BHARATHIDASAN UNIVERSITY

Tiruchirappalli- 620024

Tamil Nadu, India.

Programme: M.Sc. Statistics

Course Title: R Programming

Course Code: 23ST05CC

Unit-IV

Discrete Probability Distribution

Dr. T. Jai Sankar

Associate Professor and Head

Department of Statistics

Ms. I. Angel Agnes Mary

Guest Faculty

Department of Statistics

UNIT – IV

Discrete Probability Distribution

Distribution	MGF	Mean	Variance	PMF
Bernoulli	q + pe^{t}	P	Pq	$P(x) = p^x q^{1-x}$, $x = 0,1$
Binomial	(q + pe^{t})^{n}	Np	Npq	$P(x) = \binom{n}{x} p^x q^{n-x}$, $x = 0,1,2,....$
Poisson	Exp[λ(e^{t} – 1)	λ	λ	$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, $x = 0,1,2,....$
Negative Binomial	$\left(\frac{1-p}{1-pe^t}\right)^r$	$\frac{pr}{(1-p)}$	$\frac{pr}{(1-p)^2}$	$f(x) = \frac{(x-1)+r}{(x-1)+1} \times q^x \times f(x-1)$, $x = 0,1,2,....$
Geometric	$\left(\frac{p}{1-(1-p)e^t}\right)$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$	$P(x) = q^x \times p$, $x = 0,1,2,....$
Hyper geometric	$\frac{\binom{N-k}{n} 2F_1(-n,-k;N-k-n+1;e^t)}{\binom{N}{n}}$	$\frac{nk}{N}$	$n\left(\frac{N-n}{N-1}\right)\left(\frac{k}{N}\right)\left(\frac{N-k}{N}\right)$	$P(x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$, $x = 0,1,2,....$

10. Fitting of Bernoulli distribution

Fitting of Bernoulli distribution for the following data and test for the goodness.

No. of cells	0	1
Frequency	56	44

Aim

Fit an appropriate Bernoulli distribution to the above data and calculate theoretical frequency and testing the goodness of fit of a Bernoulli distribution.

Procedure

- Define the Null Hypothesis (H_0) and Alternate Hypothesis (H_1) for given data.
- For a given Bernoulli Distribution Calculate respecting Mean and then compute successive event (p).
- Compute the respective probability using the recurrence relation of Bernoulli Distribution.

  $$P(x) = p^x q^{1-x}, x = 0,1$$

- Compute the Expected Frequency [E_i].
- Compute the test Statistic of Chi-Square Test,

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

- Taking decision about the hypothesis H_0 with respect to the level of significance and degree of freedom.

Calculation

Hypothesis

H_0: The data is good fit for the Bernoulli distribution.

H_1: The data is not good fit for the Bernoulli Binomial distribution.

Calculate mean for appropriate formula of Discrete Distribution.

x	Y=f	fx
0	56	0
1	44	44
Total	100	44

$$\mu = \frac{\sum_{i=1}^{n} fx}{\sum_{i=1}^{n} f}$$

$$= \frac{44}{100}$$

$$= 0.44$$

To find Successive Event (p),

$$Mean = np$$

$$p = \frac{Mean}{n}$$

$$p = \frac{0.44}{1}$$

$$= 0.44$$

Then, q = 1 – p,

q = 1 – 0.44

= 0.56

For a Bernoulli distribution the probability mass function is given by,

$$P(x) = p^x q^{1-x}, x = 0,1$$

$$P(0) = (0.44)^0 (0.56)^{1-0}$$

= 0.56

$$P(1) = (0.44)^1 (0.56)^{1-1}$$

= 0.44

Calculate the Expected Frequency,

x	N . p(x)	Expected Frequency (E_i)
0	100 x 0.56	56
1	100 x 0.44	44

Compute the test Statistic of Chi-Square Test,

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

Frequency (O_i)	Expected Frequency (E_i)	O_i - E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
56	56	0	0	0
44	44	0	0	0
			Total	0

Degrees of Freedom

Calculate the Degrees of Freedom = n – p – 1

= 2 – 1 – 1

= 0

Table Value from Chi-Square Table in 0df = 0

Calculated Value = 0

Conclusion

Since Calculated Value is equal the Table Value (i.e. 0 = 0). So, we accept the null Hypothesis. Hence we conclude that the data is good fit for the Bernoulli distribution.

