Math 362: Mathematical Statistics II

Le Chen
le.chen@emory.edu
chenle02@gmail.com

Emory University
Atlanta, GA

Last updated on Spring 2021
Last compiled on January 15, 2023

2021 Spring

Chapter 10. Goodness-of-fit Tests

Rationale

! We want to test if the c.d.f. $F_Y(\cdot)$ is given by the true c.d.f. $F_0(\cdot)$, i.e.,

$$H_0 : F_Y(y) = F_0(y) \quad v.s. \quad H_1 : F_Y(y) \neq F_0(y)$$

$\sim$ By properly partitioning the domain, the random sample should follow *an induced multinomial distribution*.

$\Longrightarrow$ Then testing $F_Y(\cdot) = F_0(\cdot)$ reduces to testing the induced multinomial distribution of the following form:

$$H_0 : p_1 = p_1', \cdots, p_n = p_n'$$

$$v.s.$$

$$H_1 : p_i \neq p_i' \quad \text{for at least one } i$$

How

1. Suppose we are sampling from the c.d.f. $F(y)$

2. Divide the range of the distribution into k mutually exclusive and exhausive intervals, say $I_1, \cdots, I_k$.

3. Let $\pi_i = \mathbb{P}(X \in I_i)$, $i = 1, \cdots, k$.

4. Let $O_1, \cdots, O_k$ be the respective observed numbers of the observations $X_1, \cdots, X_n$ in the intervals $I_1, \cdots, I_k$.

5. Then $O = (O_1, \cdots, O_k) \sim$ multinomial distribution with $(\pi_1, \cdots, \pi_k)$, i.e.,

$$\mathbb{P}\left(O_1 = o_1, \cdots, O_k = o_k\right) = \frac{n!}{\prod_{i=1}^k o_i!} \prod_{i=1}^k \pi_i^{o_i}$$

   with $\sum_{i=1}^k \pi_i = 1$, $\sum_{i=1}^k o_i = n$, and

$$\mathbb{E}[O_i] = n\pi_i =: e_i, \quad \mathrm{Var}(O_i) = n\pi_i(1 - \pi_i)$$

6. When $k = 2$, by CLT, as $n \to \infty$,

$$\frac{O_1 - n\pi_1}{\sqrt{n\pi_1(1-\pi_1)}} \xrightarrow{d} N(0,1) \quad \Longrightarrow \quad \frac{(O_1 - n\pi_1)^2}{n\pi_1(1-\pi_1)} \xrightarrow{d} \chi_1^2$$

$$\|$$

$$\frac{(O_1 - n\pi_1)^2}{n\pi_1} + \frac{(O_2 - n\pi_2)^2}{n\pi_2}$$

$$\|$$

$$\frac{(O_1 - e_1)^2}{e_1} + \frac{(O_2 - e_2)^2}{e_2}$$

Hence, as $n \to \infty$,

$$\sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} \xrightarrow{d} \chi_{k-1}^2$$

7. For general k,

$$\sum_{i=1}^{k} \frac{(O_i - n\pi_i)^2}{n\pi_i} = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$$

follows a complicated, but exact, distribution, from which, one can show

$$\sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} \xrightarrow{d} \chi^2_{k-1}$$

$$\Downarrow$$

Thm. When n is large enough, namely, when $n\pi_i \geq 5$ for all i,

$$D = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} \overset{appr.}{\sim} \chi^2_{k-1}.$$

Rmk: The above is called *Pearson's chi-square test*. It is asymptotically equivalent to the generalized likelihood ratio test.

Alternative: G-test

– the likelihood ration test for multinomial model

1. Under $H_0 : \pi_i = p_i$, $i = 1, \cdots, k$, the MLE of π_i are

$$\widetilde{\pi}_i = p_i = \frac{np_i}{n} = \frac{e_i}{n}, \qquad \forall i.$$

2. When there are no constraints, for $i = 1, \cdots, k-1$,

$$\frac{\partial}{\partial \pi_i} \ln L(\pi_1, \cdots, \pi_{k-1} | o_1, \cdots, o_k) = 0, \quad 1 \le i \le k-1$$

$$\Updownarrow$$

$$\frac{o_i}{\widehat{\pi}_i} = \frac{o_k}{1 - \widehat{\pi}_1 - \cdots - \widehat{\pi}_{k-1}}, \quad 1 \le i \le k-1$$

