MATH 536 Homework 3

Gustavo Esparza

9/25/2019

1

Generate 10,000 random observations from a Poisson distribution with known λ. Run a χ^2 goodness of fit test for testing whether the random observations were indeed generated from a Poisson distribution with the predetermined mean.

Here is a table that displays our 10,000 generated poisson observations with $\lambda = 5$ as well as their frequency:

data	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Freq	68	315	835	1405	1822	1715	1486	1020	657	338	184	91	37	19	7	1

Here is a table that displays the expected frequency of each count for the poisson distribution with λ =5:

values	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
expected	67	337	842	1404	1755	1755	1462	1044	653	363	181	82	34	13	5	2

Finally, here are the results from our Chi Square test:

```
##
##  Chi-squared test for given probabilities with simulated p-value
##  (based on 2000 replicates)
##
## data:  c(freq, 0)
## X-squared = 13.315, df = NA, p-value = 0.6562
```

As expected, our p-value is quite large and therefore suggests that our data does indeed originate from a poisson distirbution with $\lambda = 5$.

2

For testing the hypothesis $H_0 : p_i = p_i^0,$ for $i = 1, \ldots, k$,the χ^2 goodness of fittest is:

$$Q = \sum_{i=1}^{k} \frac{(N_i - np_i^0)^2}{np_i^0}$$

Show that if $p_i^0 = 1/k$ for $i = 1, \ldots, k$, then the χ^2 goodness of fit test can be written as:

$$Q = \left(\frac{k}{n} \sum_{i=1}^{K} N_i^2\right) - n$$

Here, we have the following result:

$$Q = \sum_{i=1}^{k} \frac{(N_i - n \times \frac{1}{k})^2}{\frac{n}{k}} = \frac{k}{n} \sum_{i=1}^{k} \left(N_i^2 - \frac{2N_i n}{k} + \frac{n^2}{k^2}\right) = \sum_{i=1}^{K} \left(\frac{kN_i^2}{n} - 2N_i + \frac{n}{k}\right)$$

Performing the designated sums leads to the following result:

$$\frac{k}{n}\sum_{i=1}^{K} N_i^2 - 2n + n = \frac{k}{n}\sum_{i=1}^{K} N_i^2 - n$$

3

Consider a sequence of n Bernoulli trials with unknown probability of success p on each trial. We want to test the hypothesis:

$$H_0 : p = p_0$$
$$H_a : p \neq p_0$$

where p_0 is a given number on the interval (0, 1).

A

Show that the χ^2 goodness-of-fit test:

$$Q = \sum_{i=1}^{k} \frac{(N_i - np_i^0)^2}{np_i^0}$$

can be rewritten as

$$Q = \frac{n(\bar{X} - p_0)^2}{p_0(1 - p_0)}$$

where $\bar{X}$ denotes the proportion of successes in the n trials.

Since we have a binomial distribution, our summation of k involves two outcomes: success or failure. Thus, we have:

$$Q = \frac{(N_1 - np_0)^2}{np_0} + \frac{(N_2 - n(1 - P_0))^2}{n(1 - P_0)}$$

Now, considering $\bar{X}$, we have

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} \rightarrow n\bar{X} = N_1 \text{ number of successes, } n - n\bar{X} = N_2 \text{ number of failures}$$

Thus, our test statistic can be expressed as follows

$$Q = \frac{(n\bar{x} - np_0)^2}{np_0} + \frac{n - n\bar{X} - n + np_0}{n(1 - p_0)} = \frac{(n\bar{x} - np_0)^2}{np_0} + \frac{(n - np_0 - n\bar{X})^2}{n(1 - p_0)} = \frac{n(\bar{x} - p_0)^2}{p_0} + \frac{n(p_0 - \bar{x})^2}{n(1 - p_0)}$$

combining our two fractions gives us the following result:

$$\frac{n(\bar{X} - p_0)^2 - p_0 n(\bar{X} - p_0)^2 + np_0(p_0 - \bar{X})^2}{p_0(1 - p_0)} = \frac{n(\bar{X} - p_0)^2}{p_0(1 - p_0)}$$

As desired.