R coding

 # Import Rlab library

library(Rlab)

 # Given data

n<-1

x<-c(0,1)

frequency<-c(56,44)

 # Calculating Observed probability of Success

```
mean<- sum(x*frequency)/sum(frequency)

mean

p <- mean
```

 # Calculations for Fitting of Distribution by Chi-square Method

```
prob = round(dbern(x, p),6)

prob

exp_freq<-prob*(sum(frequency))

oi_ei_sq<-(frequency-exp_freq)^2

oi_ei<-round(oi_ei_sq/exp_freq,4)
```

 # Output Table, Chi-square test statistics & P-Value

```
out_table<- cbind(x,prob, exp_freq,oi_ei)

out_table

chi_square<- sum(oi_ei)

chi_square

pvalue<- pchisq(chi_square,n+1-1-1,lower.tail=FALSE)

pvalue
```

<u>R Output</u>

mean_x

[1] 0.44

p

[1] 0.44

Output

x	prob_x	exp_freq	oi_ei
0	0.56	56	0
1	0.44	44	0

> chi_square

[1] 0

> pvalue

[1] 1

Interpretation

Here the p-value is greater than the specified level of significance (i.e.) **1 > 0.05**, so we accept the null hypothesis. Hence, we conclude that the data is good fit for the Bernoulli distribution.

Observation Problem

The following data are the number of seeds germinating out of 2 a damp fitter for 50 seeds fit a Bernoulli distribution to these data.

x	0	1
Frequency	15	35

Also find goodness of fit for the given data.

11. Fitting of Binomial Distribution

A lorry carrying a large number of Eggs. Which involved in an accident each box contain six eggs. After accident occur a random sample of 100 boxes are examine and the number of eggs broken are recorded. The number of boxes containing various number of the broken eggs is given in the following data.

X	0	1	2	3	4	5	6
Y	31	37	22	7	2	1	0

Test whether Binomial distribution is a good fit for the given data

Aim

Fit an appropriate Binomial Distribution to the above data and calculate theoretical frequency and testing the goodness of fit of a Binomial Distribution.

Procedure

- Define the Null Hypothesis (H_0) and Alternate Hypothesis (H_1) for given data.
- For a given Binomial Distribution Calculate respecting Mean and then compute successive event (p).
- Compute the respective probability using the recurrence relation of Binomial Distribution.

$$P(x) = \binom{n}{x} p^x q^{n-x}, \qquad x = 0, 1, 2, \ldots.$$

- Compute the Expected Frequency [E_i].
- Compute the test Statistic of Chi-Square Test,

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

- Taking decision about the hypothesis H_0 with respect to the level of significance and degree of freedom.

Calculation

Hypothesis

H_0: The data is good fit for the Binomial Distribution.

H_1: The data is not good fit for the Binomial Distribution.

Calculate mean for appropriate formula of Discrete Distribution.

X	Y=f	fX
0	31	0
1	37	37
2	22	44
3	7	21
4	2	8
5	1	5
6	0	0
Total	100	115

$$\mu = \frac{\sum_{i=1}^{n} fx}{\sum_{i=1}^{n} f}$$

$$= \frac{115}{100}$$

$$= 1.15$$

To find Successive Event (p),

$$Mean = np$$

$$p = \frac{Mean}{n}$$

$$p = \frac{1.15}{6}$$

$$= 0.1917$$

Then, q = 1 – p,

q = 1 – 0.1917 = 0.8083

For a Binomial Distribution the probability mass function is given by,

$$P(x) = \binom{n}{x} p^x q^{n-x}, \qquad x = 0, 1, 2, \ldots.$$

$$P(0) = \binom{6}{0} (0.1917)^0 (0.8083)^{6-0}$$

= 0.2789

$$P(1) = \binom{6}{1} (0.1917)^1 (0.8083)^{6-1}$$

= 6 x 0.1917 x 0.3450

= 0.3968

$$P(2) = \binom{6}{2}(0.1917)^2(0.8083)^{6-2}$$

= 15 x 0.0367 x 0.4269

= 0.2353

$$P(3) = \binom{6}{3}(0.1917)^3(0.8083)^{6-3}$$

= 20 x 0.0070 x 0.5281

= 0.0740

$$P(4) = \binom{6}{4}(0.1917)^4(0.8083)^{6-4}$$

= 15 x 0.00135 x 0.65335

= 0.01323

$$P(5) = \binom{6}{5}(0.1917)^5(0.8083)^{6-5}$$

= 15 x 0.000259 x 0.8083

= 0.001255

$$P(6) = \binom{6}{6}(0.1917)^6(0.8083)^{6-6}$$

= 1 x 0.0000496 x 1

= 0.000050

Calculate the Expected Frequency,

x	N . p(x)	Expected Frequency (E_i)
0	100 x 0.2789	27.89
1	100 x 0.3968	39.68
2	100 x 0.2353	23.53
3	100 x 0.0744	7.44
4	100 x 0.0132	1.32
5	100 x 0.00125	0.125
6	100 x 0.000050	0.0050

