F_n(x)$ is an unbiased estimator of the population distribution $F(x)$. That is,

$$E(F_n(x)) = F(x) \tag{1}$$

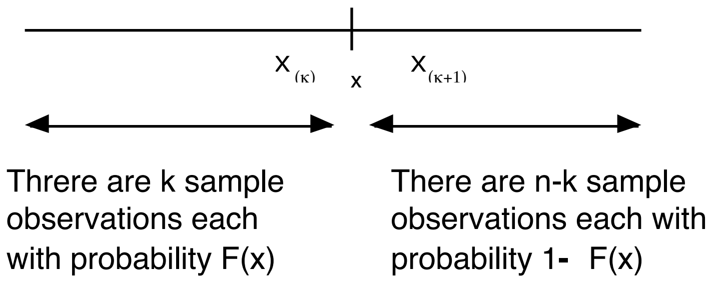
for a fixed value of x. To establish (1), we need the probability density function of the random variable $F_n(x)$. From the definition of the empirical distribution we see that if exactly k observations are less than or equal to x, then

$$F_n(x) = \frac{k}{n}$$

which is

$$n\,F_n(x) = k.$$

The probability that an observation is less than or equal to x is given by $F(x)$.


X (K)
x
X (K+1)
Threre are k sample observations each with probability F(x)
There are n-k sample observations each with probability 1- F(x)

Distribution of the Empirical Distribution Function

Hence (see figure above)

$$\begin{aligned} P(n\,F_n(x) = k) &= P\left(F_n(x) = \frac{k}{n}\right) \\ &= \binom{n}{k} \left[F(x)\right]^k \left[1 - F(x)\right]^{n-k} \end{aligned}$$

for $k = 0, 1, ..., n$. Thus

$$n\,F_n(x) \sim BIN(n, F(x)).$$

Thus the expected value of the random variable $n\,F_n(x)$ is given by

$$\begin{aligned} E(n\,F_n(x)) &= n\,F(x) \\ n\,E(F_n(x)) &= n\,F(x) \\ E(F_n(x)) &= F(x). \end{aligned}$$

This shows that, for a fixed x, $F_n(x)$, on an average, equals to the population distribution function $F(x)$. Hence the empirical distribution function $F_n(x)$ is an unbiased estimator of $F(x)$.

Since $n\,F_n(x) \sim BIN(n, F(x))$, the variance of $n\,F_n(x)$ is given by

$$Var(n\,F_n(x)) = n\,F(x)\,[1 - F(x)].$$

Hence the variance of $F_n(x)$ is

$$Var(F_n(x)) = \frac{F(x)\,[1 - F(x)]}{n}.$$

It is easy to see that $Var(F_n(x)) \to 0$ as $n \to \infty$ for all values of x. Thus the empirical distribution function $F_n(x)$ and $F(x)$ tend to be closer to each other with large n. As a matter of fact, Glivenkno, a Russian mathematician, proved that $F_n(x)$ converges to $F(x)$ uniformly in x as $n \to \infty$ with probability one.

Because of the convergence of the empirical distribution function to the theoretical distribution function, it makes sense to construct a goodness of fit test based on the closeness of $F_n(x)$ and hypothesized distribution $F(x)$.

Let

$$D_n = \max_{x \in \mathbb{R}} |F_n(x) - F(x)|.$$

That is D_n is the maximum of all pointwise differences $|F_n(x) - F(x)|$. The distribution of the Kolmogorov-Smirnov statistic, D_n can be derived. However, we shall not do that here as the derivation is quite involved. In stead, we give a closed form formula for $P(D_n \leq d)$. If $X_1, X_2, ..., X_n$ is a sample from a population with continuous distribution function $F(x)$, then

$$P(D_n \leq d) = \begin{cases} 0 & \text{if } d \leq \frac{1}{2\,n} \\ n! \displaystyle\prod_{i=1}^{n} \int_{2\,i-d}^{2\,i-\frac{1}{n}+d} du & \text{if } \frac{1}{2n} < d < 1 \\ 1 & \text{if } d \geq 1 \end{cases}$$

where $du = du_1 du_2 \cdots du_n$ with $0 < u_1 < u_2 < \cdots < u_n < 1$. Further,

$$\lim_{n\to\infty} P(\sqrt{n}\, D_n \leq d) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} e^{-2\,k^2\,d^2}.$$

These formulas show that the distribution of the Kolmogorov-Smirnov statistic D_n is distribution free, that is, it does not depend on the distribution F of the population.