$$\Updownarrow$$

$$\widehat{\pi}_i = \frac{o_i}{n}, \quad 1 \le i \le k.$$

$\Rightarrow$

$$\lambda := \ln\left(\frac{L(\widetilde{\pi}_1, \cdots, \widetilde{\pi}_{k-1} | o_1, \cdots, o_k)}{L(\widehat{\pi}_1, \cdots, \widehat{\pi}_{k-1} | o_1, \cdots, o_k)}\right) = \log\left(\frac{\prod_{i=1}^k \widetilde{\pi}_i^{o_i}}{\prod_{i=1}^k \widehat{\pi}_i^{o_i}}\right)$$

$$= \sum_{i=1}^{k} o_i \ln\left(\frac{\widetilde{\pi}_i}{\widehat{\pi}_i}\right)$$

$$= \sum_{i=1}^{k} o_i \ln\left(\frac{e_i}{o_i}\right)$$

Critical region: $\lambda < \lambda_* < 0$.

Def.

$$G := -2\lambda = -2\sum_{i=1}^{k} o_i \ln\left(\frac{e_i}{o_i}\right) = 2\sum_{i=1}^{k} o_i \ln\left(\frac{o_i}{e_i}\right)$$

$G \overset{approx.}{\sim} \chi^2_{k-1}$ for large n.

Critical region: $G \geq G_* = \chi^2_{1-\alpha,k-1}$.

Relation G-test and Pearson's Chi square test

By second order Taylor expanson around 1,

$$\begin{aligned} G &= -2\sum_{i=1}^{k} o_i \ln\left(\frac{e_i}{o_i}\right) \\ &\approx -2\sum_{i=1}^{k} o_i \left[\left(\frac{e_i}{o_i} - 1\right) - \frac{1}{2}\left(\frac{e_i}{o_i} - 1\right)^2\right] \\ &= -2\sum_{i=1}^{k}(e_i - o_i) + \sum_{i=1}^{k} o_i \left(\left(1 - \frac{o_i}{e_i}\right) + \frac{o_i}{e_i}\right)\left(\frac{e_i}{o_i} - 1\right)^2 \\ &= 0 + \sum_{i=1}^{n} \frac{o_i^2}{e_i}\left(1 - \frac{o_i}{e_i}\right)^3 + \sum_{i=1}^{k} \frac{(e_i - o_i)^2}{e_i} \\ &\approx \sum_{i=1}^{k} \frac{(e_i - o_i)^2}{e_i} \\ &\qquad\quad || \\ &\qquad\quad D \end{aligned}$$

$\therefore$ Pearson's Chi-square test is an approximation of G-test.

E.g. 1 *Benford's law*:

Table 10.3.1

Digit, i	$\log_{10}(i+1) - \log_{10}(i)$
1	0.301
2	0.176
3	0.125
4	0.097
5	0.079
6	0.067
7	0.058
8	0.051
9	0.046

Initial digits

Digit	Observed, k_i
1	111
2	60
3	46
4	29
5	26
6	22
7	21
8	20
9	20
	355

Use this law to check whether the bookkeepers have made up entries.

Assume that bookkeepers are not aware of Benford's law.

Sol. The test should be

$$
\begin{aligned}
&H_0 : p_1 = p_{10}, \cdots, p_9 = p_{90} \\
&\qquad\qquad v.s. \\
&H_1 : p_i \neq p_{i0} \quad \text{for at least one } i = 1, \cdots, 9.
\end{aligned}
$$

Critical region: $(\chi^2_{.95,8}, \infty) = (15.507, \infty)$.