Compute the test Statistic of Chi-Square Test,

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

Frequency (O_i)	Expected Frequency (E_i)	O_i - E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
31	27.89	3.11	9.6721	0.3468
37	39.68	-2.68	7.1824	0.1810
22	23.53	-1.53	2.3409	0.0995
7	7.44	-0.44	0.1936	0.0260
2	1.32	0.67	0.4489	0.3401
1	0.125	0.875	0.7656	6.1248
0	0.0050	-0.0050	0.000025	0.005
			Total	7.1232

Degrees of Freedom

Calculate the Degrees of Freedom = n – p – 1

= 7 – 1 – 1

= 5

Table Value from Chi-Square Table in 5df = 11.070

Calculated Value = 7.1232

Conclusion

Since Calculated Value is less than the Table Value (i.e. 7.1232 < 11.070). So, we accept the null Hypothesis. Hence we conclude that the data is good fit for the Binomial Distribution.

R Coding

 # Given data

```
n <- 6

x <- c(0,1,2,3,4,5,6)

frequency<- c(31,37,22,7,2,1,0)
```

 # Calculating Observed probability of Success

```
mean<- sum(x* frequency)/sum(frequency)

p <- round((mean/n),4)
```

 # Calculations for Fitting of Distribution by Chi-square Method

```
prob = round(dbinom(x, n, p),6)

prob

exp_freq<-prob*(sum(frequency))

oi_ei_sq<-(frequency-exp_freq)^2

oi_ei<-round(oi_ei_sq/exp_freq,4)
```

 # Output Table, Chi-square test statistics & P-Value

```
out_table<- cbind(x,prob, exp_freq,oi_ei)

out_table

chi_square<- sum(oi_ei)

chi_square

pvalue<- pchisq(chi_square,n+1-1-1,lower.tail=FALSE)

pvalue
```

<u>R Output</u>

mean_x

[1] 1.15

p

[1] 0.1917

Output

x	prob_x	exp_freq	oi_ei
0	0.278892	27.8892	0.3470
1	0.396859	39.6859	0.1818
2	0.235302	23.5302	0.0995
3	0.074407	7.4407	0.0261
4	0.013235	1.3235	0.3458
5	0.001256	0.1256	6.0874
6	0.000050	0.0050	0.0050

> chi_square

[1] 7.0926

> pvalue

[1] 0.2138441

Interpretation

Here the p-value is greater than the specified level of significance (i.e.) **0.213** > **0.05**, so we accept the null hypothesis. Hence, we conclude that the data is good fit for the Binomial Distribution.

Observation Problem

Screws produced by a certain machine were checked by examine samples of 12. The following table shows the distribution of 128 samples according to the no. of defectives items the contained.

Defectives	0	1	2	3	4	5	6	7
Samples	7	6	19	35	30	23	7	1

Test whether Binomial distribution is a good fit for the given data.

12. Fitting of Poisson distribution

Fitting of Poisson distribution for the following data and test for the goodness.

Arrivals	0	1	2	3	4	5	6	7
Frequency	16	30	45	39	27	19	8	6

Aim

Fit a Poisson distribution for the above data and calculate theoretical frequency and testing the goodness of fit of a Poisson distribution.

Procedure

- Define the Null Hypothesis (H_0) and Alternate Hypothesis (H_1) for given data.
- For a given Poisson Distribution Calculate respecting Mean.

$$Mean\ (\lambda) = \frac{\sum fx}{\sum f}$$

- Compute the respective probability using the recurrence relation of Poisson Distribution.

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!},\ x = 0,1,2,\ldots.$$

- Compute the Expected Frequency [E_i].
- Compute the test Statistic of Chi-Square Test,

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

- Taking decision about the hypothesis H_0 with respect to the level of significance and degree of freedom.

Calculation

Hypothesis

H_0: The data is good fit for the Poisson distribution.

H_1: The data is not good fit for the Poisson distribution.