B

Assume that H_0 is true, prove that as $n \to \infty$, the cumulative distribution function of Q converges to the cumulative distribution function of the χ^2 distribution with 1 degree of freedom.

We will begin by defining our CDF:

$$P(Q \leq q) = p(\frac{n(\bar{X} - p_0)^2}{p_0(1-p_0)} \leq q) = p\left(\left[\frac{\sqrt{n}(\bar{X} - p_0)^2}{\sqrt{p_0(1-p_0)}}\right]^2 \leq q\right) = p\left(\left[\frac{(\bar{X} - p_0)}{\sqrt{p_0(1-p_0)/n}}\right]^2 \leq q\right)$$

Since the quantity inside the exponent is the standardization of our mean $\bar{X}$ (since $E(\bar{X}) = p_0$ and $V(\bar{X}) = p_0(1-p_0)/n)$), we know that by the Central Limit Theorem:

$$\frac{(\bar{X} - p_0)}{\sqrt{p_0(1-p_0)/n}} \longrightarrow N(0,1)$$

and since the square of the standard normal approaches χ^2_1 in probability, we have shown that the cumulative distribution function of Q converges to the cumulative distribution function of the χ^2 distribution with 1 degree of freedom.

4

Consider an $R \times C$ contingency table and the following hypothesis to test the independence of the row variable and the column variable:

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$$
$$H_a : \text{ otherwise}$$

A

Show that the χ^2 test statistic can be written as:

$$Q = \Big(\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{N_{ij}^2}{E_{ij}}\Big) - n$$

$$\sum_{i=1}^{R}\sum_{i=1}^{C}\frac{(N_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{R}\sum_{i=1}^{C}\Big(\frac{N_{ij}^2 - 2N_{ij}E_{ij} + E_{ij}^2}{E_{ij}}\Big)$$
$$= \sum_{i=1}^{R}\sum_{i=1}^{C}\frac{N_{ij}^2}{E_{ij}} - 2\sum_{i=1}^{R}\sum_{i=1}^{C}N_{ij} + \sum_{i=1}^{R}\sum_{i=1}^{C}E_{ij}$$
$$= \sum_{i=1}^{R}\sum_{i=1}^{C}\frac{N_{ij}^2}{E_{ij}} - 2n + n = \sum_{i=1}^{R}\sum_{i=1}^{C}\frac{N_{ij}^2}{E_{ij}} - n$$

B

Show that if $C = 2$, then Q can be written as:

$$\frac{n}{N_{+2}}\Big(\sum_{i=1}^{R}\frac{N_{i1}^2}{E_{i1}} - N_{+1}\Big)$$

For $C = 2$, we have the following using part A

$$Q = \sum_{i=1}^{R}\frac{N_{i1}^2}{E_{i1}} + \sum_{i=1}^{R}\frac{N_{i2}^2}{E_{i2}} - n$$

Now focusing on our second summation, we have

$$\sum_{i=1}^{R}\frac{N_{i2}^2}{E_{i2}} = \sum_{i=1}^{R}\frac{(N_{i+} - N_{i1})^2}{E_{i2}} = \sum_{i=1}^{R}\frac{N_{i+}^2}{E_{i2}} - 2\sum_{i=1}^{R}\frac{N_{i+}N_{i1}}{E_{i2}} + \sum_{i=1}^{R}\frac{N_{i1}^2}{E_{i2}}$$

With this expression we will make the following substitution for E_{i2}:

$$E_{i2} = \frac{N_{i+}N_{+2}}{n}$$

This gives us the following expression:

$$\sum_{i=1}^{R}\frac{N_{i2}^2}{E_{i2}} = \sum_{i=1}^{R}\frac{N_{i+}^2 \times n}{N_{i+}N_{+2}} - 2\sum_{i=1}^{R}\frac{N_{i+}N_{i1} \times n}{N_{i+}N_{+2}} + \sum_{i=1}^{R}\frac{N_{i1}^2 \times n}{N_{i+}N_{+2}}$$