For most situations, it is sufficient to use the following approximations due to Kolmogorov:

$$P(\sqrt{n}\, D_n \leq d) \approx 1 - 2e^{-2nd^2} \qquad \text{for } d > \frac{1}{\sqrt{n}}.$$

If the null hypothesis $H_o : X \sim F(x)$ is true, the statistic D_n is small. It is therefore reasonable to reject H_o if and only if the observed value of D_n is larger than some constant d_n. If the level of significance is given to be α, then the constant d_n can be found from

$$\alpha = P(D_n > d_n \,/\, H_o \text{ is true}) \approx 2e^{-2nd_n^2}.$$

This yields the following hypothesis test: Reject H_o if $D_n \geq d_n$ where

$$d_n = \sqrt{-\frac{1}{2n} \ln\left(\frac{\alpha}{2}\right)}$$

is obtained from the above Kolmogorov's approximation. Note that the approximate value of d_{12} obtained by the above formula is equal to 0.3533 when $\alpha = 0.1$, however more accurate value of d_{12} is 0.34.

Next we address the issue of the computation of the statistics D_n. Let us define

$$D_n^+ = \max_{x\in\mathbb{R}} \{F_n(x) - F(x)\}$$

and

$$D_n^- = \max_{x\in\mathbb{R}} \{F(x) - F_n(x)\}.$$

Then it is easy to see that

$$D_n = \max\{D_n^+,\, D_N^-\}.$$

Further, since $F_n(x_{(i)}) = \frac{i}{n}$. it can be shown that

$$D_n^+ = \max\left\{ \max_{1\leq i\leq n} \left[\frac{i}{n} - F(x_{(i)})\right],\, 0 \right\}$$

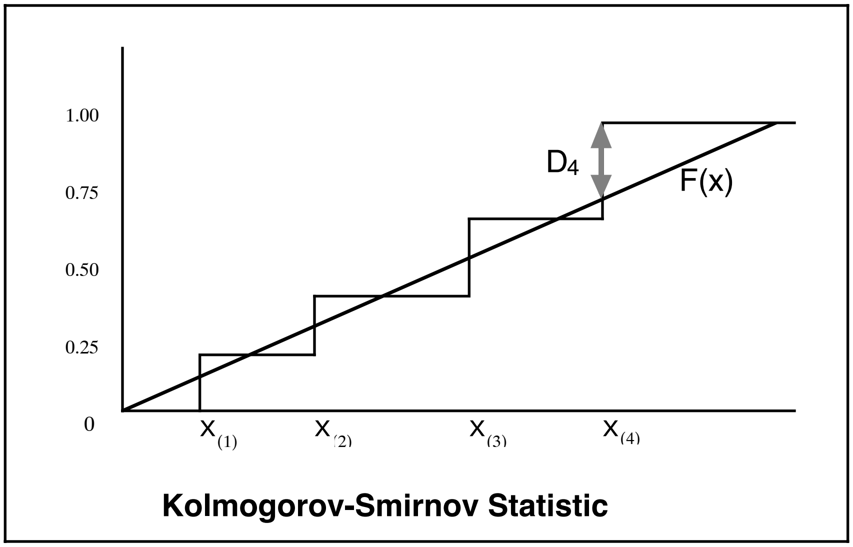
and

$$D_n^- = \max\left\{ \max_{1\leq i\leq n} \left[F(x_{(i)}) - \frac{i-1}{n} \right], 0 \right\}.$$

Therefore it can also be shown that

$$D_n = \max_{1\leq i\leq n} \left\{ \max\left[\frac{i}{n} - F(x_{(i)}),\, F(x_{(i)}) - \frac{i-1}{n} \right] \right\}.$$

The following figure illustrates the Kolmogorov-Smirnov statistics D_n when $n = 4$.



Kolmogorov-Smirnov Statistic

Example 21.1. The data on the heights of 12 infants are given below: 18.2, 21.4, 22.6, 17.4, 17.6, 16.7, 17.1, 21.4, 20.1, 17.9, 16.8, 23.1. Test the hypothesis that the data came from some normal population at a significance level $\alpha = 0.1$.