Calculate mean (λ)

X	F	FX
0	16	0
1	30	30
2	45	90
3	39	117
4	27	108
5	19	95
6	8	48
7	6	42
Total	190	530

$$\lambda = \frac{\sum_{i=1}^{n} fx}{\sum_{i=1}^{n} f}$$

$$= \frac{530}{190}$$

$$= 2.7895$$

For a Poisson distribution the probability mass function is given by,

$$P(x) = \frac{e^{-\lambda} \lambda^{x}}{x!}, \; x = 0,1,2,....$$

$$P(0) = \frac{e^{(-2.7895)} (2.7895)^{0}}{0}$$

$$= 0.0615$$

$$P(1) = \frac{e^{(-2.7895)} (2.7895)^{1}}{1!}$$

$$= 0.1714$$

$$P(2) = \frac{e^{(-2.7895)} (2.7895)^{2}}{2!}$$

$$= 0.2391$$

$$P(3) = \frac{e^{(-2.7895)}(2.7895)^3}{3!}$$

$$= 0.2233$$

$$P(4) = \frac{e^{(-2.7895)}(2.7895)^4}{4!}$$

$$= 0.1550$$

$$P(5) = \frac{e^{(-2.7895)}(2.7895)^5}{5!}$$

$$= 0.0865$$

$$P(6) = \frac{e^{(-2.7895)}(2.7895)^6}{6!}$$

$$= 0.0402$$

$$P(7) = \frac{e^{(-2.7895)}(2.7895)^7}{7!}$$

$$= 0.0160$$

Calculate the Expected Frequency,

X	F	N . p(x)	Expected Frequency (E_i)
0	16	190 x 0.0615 = 11.68	12
1	30	190 x 0.1714 = 32.57	33
2	45	190 x 0.2391 = 45.43	45
3	39	190 x 0.2233 = 42.23	42
4	27	190 x 0.1550 = 29.45	29
5	19	190 x 0.0865 = 16.43	16
6	8	190 x 0.0402 = 7.63	8
7	6	190 x 0.0160 = 3.04	3

Compute the test Statistic of Chi-Square Test,

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

Frequency (O_i)	Expected Frequency (E_i)	O_i - E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
16	12	4	16	1.3333
30	33	-3	9	0.2727
45	45	0	0	0
39	42	-3	9	0.2143
27	29	-2	4	0.1379
19	16	3	9	0.5625
8	8	0	0	0
6	3	3	9	3
			Total	5.5207

Degrees of Freedom

Calculate the Degrees of Freedom = n – p – 1

= 8 – 1 – 1

= 6

Table Value from Chi-Square Table in 6df = 12.592

Calculated Value = 5.5207

Conclusion

Since Calculated Value is less than the Table Value (i.e. 5.5207 < 12.592). So, we accept the null Hypothesis. Hence we conclude that the data is good fit for the Poisson distribution.

<u>R Coding</u>

 # Given data

```
x <- c(0,1,2,3,4,5,6,7)

obs_freq<- c(16,30,45,39,27,19,8,6)
```

 # Calculating Observed probability of Success

```
mean_x<- sum(x*obs_freq)/sum(obs_freq)

lambda<- mean_x
```

 # creating required storage Space

```
prob_x<- seq(1,8,1)

exp_x<- seq(1,8,1)

oi_ei<- seq(1,8,1)
```

 # Calculations for Fitting of Distribution by Chi-square Method

```
for (i in 1:8)
 {
 prob_x[i] <- round(exp(-(lambda))*(lambda^x[i])/factorial(x[i]),4)
 exp_x[i] <- round((sum(obs_freq)*prob_x[i]),0)
oi_ei[i] <- round(((obs_freq[i]-exp_x[i])^2)/exp_x[i],4)
 }
```

 # Output Table, Chi-square test statistics & P-Value

```
output<- cbind(x,prob_x,exp_x,oi_ei)
output
chi_square<- sum(oi_ei)
pvalue<- pchisq(chi_square,8-1-1,lower.tail=FALSE)
pvalue
```

<u>R Output</u>

mean_x

[1] 2.789474

Output

x	prob_x	exp_freq	oi_ei
0	0.0615	12	1.3333
1	0.1714	33	0.2727
2	0.2391	45	0.0000
3	0.2223	42	0.2143
4	0.1550	29	0.1379
5	0.0865	16	0.5625
6	0.0402	8	0.0000
7	0.0160	3	3.0000

> chi_square

[1] 5.5207

> pvalue

[1] 0.4789584

Interpretation

Here the p-value is greater than the specified level of significance (i.e.) **0.479 > 0.05**, so we accept the null hypothesis. Hence, we conclude that the data is good fit for the Poisson distribution.