Now, we will make the following substitution for N_{i+} in the last term:

$$N_{i+} = \frac{nE_{i1}}{N_{+1}}$$

This gives the following result:

$$\sum_{i=1}^{R}\frac{N_{i2}^2}{E_{i2}} = \sum_{i=1}^{R}\frac{N_{i+}^2 \times n}{N_{i+}N_{+2}} - 2\sum_{i=1}^{R}\frac{N_{i+}N_{i1} \times n}{N_{i+}N_{+2}} + \sum_{i=1}^{R}\frac{N_{i1}^2 \times n \times N_{+1}}{N_{+2}E_{i1}n}$$

Simplifying this expression and rewriting Q leads to the following:

$$Q = \sum_{i=1}^{R}\frac{N_{i1}^2}{E_{i1}} + \frac{n}{N_{+2}}\sum_{i=1}^{R}N_{i+} - \frac{2n}{N_{+2}}\sum_{i=1}^{R}N_{i1} + \frac{N_{+1}}{N_{+2}}\sum_{i=1}^{R}\frac{N_{i1}^2}{E_{i1}} - n$$

After combining like terms and simplifying summations, we obtain the expression for Q

$$\Big(1 + \frac{N_{+1}}{N_{+2}}\Big)\sum_{i=1}^{R}\left(\frac{N_{i1}^2}{E_{i1}}\right) + \frac{n^2}{N_{+2}} - \frac{2nN_{+1}}{N_{+2}} - n = \Big(1 + \frac{N_{+1}}{N_{+2}}\Big)\sum_{i=1}^{R}\left(\frac{N_{i1}^2}{E_{i1}}\right) + \frac{n}{N_{+2}}(n - 2N_{+1} - N_{+2})$$

since $\Big(1 + \frac{N_{+1}}{N_{+2}}\Big) = \Big(\frac{N_{+2}}{N_{+2}} + \frac{N_{+1}}{N_{+2}}\Big) = \frac{n}{N_{+2}}$, we have a final result given by:

$$\frac{n}{N_{+}2}\left[\sum_{i=1}^{R}\left(\frac{N_{i1}^2}{E_{i1}}\right)-N_{+1}\right]$$

5

Suppose that 400 people are chosen at random from a large population, and that each person in the sample specifies which one of five breakfast cereals she/he most prefers. For $i = 1, \ldots, 5$, let p_i denote the proportion of the population that prefers cereal i, and let N_i denote the number of persons in the sample who prefer cereal i. If we test the following hypotheses at the $\alpha = 0.01$ level:

$$H0 : p_1 = p_2 = \cdots = p_5 \; H_a : otherwise$$

What values of

$$\sum_{i=1}^{5} N_i^2$$

would reject the null hypothesis?

We have the following information:

$$Q = \sum_{i=1}^{5} \frac{(N_i - E_i)^2}{E_i} \; , \; \chi^2_{.01,4} = 13.277 \; , \; E_i = 400/5 = 80$$

Using this knowledge, we have

$$\sum_{i=1}^{5} \frac{(N_i - 80)^2}{80} \geq 13.277 \rightarrow \frac{1}{80}\sum_{i=1}^{5}(N_i^2 - 160N_i + 80^2) \geq 13.277$$

Expanding our summation gives us the following:

$$\frac{1}{80}\sum_{i=1}^{5}(N_i^2) - 2\sum_{i=1}^{5} N_i + \sum_{i=1}^{5} 80 \geq 13.277 \rightarrow \frac{1}{80}\sum_{i=1}^{5} N_i^2 - 2(400) + 400 \geq 13.277$$

Finally, solving for our sum provides the following:

$$\sum_{i=1}^{5}(N_i)^2 \geq 33062.16$$

Thus, any value for our summation that exceeds this value of 33062.16 will reject the null hypothesis and suggest the 5 proportions are not all equal.