Answer: Here, the null hypothesis is

$$H_o : X \sim N(\mu, \sigma^2).$$

First we estimate μ and σ^2 from the data. Thus, we get

$$\overline{x} = \frac{230.3}{12} = 19.2.$$

and

$$s^2 = \frac{4482.01 - \frac{1}{12}\,(230.3)^2}{12-1} = \frac{62.17}{11} = 5.65.$$

Hence $s = 2.38$. Then by the null hypothesis

$$F(x_{(i)}) = P\left(Z \leq \frac{x_{(i)} - 19.2}{2.38} \right)$$

where $Z \sim N(0,1)$ and $i = 1, 2, ..., n$. Next we compute the Kolmogorov-Smirnov statistic D_n the given sample of size 12 using the following tabular form.

i	$x_{(i)}$	$F(x_{(i)})$	$\frac{i}{12} - F(x_{(i)})$	$F(x_{(i)}) - \frac{i-1}{12}$
1	16.7	0.1469	-0.0636	0.1469
2	16.8	0.1562	0.0105	0.0729
3	17.1	0.1894	0.0606	0.0227
4	17.4	0.2236	0.1097	-0.0264
5	17.6	0.2514	0.1653	-0.0819
6	17.9	0.2912	0.2088	-0.1255
7	18.2	0.3372	0.2461	-0.1628
8	20.1	0.6480	0.0187	0.0647
9	21.4	0.8212	0.0121	0.0712
10	21.4			
11	22.6	0.9236	-0.0069	0.0903
12	23.1	0.9495	0.0505	0.0328

Thus

$$D_{12} = 0.2461.$$

From the tabulated value, we see that $d_{12} = 0.34$ for significance level $\alpha = 0.1$. Since D_{12} is smaller than d_{12} we accept the null hypothesis $H_o : X \sim N(\mu, \sigma^2)$. Hence the data came from a normal population.

Example 21.2. Let $X_1, X_2, ..., X_{10}$ be a random sample from a distribution whose probability density function is

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Based on the observed values 0.62, 0.36, 0.23, 0.76, 0.65, 0.09, 0.55, 0.26, 0.38, 0.24, test the hypothesis $H_o : X \sim UNIF(0,1)$ against $H_a : X \not\sim UNIF(0,1)$ at a significance level $\alpha = 0.1$.

Answer: The null hypothesis is $H_o : X \sim UNIF(0,1)$. Thus

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1. \end{cases}$$

Hence

$$F(x_{(i)}) = x_{(i)} \qquad \text{for } i = 1, 2, ..., n.$$

Next we compute the Kolmogorov-Smirnov statistic D_n the given sample of size 10 using the following tabular form.

i	$x_{(i)}$	$F(x_{(i)})$	$\frac{i}{10} - F(x_{(i)})$	$F(x_{(i)}) - \frac{i-1}{10}$
1	0.09	0.09	0.01	0.09
2	0.23	0.23	-0.03	0.13
3	0.24	0.24	0.06	0.04
4	0.26	0.26	0.14	-0.04
5	0.36	0.36	0.14	-0.04
6	0.38	0.38	0.22	-0.12
7	0.55	0.55	0.15	-0.05
8	0.62	0.62	0.18	-0.08
9	0.65	0.65	0.25	-0.15
10	0.76	0.76	0.24	-0.14

Thus

$$D_{10} = 0.25.$$

From the tabulated value, we see that $d_{10} = 0.37$ for significance level $\alpha = 0.1$. Since D_{10} is smaller than d_{10} we accept the null hypothesis

$$H_o : X \sim UNIF(0,1).$$

21.2 Chi-square Test

The chi-square goodness of fit test was introduced by Karl Pearson in 1900. Recall that the Kolmogorov-Smirnov test is only for testing a specific continuous distribution. Thus if we wish to test the null hypothesis

$$H_o : X \sim BIN(n,p)$$

against the alternative $H_a : X \not\sim BIN(n,p)$, then we can not use the Kolmogorov-Smirnov test. Pearson chi-square goodness of fit test can be used for testing of null hypothesis involving discrete as well as continuous distribution. Unlike Kolmogorov-Smirnov test, the Pearson chi-square test uses the density function the population X.