Observation Problem

The actual arrival per minute was observed in 200 one-minute periods over the course of a week. The results are summarized below:

Arrivals	0	1	2	3	4	5	6	7
Frequency	25	23	45	28	32	22	15	10

Test whether Poisson distribution is a good fit for the given data

13. Fitting of Negative Binomial Distribution

Fitting of Negative Binomial distribution for the following data and test for the goodness.

No. of cells	0	1	2	3	4	5
Frequency	213	128	37	18	3	1

Aim

Fit an appropriate Negative Binomial distribution for the above data and calculate theoretical frequency and testing the goodness of fit of a Negative Binomial distribution.

Procedure

- Define the Null Hypothesis (H_0) and Alternate Hypothesis (H_1) for given data.
- For a given Negative Binomial Distribution Calculate Mean and variance. And also calculate Successes and Failure events p, q and r.

$$Mean = \frac{\sum fx}{\sum f}$$

$$Variance = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2$$

- Compute the respective probability using the recurrence relation of Negative Binomial Distribution.

$$f(x+1;r,p) = \frac{x+r}{x+1} q . f(x;r,p), \qquad x = 0,1,2,....$$

- Compute the Expected Frequency [E_i].
- Compute the test Statistic of Chi-Square Test,

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

- Taking decision about the hypothesis H_0 with respect to the level of significance and degree of freedom.

Calculation

Hypothesis

H_0: The data is good fit for the Negative Binomial Distribution.

H_1: The data is not good fit for the Negative Binomial Distribution.

Calculate Mean and Variance for the data.

x	f	fx	x^2	fx^2
0	213	0	0	0
1	128	128	1	128
2	37	74	4	148
3	18	54	9	162
4	3	12	16	48
5	1	5	25	25
Total	400	273		511

$$\mu = \frac{\sum_{i=1}^{n} fx}{\sum_{i=1}^{n} f} = \frac{273}{400} = 0.6825$$

$$\sigma^2 = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2$$

$$= \frac{511}{400} - \left(\frac{273}{400}\right)^2$$

$$= 1.2775 - (0.6825)^2$$

$$= 1.2775 - 0.4658$$

$$= 0.8117$$

$$p = \frac{Mean}{Variance}$$

$$= \frac{0.6825}{0.8117}$$

$$= 0.8408$$

$$q = 1 - p$$

$$= 1 - 0.8408$$

$$= 0.1592$$

$$r = \frac{p \times Mean}{q}$$

$$= \frac{0.8408 \times 0.6825}{0.1592}$$

$$= \frac{0.5738}{0.1592}$$

$$= 3.6046 \simeq 4$$

For a Negative Binomial Distribution the probability mass function is given by,

$$f(x) = \frac{(x-1)+r}{(x-1)+1} \times q^x \times f(x-1), \qquad x = 0,1,2,\ldots.$$

$$f(0) = p^r = (0.8408)^4$$

= 0.4998

$$f(1) = \frac{0+4}{0+1}(0.1592) \times (0.4998)$$

= 4 x (0.1592) x (0.4998)

= 0.3183

$$f(2) = \frac{1+4}{1+1}(0.1592) \times (0.3183)$$

= 2.5 x (0.1592) x (0.3183)

= 0.1267

$$f(3) = \frac{2+4}{2+1}(0.1592) \times (0.1267)$$

= 2 x (0.1592) x (0.1267)

= 0.0403

$$f(4) = \frac{3+4}{3+1}(0.1592) \times (0.0403)$$

= 1.75 x (0.1592) x (0.0403)

= 0.0112

$$f(5) = \frac{4+4}{4+1}(0.1592) \times (0.0112)$$

= 1.6 x (0.1592) x (0.0112)

= 0.0029

Calculate the Expected Frequency,

X	F	f(x)	N . f(x)	Expected Frequency (E_i)
0	213	0.4998	400 x 0.4998 = 199.92	200
1	128	0.3183	400 x 0.3183 = 127.32	127
2	37	0.1267	400 x 0.1267 = 50.68	51
3	18	0.0403	400 x 0.0403 = 16.12	16
4	3	0.0112	400 x 0.0112 = 4.48	5
5	1	0.0029	400 x 0.0029 = 1.16	1

Compute the test Statistic of Chi-Square Test,

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

Frequency (O_i)	Expected Frequency (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
213	200	13	169	0.8450
128	127	1	1	0.0079
37	51	-14	196	3.8431
18	16	2	4	0.2500
3	5	-2	4	0.8000
1	1	0	0	0
			Total	5.7460

Degrees of Freedom

Calculate the Degrees of Freedom = n – p – 1

= 6 – 2 – 1

= 3

Table Value from Chi-Square Table in 3df = 7.815

Calculated Value = 5.7460

Conclusion

Since Calculated Value is less than the Table Value (i.e. 5.7460 < 7.815). So, we accept the null Hypothesis. Hence we conclude that the data is good fit for the Negative Binomial Distribution.