6

Suppose that a χ^2 test of independence is to be applied to the elements of a 2×2 table. Show that the χ^2 test statistic Q can be wriiten as follows:

$$Q = \frac{n(N_{11}N_{22} - N_{12}N_{21})^2}{N_{1+}N_{2+}N_{+1}N_{+2}}$$

We have

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(N_{ij}-E_{ij})^2}{E_{ij}},\ \text{where } E_{ij}=\frac{N_{i+}N_{+j}}{n}$$

We will first consider the quanitity $(N_{ij}-E_{ij})^2$. Leaving i and j as variables that can represent the values of 1 and 2, we can obtain the following:

$$(N_{ij}-E_{ij})^2=\left(N_{ij}-\frac{N_{i+}N_{+j}}{n}\right)^2=\frac{1}{n^2}(n\times N_{ij}-N_{i+}N_{+j})^2$$

Expanding n as the sum of all elements and expressing the marginal distributions as sums of two columns/rows, we find the following result:

$$\begin{aligned}(N_{ij}-E_{ij})^2&=\frac{1}{n^2}\Big[(N_{ii}+N_{ij}+N_{ji}+N_{jj})N_{ij}-(N_{ii}+N_{ij})(N_{ij}+N_{jj})\Big]^2\\&=\frac{1}{n^2}[N_{ii}N_{ij}+N_{ij}N_{ij}+N_{ji}N_{ij}+N_{jj}N_{ij}-N_{ii}N_{ij}-N_{ii}N_{jj}-N_{ij}N_{ij}-N_{ij}N_{jj}]\\&=\frac{1}{n^2}[-(N_{ii}N_{jj}-N_{ji}N_{ij})]^2=\frac{1}{n^2}[N_{ii}N_{jj}-N_{ji}N_{ij}]^2\end{aligned}$$

Which is the desired result. As for $\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{E_{ij}}$, we have the following sum:

$$\frac{n}{N_{1+}N_{+1}}+\frac{n}{N_{1+}N_{+2}}+\frac{n}{N_{2+}N_{+1}}+\frac{n}{N_{2+}N_{+2}}=\frac{n(N_{2+}N_{+2}+N_{2+}N_{+1}+N_{1+}N_{+2}+N_{1+}N_{+1})}{N_{1+}N_{+1}N_{2+}N_{+2}}$$

After combining like terms and adding row/column totals, we have the following:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{E_{ij}}&=\frac{n[(N_{2+})(N_{+2}+N_{+1})+(N_{1+})(N_{+2}+N_{+1})]}{N_{1+}N_{+1}N_{2+}N_{+2}}\\&=\frac{n[(N_{2+}+N_{1+})(N_{+2}+N_{+1})]}{N_{1+}N_{+1}N_{2+}N_{+2}}=\frac{n(n*n)}{N_{1+}N_{+1}N_{2+}N_{+2}}=\frac{n^3}{N_{1+}N_{+1}N_{2+}N_{+2}}\end{aligned}$$

Now, combining our two realized double summations, we have the final result of:

$$\frac{1}{n^2}[N_{ii}N_{jj}+N_{ji}N_{ij}]^2\times\frac{n^3}{N_{1+}N_{+1}N_{2+}N_{+2}}=\frac{n(N_{11}N_{22}-N_{12}N_{21})^2}{N_{1+}N_{2+}N_{+1}N_{+2}}$$

This final form is the desired quantity for our 2×2 contingency table.

Appendix

1

```
set.seed(100)
data = rpois(10000,5)
df = mutate_all(data.frame(table(data)),function(x) as.numeric(as.character(x)))

freq = df$Freq
values = df$data

t(df) %>%
kable() %>%
kable_styling()
```

Here is a table that displays the expected frequency of each count for the poisson distribution with λ =5:

```
probs = dpois(values, lambda=5)
comp = 1-sum(probs)

expected = round(10000*probs)
expected_table = cbind(values,expected)

t(expected_table) %>%
kable() %>%
kable_styling()
```

Finally, here are the results from our Chi Square test:

```
#chisq = sum(( freq-expected)^2/(expected))
chisq.test(x=c(freq,0), p=c(probs,comp),simulate.p.value = T)
```