Let $X_1, X_2, ..., X_n$ be a random sample from a population X with probability density function $f(x)$. We wish to test the null hypothesis

$$H_o : X \sim f(x)$$

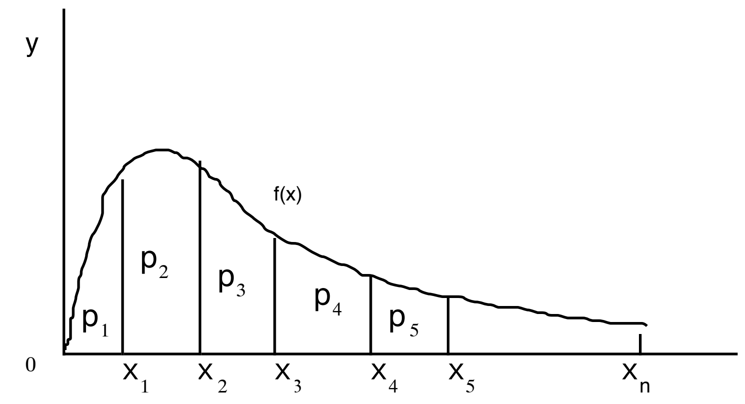
against

$$H_a : X \not\sim f(x).$$

If the probability density function $f(x)$ is continuous, then we divide up the abscissa of the probability density function $f(x)$ and calculate the probability p_i for each of the interval by using

$$p_i = \int_{x_{i-1}}^{x_i} f(x)\, dx,$$

where $\{x_0, x_1, ..., x_n\}$ is a partition of the domain of the $f(x)$.



Discretization of continuous density function

Let $Y_1, Y_2, ..., Y_m$ denote the number of observations (from the random sample $X_1, X_2, ..., X_n$) is $1^{\text{st}}, 2^{\text{nd}}, 3^{\text{rd}}, ..., m^{\text{th}}$ interval, respectively.

Since the sample size is n, the number of observations expected to fall in the i^{th} interval is equal to np_i. Then

$$Q = \sum_{i=1}^{m} \frac{(Y_i - np_i)^2}{np_i}$$

measures the closeness of observed Y_i to expected number np_i. The distribution of Q is chi-square with $m-1$ degrees of freedom. The derivation of this fact is quite involved and beyond the scope of this introductory level book.

Although the distribution of Q for $m > 2$ is hard to derive, yet for $m = 2$ it not very difficult. Thus we give a derivation to convince the reader that Q has χ^2 distribution. Notice that $Y_1 \sim BIN(n, p_1)$. Hence for large n by the central limit theorem, we have

$$\frac{Y_1 - n\,p_1}{\sqrt{n\,p_1\,(1-p_1)}} \sim N(0,1).$$

Thus

$$\frac{(Y_1 - n\,p_1)^2}{n\,p_1\,(1-p_1)} \sim \chi^2(1).$$

Since

$$\frac{(Y_1 - n\,p_1)^2}{n\,p_1\,(1-p_1)} = \frac{(Y_1 - n\,p_1)^2}{n\,p_1} + \frac{(Y_1 - n\,p_1)^2}{n\,(1-p_1)},$$

we have This implies that

$$\frac{(Y_1 - n\,p_1)^2}{n\,p_1} + \frac{(Y_1 - n\,p_1)^2}{n\,(1-p_1)} \sim \chi^2(1)$$

which is

$$\frac{(Y_1 - n\,p_1)^2}{n\,p_1} + \frac{(n - Y_2 - n + n\,p_2)^2}{n\,p_2} \sim \chi^2(1)$$

due to the facts that $Y_1 + Y_2 = n$ and $p_1 + p_2 = 1$. Hence

$$\sum_{i=1}^{2} \frac{(Y_i - n\,p_i)^2}{n\,p_i} \sim \chi^2(1),$$

that is, the chi-square statistic Q has approximate chi-square distribution.