Compute the D and G scores:

Digit	o_i	p_i	e_i	$(o_i - e_i)^2/e_i$	$2o_i \ln(e_i/o_i)$
1	111	0.301			
2	60	0.176			
3	46	0.125			
4	29	0.097			
5	26	0.079			
6	22	0.067			
7	21	0.058			
8	20	0.051			
9	20	0.046			
sum	355	1	355	$d =$ ____	$g =$ ____

Digit	o_i	p_i	e_i	$(o_i - e_i)^2/e_i$	$2o_i \ln(e_i/o_i)$
1	111	0.301	106.9	0.16	8.449
2	60	0.176	62.5	0.10	-4.860
3	46	0.125	44.4	0.06	3.309
4	29	0.097	34.4	0.86	-9.963
5	26	0.079	28.0	0.15	-3.937
6	22	0.067	23.8	0.13	-3.433
7	21	0.058	20.6	0.01	0.828
8	20	0.051	18.1	0.20	3.982
9	20	0.046	16.3	0.82	8.109
sum	355	1	355	$d = \underline{2.49}$	$g = \underline{2.48}$

Conclusion: Fail to reject.

```
> #   EX 10.3.2
> library(data.table)
> mydat <- fread('http://math.emory.edu/~lchen41/teaching/2020_Spring/Case_10-3-2.data')
trying URL 'http://math.emory.edu/~lchen41/teaching/2020_Spring/Case_10-3-2.data'
Content type 'unknown' length 153 bytes
==================================================
downloaded 153 bytes

> head(mydat)
   Digit  Oi    Pi
1:     1 111 0.301
2:     2  60 0.176
3:     3  46 0.125
4:     4  29 0.097
> pi = mydat[,3]
> oi = mydat[,2]
> n = sum(oi)
> ei = n*pi
> di = (ei-oi)^2/ei
> gi =  2*oi*log(oi/ei)
> print(paste("Using Pearson's test, D value is equal to ", round(sum(di),3)))
[1] "Using Pearson's test, D value is equal to  2.491"
> print(paste("Using the G-test, G value is equal to ", round(sum(gi),3)))
[1] "Using the G-test, G value is equal to  2.484"
```

Codes available

`http://math.emory.edu/~lchen41/teaching/2020_Spring/Case_10-3-2.R`

E.g. 2 Test for randomness

Is the following sample of size 40 from $f_Y(y) = 6y(1-y)$, $y \in [0,1]$?

Table 10.3.4

0.18	0.06	0.27	0.58	0.98
0.55	0.24	0.58	0.97	0.36
0.48	0.11	0.59	0.15	0.53
0.29	0.46	0.21	0.39	0.89
0.34	0.09	0.64	0.52	0.64
0.71	0.56	0.48	0.44	0.40
0.80	0.83	0.02	0.10	0.51
0.43	0.14	0.74	0.75	0.22

Sol. Test continuous pdf $\rightarrow$ reduce to a set of classes:

Table 10.3.5

Class	Observed Frequency, k_i	P_{i_o}	$40p_{i_o}$
$0 \le y < 0.20$	8	0.104	4.16
$0.20 \le y < 0.40$	8	0.248	9.92
$0.40 \le y < 0.60$	14	0.296	11.84
$0.60 \le y < 0.80$	5	0.248	9.92
$0.80 \le y < 1.00$	5	0.104	4.16

Table 10.3.6

Class	Observed Frequency, k_i	P_{i_o}	$40p_{i_o}$
$0 \le y < 0.40$	16	0.352	14.08
$0.40 \le y < 0.60$	14	0.296	11.84
$0.60 \le y \le 1.00$	10	0.352	14.08

$$d = \cdots = 1.84.$$

Critical region: $(\chi^2_{.95,2}, \infty) = (5.992, \infty)$.