R Coding

#Input Data

```
x<-c(0,1,2,3,4,5)

frequency<-c(213,128,37,18,3,1)
```

#obtaining theoretical probability

```
mean<- sum(x*frequency)/sum(frequency)

mean

variance<-(sum(frequency*(x^2))/sum(frequency))- (sum(frequency*(x))/sum(frequency))^2

variance

p<-mean/variance

q<-1-p

r=round((p*mean)/q)

cumprob<-round(c(pnbinom(x,size=r,prob=p,lower.tail=TRUE)),4)

cumprob
```

#obtaining required columns in chi-square goodness of fit

```
prob<-seq(1,6)
for(i in 2:6)
{
prob[1]<-cumprob[1]
prob[i]<-round(cumprob[i]-cumprob[i-1],4)
}
exp_freq<-prob*(sum(frequency))
oi_ei_sq<-(frequency-exp_freq)^2
oi_ei<-round(oi_ei_sq/exp_freq,2)
```

 # Output Table, Chi-square test statistics & P-Value

```
out<-data.frame(cbind(x,frequency,cumprob,prob,exp_freq,oi_ei_sq,oi_ei))
chi_sq<-sum(oi_ei)
pvalue<-pchisq(chi_sq,6-2-1,lower.tail=FALSE)
pvalue
```

<u>R Output</u>

>mean

```
[1] 0.6825
```

>variance

```
[1] 0.8116938
```

>p

```
[1] 0.8408344
```

>r

[1] 4

Output

x	frequency	cumprob	prob	exp_freq	oi_ei_sq	oi_ei
0	213	0.4999	0.4999	200	169	0.8450
1	128	0.8181	0.3182	127	1	0.0079
2	37	0.9447	0.1266	51	196	3.8431
3	18	0.9850	0.0403	16	4	0.2500
4	3	0.9963	0.0113	5	4	0.8000
5	1	0.9991	0.0028	1	0	0

> chi_sq

[1] 5.746

> pvalue

[1] 0.1246435

Interpretation

Here the p-value is greater than the specified level of significance (i.e.) **0.125 > 0.05**, so we accept the null hypothesis. Hence, we conclude that the data is good fit for the Negative Binomial Distribution.

Observation Problem

Data was collected over a period of 10 years showing number of deaths from horse kicks in each of the 20 army crops from the 250 crops years. The distribution of deaths was follows.

No. of deaths	0	1	2	3	4	5	6
Pre-frequency	55	40	32	38	32	23	30

Test whether Negative Binomial distribution is a good fit for the given data

14. Fitting of Geometric Distribution

The following distribution relates to the number of accidents to 650 women working on highly explosive shells during 5-weeks period. Find Geometric distribution and gives very good fit to the data.

No. of seizceres	0	1	2	3	4	5
No. of People	450	132	41	22	3	2

Aim

Fit an appropriate Geometric distribution for the above data and calculate theoretical frequency and testing the goodness of fit of a Geometric Distribution.

Procedure

- Define the Null Hypothesis (H_0) and Alternate Hypothesis (H_1) for given data.
- For a given Geometric Distribution Calculate respecting Mean and variance. And also calculate p and q.

$$Mean\ (\lambda) = \frac{\sum fx}{\sum f}$$

$$Variance = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2$$

- Compute the respective probability Mass function is,

$$P(x) = q^x \bullet p,\ x = 0,1,2,....$$

- Compute the respective Cumulative Density function is,

$$P(x) = 1 - q^{x+1}, \qquad x = 0,1,2,....$$

- Compute the Expected Frequency [E_i].
- Compute the test Statistic of Chi-Square Test,

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

- Taking decision about the hypothesis H_0 with respect to the level of significance and degree of freedom.

Calculation

Hypothesis

H_0: The data is good fit for the Geometric Distribution.