Now the simple null hypothesis

$$H_0 : p_1 = p_{10}, \;\; p_2 = p_{20}, \;\; \cdots \;\; p_m = p_{m0}$$

is to be tested against the composite alternative

$$H_a : \text{at least one } p_i \text{ is not equal to } p_{i0} \text{ for some } i.$$

Here $p_{10}, p_{20}, ..., p_{m0}$ are fixed probability values. If the null hypothesis is true, then the statistic

$$Q = \sum_{i=1}^{m} \frac{(Y_i - n\,p_{i0})^2}{n\,p_{i0}}$$

has an approximate chi-square distribution with $m-1$ degrees of freedom. If the significance level α of the hypothesis test is given, then

$$\alpha = P\left(Q \geq \chi^2_{1-\alpha}(m-1)\right)$$

and the test is "Reject H_o if $Q \geq \chi^2_{1-\alpha}(m-1)$." Here $\chi^2_{1-\alpha}(m-1)$ denotes a real number such that the integral of the chi-square density function with $m-1$ degrees of freedom from zero to this real number $\chi^2_{1-\alpha}(m-1)$ is $1-\alpha$. Now we give several examples to illustrate the chi-square goodness-of-fit test.

Example 21.3. A die was rolled 30 times with the results shown below:

Number of spots	1	2	3	4	5	6
Frequency (x_i)	1	4	9	9	2	5

If a chi-square goodness of fit test is used to test the hypothesis that the die is fair at a significance level $\alpha = 0.05$, then what is the value of the chi-square statistic and decision reached?

Answer: In this problem, the null hypothesis is

$$H_o : p_1 = p_2 = \cdots = p_6 = \frac{1}{6}.$$

The alternative hypothesis is that not all p_i's are equal to $\frac{1}{6}$. The test will be based on 30 trials, so $n = 30$. The test statistic

$$Q = \sum_{i=1}^{6} \frac{(x_i - n\,p_i)^2}{n\,p_i},$$

where $p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$. Thus

$$n\,p_i = (30)\,\frac{1}{6} = 5$$

and

$$\begin{aligned} Q &= \sum_{i=1}^{6} \frac{(x_i - n\,p_i)^2}{n\,p_i} \\ &= \sum_{i=1}^{6} \frac{(x_i - 5)^2}{5} \\ &= \frac{1}{5}\,[16 + 1 + 16 + 16 + 9] \\ &= \frac{58}{5} = 11.6. \end{aligned}$$

The tabulated χ^2 value for $\chi^2_{0.95}(5)$ is given by

$$\chi^2_{0.95}(5) = 11.07.$$

Since

$$11.6 = Q > \chi^2_{0.95}(5) = 11.07$$

the null hypothesis $H_o : p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$ should be rejected.

Example 21.4. It is hypothesized that an experiment results in outcomes K, L, M and N with probabilities $\frac{1}{5}$, $\frac{3}{10}$, $\frac{1}{10}$ and $\frac{2}{5}$, respectively. Forty independent repetitions of the experiment have results as follows:

Outcome	K	L	M	N
Frequency	11	14	5	10

If a chi-square goodness of fit test is used to test the above hypothesis at the significance level $\alpha = 0.01$, then what is the value of the chi-square statistic and the decision reached?

Answer: Here the null hypothesis to be tested is

$$H_o : p(K) = \frac{1}{5},\ p(L) = \frac{3}{10},\ p(M) = \frac{1}{10},\ p(N) = \frac{2}{5}.$$

The test will be based on $n = 40$ trials. The test statistic

$$\begin{aligned} Q &= \sum_{k=1}^{4} \frac{(x_k - np_k)^2}{n\,p_k} \\ &= \frac{(x_1-8)^2}{8} + \frac{(x_2-12)^2}{12} + \frac{(x_3-4)^2}{4} + \frac{(x_4-16)^2}{16} \\ &= \frac{(11-8)^2}{8} + \frac{(14-12)^2}{12} + \frac{(5-4)^2}{4} + \frac{(10-16)^2}{16} \\ &= \frac{9}{8} + \frac{4}{12} + \frac{1}{4} + \frac{36}{16} \\ &= \frac{95}{24} = 3.958. \end{aligned}$$

From chi-square table, we have

$$\chi^2_{0.99}(3) = 11.35.$$

Thus

$$3.958 = Q < \chi^2_{0.99}(3) = 11.35.$$

Therefore we accept the null hypothesis.