Conclusion: Fail to reject.

```
> # Case Study 10.3.2
> # Read data from the URL link
>  library (data.table)
> mydat <- fread('http://math.emory.edu/~lchen41/teaching/2020_Spring/EX_10-3-1.data')
 trying  URL 'http://math.emory.edu/~lchen41/teaching/2020_Spring/EX_10-3-1.data'
Content type 'unknown' length 234 bytes
==================================================
downloaded 234 bytes

>d(mydat)
    Col1 Col2 Col3 Col4 Col5
    1:  0.18 0.06 0.27 0.58 0.98
    2:  0.55 0.24 0.58 0.97 0.36
    3:  0.48 0.11 0.59 0.15 0.53
    4:  0.29 0.46 0.21 0.39 0.89
    5:  0.34 0.09 0.64 0.52 0.64
    6:  0.71 0.56 0.48 0.44 0.40
# Conditions for  lower bounds
> lb=c(0,0.40,0.60)
> # Conditions for  upper bounds
> up=c(0.40,0.60,1.00)
> # Store the  results  in  d
> oi  <- seq(1:length(lb))
> pi  <- seq(1:length(lb))
> integrand <- function(y)  {6*y*(1-y)}
> for  (i  in  c(1:length(lb)))  {
+    oi[i]  <- table(mydat>=lb[i]  & mydat<up[i])[2]
+    pi[i]  <- integrate(integrand,    lb[i],  up[i])$value[1]
+    print(paste("the",  i,"th  bin  has",  oi[i],
+          "entries  and pi  is  equal to",  pi[i]))
+ }
```

```
[1] "the 1 th bin has 16 entries and pi is equal to 0.352"
[1] "the 2 th bin has 14 entries and pi is equal to 0.296"
[1] "the 3 th bin has 10 entries and pi is equal to 0.352"
> pi <- unlist(pi)
> n <- sum(oi)
> ei <- n*pi
> di <- (ei-oi)^2/ei
> gi <- 2*oi*log(oi/ei)
> rbind(oi,pi,ei,di,gi)
         [,1]       [,2]      [,3]
oi 16.0000000 14.0000000 10.000000
pi  0.3520000  0.2960000  0.352000
ei 14.0800000 11.8400000 14.080000
di  0.2618182  0.3940541  1.182273
gi  4.0906679  4.6920636 -6.843405
> print(paste("Using Pearson's test, D value is equal to ",round(sum(di),3)))
[1] "Using Pearson's test, D value is equal to  1.838"
> print(paste("Using the G-test, G value is equal to ", round(sum(gi),3)))
[1] "Using the G-test, G value is equal to  1.939"<Paste>
```

http://math.emory.edu/~lchen41/teaching/2020_Spring/EX_10-3-1.R

E.g. 3 Fisher's suspicion on Mendel's experiments on 1866:

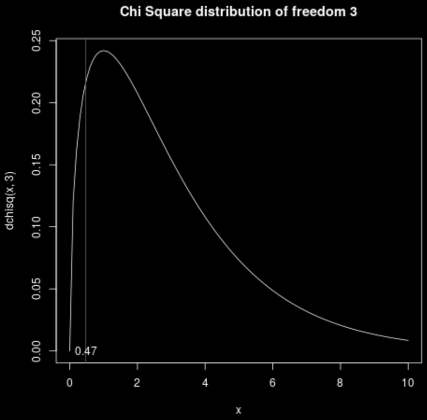
Table 10.3.7

Phenotype	Obs. Freq.	Mendel's Model	Exp. Freq.
(round, yellow)	315	9/16	312.75
(round, green)	108	3/16	104.25
(angular, yellow)	101	3/16	104.25
(angular, green)	32	1/16	34.75

$$d = ... = 0.47$$

$$P\text{-value} = \mathbb{P}(\chi_3^2 \leq 0.47) = 0.0746.$$

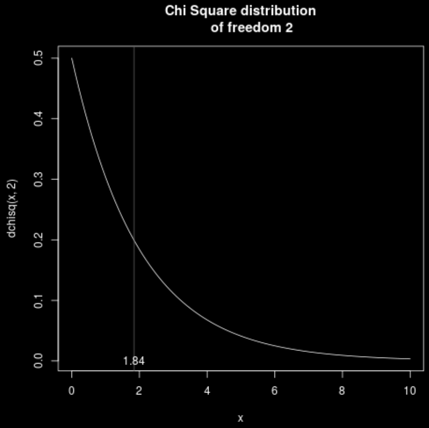
```
> # Case Study 10.3.3
> x=seq(0,10,0.1)
> plot(x,dchisq(x,3),type = "l")
> abline(v=0.47,col = "gray60")
> text(0.47,0,"0.47")
>  title ("Chi Square distribution
+          of freedom 3")
> pchisq(0.47,3)
 [1]  0.07456892
```



E.g. 2' A second look at the random generator in E.g. 2.

Does it fit the model too well? Find the *P*-value.

```
> # Example 10.3.1
> x=seq(0,10,0.1)
> plot(x,dchisq(x,2),type = "l")
> abline(v=1.84,col = "gray60")
> text(1.84,0,"1.84")
>  title ("Chi Square distribution
+           of freedom 2")
> pchisq(1.84,2)
[1]  0.601481
```



$$P\text{-value} = 0.601 \quad \Longrightarrow \quad \text{No.}$$