H_1: The data is not good fit for the Geometric Distribution.

Calculate mean and Variance,

x	f	fx	x^2	fx^2
0	450	0	0	0
1	132	132	1	132
2	41	82	4	164
3	22	66	9	198
4	3	12	16	48
5	2	10	25	50
Total	N = 650	302		592

$$\mu = \frac{\sum_{i=1}^{n} fx}{\sum_{i=1}^{n} f}$$

$$= \frac{302}{650}$$

$$= 0.4646 = \frac{q}{p} \longrightarrow (1)$$

$$\sigma^2 = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2$$

$$= \frac{592}{650} - \left(\frac{302}{650}\right)^2$$

$$= 0.9108 - (0.4646)^2$$

$$= 0.9108 - 0.2159$$

$$= 0.6949 = \frac{q}{p^2} \longrightarrow (2)$$

$$\frac{Mean}{Variance} = \frac{q}{p} \times \frac{p^2}{q}$$

$$\frac{0.4646}{0.6949} = \frac{q}{p} \times \frac{p^2}{q}$$

$$p = 0.6686$$

$$q = 1 - p$$

$$= 1 - 0.6689$$

$$= 0.3314$$

Calculate Cumulative Density Function for Geometric Distribution is given by,

$$P(x) = 1 - q^{x+1}, \qquad x = 0,1,2,....$$

$P(0) = 1 - (0.3314)^{0+1}$

= 0.6686

$P(1) = 1 - (0.3314)^{1+1}$

= 0.8902

$P(2) = 1 - (0.3314)^{2+1}$

= 0.9636

$P(3) = 1 - (0.3314)^{3+1}$

= 0.9879

$P(4) = 1 - (0.3314)^{4+1}$

= 0.9960

$P(5) = 1 - (0.3314)^{5+1}$

= 0.9987

Calculate the Expected Frequency,

X	p(x)	Δp(x)	N . Δp(x)	Expected Frequency (E_i)
0	0.6686	0.6686	434.59	435
1	0.8902	0.8902 - 0.6686 = 0.2216	144.04	144
2	0.9636	0.9636 - 0.8902 = 0.0734	47.71	48
3	0.9879	0.9879 - 0.9636 = 0.0243	15.79	16
4	0.9960	0.9960 - 0.9879 = 0.0081	5.27	5
5	0.9987	0.9987 - 0.9960 = 0.0027	1.76	2

Compute the test Statistic of Chi-Square Test,

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

Frequency (O_i)	Expected Frequency (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
450	435	15	225	0.52
132	144	-12	144	1
41	48	-7	49	1.021
22	16	6	36	2.25
3	5	-2	4	0.8
2	2	0	0	0
			Total	5.5880

Degrees of Freedom

Calculate the Degrees of Freedom = n – p – 1

= 6 – 2 – 1

= 3

Table Value from Chi-Square Table in 3df = 7.815

Calculated Value = 5.5880

Conclusion

Since Calculated Value is less than the Table Value (i.e. 5.5880 < 7.815). So, we accept the null Hypothesis. Hence we conclude that the data is good fit for the Geometric Distribution.

R Coding

#Input Data

```
x<-c(0,1,2,3,4,5)

frequency<-c(450,132,41,22,3,2)
```

#obtaining theoretical probability

```
mean<- sum(x*frequency)/sum(frequency)

mean

variance<-(sum(frequency*(x^2))/sum(frequency))- (sum(frequency*(x))/sum(frequency))^2

variance

p<-mean/variance

p

q<-1-p

cumprob<-round(c(pgeom(x ,prob=p,lower.tail=TRUE)),4)

cumprob
```

#obtaining required columns in chi-square goodness of fit

```
prob<-seq(1,6)

for(i in 2:6)

{

prob[1]<-cumprob[1]

prob[i]<-round(cumprob[i]-cumprob[i-1],4)

}

exp_freq<-round(prob*(sum(frequency)))

exp_freq
```

```
oi_ei_sq<-(frequency-exp_freq)^2

oi_ei_sq

oi_ei<-round(oi_ei_sq/exp_freq,4)

oi_ei
```

 # Output Table, Chi-square test statistics & P-Value

```
out<-data.frame(cbind(x,frequency,cumprob,prob,exp_freq,oi_ei_sq,oi_ei))

out

chi_sq<-sum(oi_ei)

chi_sq

pvalue<-pchisq(chi_sq,6-2-1,lower.tail=FALSE)

pvalue
```