Example 21.5. Test at the 10% significance level the hypothesis that the following data

06.88	06.92	04.80	09.85	07.05	19.06	06.54	03.67	02.94	04.89
69.82	06.97	04.34	13.45	05.74	10.07	16.91	07.47	05.04	07.97
15.74	00.32	04.14	05.19	18.69	02.45	23.69	44.10	01.70	02.14
05.79	03.02	09.87	02.44	18.99	18.90	05.42	01.54	01.55	20.99
07.99	05.38	02.36	09.66	00.97	04.82	10.43	15.06	00.49	02.81

give the values of a random sample of size 50 from an exponential distribution with probability density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}\, e^{-\frac{x}{\theta}} & \text{if } 0 < x < \infty \\ \\ 0 & \text{elsewhere,} \end{cases}$$

where $\theta > 0$.

Answer: From the data $\overline{x} = 9.74$ and $s = 11.71$. Notice that

$$H_o : X \sim EXP(\theta).$$

Hence we have to partition the domain of the experimental distribution into m parts. There is no rule to determine what should be the value of m. We assume $m = 10$ (an arbitrary choice for the sake of convenience). We partition the domain of the given probability density function into 10 mutually disjoint sets of equal probability. This partition can be found as follow.

Note that $\overline{x}$ estimate θ. Thus

$$\widehat{\theta} = \overline{x} = 9.74.$$

Now we compute the points $x_1, x_2, ..., x_{10}$ which will be used to partition the domain of $f(x)$

$$\begin{aligned} \frac{1}{10} &= \int_{x_o}^{x_1} \frac{1}{\theta}\, e^{-\frac{x}{\theta}} \\ &= -\left[e^{-\frac{x}{\theta}}\right]_0^{x_1} \\ &= 1 - e^{-\frac{x_1}{\theta}}. \end{aligned}$$

Hence

$$\begin{aligned} x_1 &= \theta\, \ln\left(\frac{10}{9}\right) \\ &= 9.74\, \ln\left(\frac{10}{9}\right) \\ &= 1.026. \end{aligned}$$

Using the value of x_1, we can find the value of x_2. That is

$$\frac{1}{10} = \int_{x_1}^{x_2} \frac{1}{\theta}\, e^{-\frac{x}{\theta}}$$

$$= e^{-\frac{x_1}{\theta}} - e^{-\frac{x_2}{\theta}}.$$

Hence

$$x_2 = -\theta \, \ln \left(e^{-\frac{x_1}{\theta}} - \frac{1}{10} \right).$$

In general

$$x_k = -\theta \, \ln \left(e^{-\frac{x_{k-1}}{\theta}} - \frac{1}{10} \right)$$

for $k = 1, 2, ..., 9$, and $x_{10} = \infty$. Using these x_k's we find the intervals $A_k = [x_k,\, x_{k+1})$ which are tabulates in the table below along with the number of data points in each each interval.

Interval A_i	Frequency (o_i)	Expected value (e_i)
$[0,\ 1.026)$	3	5
$[1.026,\ 2.173)$	4	5
$[2.173,\ 3.474)$	6	5
$[3.474,\ 4.975)$	6	5
$[4.975,\ 6.751)$	7	5
$[6.751,\ 8.925)$	7	5
$[8.925,\ 11.727)$	5	5
$[11.727,\ 15.676)$	2	5
$[15.676,\ 22.437)$	7	5
$[22.437,\ \infty)$	3	5
Total	50	50

From this table, we compute the statistics

$$Q = \sum_{i=1}^{10} \frac{(o_i - e_i)^2}{e_i} = 6.4.$$

and from the chi-square table, we obtain

$$\chi^2_{0.9}(9) = 14.68.$$

Since

$$6.4 = Q < \chi^2_{0.9}(9) = 14.68$$

we accept the null hypothesis that the sample was taken from a population with exponential distribution.

21.3. Review Exercises

1. The data on the heights of 4 infants are: $18.2, 21.4, 16.7$ and 23.1. For a significance level $\alpha = 0.1$, use Kolmogorov-Smirnov Test to test the hypothesis that the data came from some uniform population on the interval $(15, 25)$. (Use $d_4 = 0.56$ at $\alpha = 0.1$.)

2. A four-sided die was rolled 40 times with the following results

Number of spots	1	2	3	4
Frequency	5	9	10	16

If a chi-square goodness of fit test is used to test the hypothesis that the die is fair at a significance level $\alpha = 0.05$, then what is the value of the chi-square statistic?

3. A coin is tossed 500 times and k heads are observed. If the chi-squares distribution is used to test the hypothesis that the coin is unbiased, this hypothesis will be accepted at 5 percents level of significance if and only if k lies between what values? (Use $\chi^2_{0.05}(1) = 3.84$.)

4. It is hypothesized that an experiment results in outcomes A, C, T and G with probabilities $\frac{1}{16}$, $\frac{5}{16}$, $\frac{1}{8}$ and $\frac{3}{8}$, respectively. Eighty independent repetitions of the experiment have results as follows:

Outcome	A	G	C	T
Frequency	3	28	15	34

If a chi-square goodness of fit test is used to test the above hypothesis at the significance level $\alpha = 0.1$, then what is the value of the chi-square statistic and the decision reached?

5. A die was rolled 50 times with the results shown below:

Number of spots	1	2	3	4	5	6
Frequency (x_i)	8	7	12	13	4	6

If a chi-square goodness of fit test is used to test the hypothesis that the die is fair at a significance level $\alpha = 0.1$, then what is the value of the chi-square statistic and decision reached?

6. Test at the 10% significance level the hypothesis that the following data

05.88	05.92	03.80	08.85	06.05	18.06	05.54	02.67	01.94	03.89
70.82	07.97	05.34	14.45	06.74	11.07	17.91	08.47	06.04	08.97
16.74	01.32	03.14	06.19	19.69	03.45	24.69	45.10	02.70	03.14
04.79	02.02	08.87	03.44	17.99	17.90	04.42	01.54	01.55	19.99
06.99	05.38	03.36	08.66	01.97	03.82	11.43	14.06	01.49	01.81

give the values of a random sample of size 50 from an exponential distribution with probability density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}\, e^{-\frac{x}{\theta}} & \text{if } 0 < x < \infty \\ \\ 0 & \text{elsewhere,} \end{cases}$$

where $\theta > 0$.

7. Test at the 10% significance level the hypothesis that the following data

0.88	0.92	0.80	0.85	0.05	0.06	0.54	0.67	0.94	0.89
0.82	0.97	0.34	0.45	0.74	0.07	0.91	0.47	0.04	0.97
0.74	0.32	0.14	0.19	0.69	0.45	0.69	0.10	0.70	0.14
0.79	0.02	0.87	0.44	0.99	0.90	0.42	0.54	0.55	0.99
0.94	0.38	0.36	0.66	0.97	0.82	0.43	0.06	0.49	0.81

give the values of a random sample of size 50 from an exponential distribution with probability density function

$$f(x;\theta) = \begin{cases} (1+\theta)\, x^{\theta} & \text{if } 0 \leq x \leq 1 \\ \\ 0 & \text{elsewhere,} \end{cases}$$

where $\theta > 0$.

8. Test at the 10% significance level the hypothesis that the following data

06.88	06.92	04.80	09.85	07.05	19.06	06.54	03.67	02.94	04.89
29.82	06.97	04.34	13.45	05.74	10.07	16.91	07.47	05.04	07.97
15.74	00.32	04.14	05.19	18.69	02.45	23.69	24.10	01.70	02.14
05.79	03.02	09.87	02.44	18.99	18.90	05.42	01.54	01.55	20.99
07.99	05.38	02.36	09.66	00.97	04.82	10.43	15.06	00.49	02.81

give the values of a random sample of size 50 from an exponential distribution with probability density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \leq x \leq \theta \\ \\ 0 & \text{elsewhere.} \end{cases}$$

9. Suppose that in 60 rolls of a die the outcomes 1, 2, 3, 4, 5, and 6 occur with frequencies n_1, n_2, 14, 8, 10, and 8 respectively. What is the least value of $\sum_{i=1}^{2}(n_i - 10)^2$ for which the chi-square test rejects the hypothesis that the die is fair at 1% level of significance level? (Answer: $\sum_{i=1}^{2}(n_i - 10)^2 \geq 63.43$.)

10. It is hypothesized that of all marathon runners 70% are adult men, 25% are adult women, and 5% are youths. To test this hypothesis, the following data from the a recent marathon are used:

Adult Men	Adult Women	Youths	Total
630	300	70	1000

A chi-square goodness-of-fit test is used. What is the value of the statistics? (Ans: 25)