R Output

>mean

[1] 0.4646154

>variance

[1] 0.6949018

>p

[1] 0.6686058

>q

[1] 0.3313942

Output

x	frequency	cumprob	prob	exp_freq	oi_ei_sq	oi_ei
0	450	0.6686	0.6686	435	225	0.5172
1	132	0.8902	0.2216	144	144	1.0000
2	41	0.9636	0.0734	48	49	1.0208
3	22	0.9879	0.0243	16	36	2.2500
4	3	0.9960	0.0081	5	4	0.8000
5	2	0.9987	0.0027	2	0	0.0000

> chi_sq

[1] 5.588

> pvalue

[1] 0.133469

Interpretation

Here the p-value is greater than the specified level of significance (i.e.) **0.133 > 0.05**, so we accept the null hypothesis. Hence, we conclude that the data is good fit for the Geometric Distribution.

Observation Problem

If data from 100 epileptic people sampled at random in one year.

No. of seizceres	0	2	4	6	8	10
No. of People	17	21	18	11	16	17

Test whether Geometric distribution is a good fit for the given data.

15. Fitting of Hyper Geometric Distribution

A create contains 50 light bulbs of which 5 are defective and 45 are not. A Quality Control Inspector randomly samples 4 bulbs without replacement. Let X = the number of defective bulbs selected. Find the probability mass function, f(x), of the discrete random variable X = 0, 1, 2, 3, 4.

Aim

Fit an appropriate Hyper geometric Distribution to the above data.

Procedure

- Compute the respective probability using the recurrence relation of Hyper geometric Distribution.

$$P(x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}, \; x = 0,1,2,\ldots.$$

Calculation

We substituting the values,

N = 50, k = 5, n = 4, x = 0, 1, 2, 3, 4.

For a Hyper Geometric Distribution the probability mass function is given by,

$$P(x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}, \; x = 0,1,2,\ldots.$$

$$P(0) = \frac{\binom{5}{0}\binom{45}{4}}{\binom{50}{4}}$$

$$= \frac{1(148995)}{230300}$$

= 0.64696

$$P(1) = \frac{\binom{5}{1}\binom{45}{3}}{\binom{50}{4}}$$

$$= \frac{5(14190)}{230300}$$

= 0.30808

$$P(2) = \frac{\binom{5}{2}\binom{45}{2}}{\binom{50}{4}}$$

$$= \frac{10(990)}{230300}$$

= 0.04298

$$P(3) = \frac{\binom{5}{3}\binom{45}{1}}{\binom{50}{4}}$$

$$= \frac{10(45)}{230300}$$

= 0.00196

$$P(4) = \frac{\binom{5}{4}\binom{45}{0}}{\binom{50}{4}}$$

$$= \frac{5(1)}{230300}$$

= 0.00002

The Probability Table is given by,

X	p(x)
0	0.64696
1	0.30808
2	0.04298
3	0.00196
4	0.00002

Conclusion

The Probability of defective bulbs is 0.64696, 0.30808, 0.04298, 0.00196, and 0.00002.

R Coding

 # Given data

```
N <- 50

k <- 5

n <- 4

x <- c(0,1,2,3,4)

m <- (N-k)

f <- (n-x)
```

 # Calculations for Fitting of Distribution

```
cumprob<-round(c(phyper(x, k, m, n, lower.tail=TRUE)),5)
```

#obtaining required columns in chi-square goodness of fit

```
prob<-seq(1,5)

for(i in 2:5)

{

prob[1]<-cumprob[1]

prob[i]<-round(cumprob[i]-cumprob[i-1],5)

}

prob
```

 # Output Table, Chi-square test statistics & P-Value

```
out<-data.frame(cbind(x,cumprob,prob))
out
```

R Output

Output

x	cumprob	prob_x
0	0.64696	0.64696
1	0.95504	0.30808
2	0.99802	0.04298
3	0.99998	0.00196
4	1.00000	0.00002

Result

The Probability of defective bulbs is 0.64696, 0.30808, 0.04298, 0.00196, and 0.00002

Observation Problem

A company (the producer) supplies microprocessors to a manufacturer (the consumer) of electronic equipment. The microprocessors are supplied in batches of 50. The consumer regards a batch as acceptable provided that there are not more than 5 defective microprocessors in the batch. Rather than test all of the microprocessors in the batch, 10 are selected at random and tested. Find the probability that out of a sample of 10, x = 0, 1, 2, 3, 4, 5 are defective when there are actually 5 defective microprocessors in the batch.