# Math 362: Mathematical Statistics II

Le Chen
le.chen@emory.edu

Emory University
Atlanta, GA

2021 Spring

# Chapter 10. Goodness-of-fit Tests

# Chapter 10. Goodness-of-fit Tests

Statistics is the grammar of science.

(Karl Pearson)

izquotes.com

1. Karl Pearson, 1857 – 1936.
2. English mathematician and biostatistician.
3. He has been credited with establishing the discipline of mathematical statistics
4. Method of moments; p-Value; Chi-square test; Foundations of statistical hypothesis testing theory; principle component analysis ...

# Pearson's chi-squared test in one shot
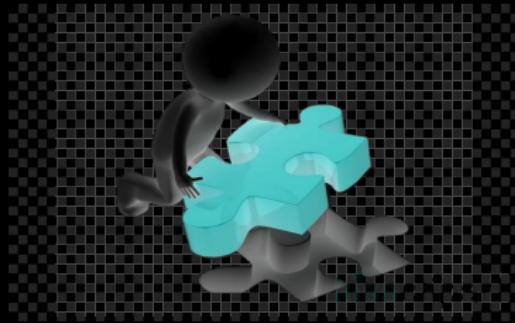


$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \sim \text{Chi Square of } df$$

$df =$ numer of classes $-$ number of estimated parameters $- 1$

All expected $\geq 5$

**Def.** Suppose one does an experiment of extracting $n$ balls of $t$ different colors from a jar, replacing the extracted ball after each draw. Balls from the same color are equivalent. Denote the variable which is the number of extracted balls of color $i$ ($i = 1, ..., t$) as $X_i$, and denote as $p_i$ the probability that a given extraction will be in color $i$. The probability distribution function of the vector $(X_1, \cdots, X_t)$ is called the multinomial distribution, which is equal to

$$p_{X_1, \cdots, X_t}(k_1, \cdots, k_t) = \mathbb{P}(X_1 = k_1, \cdots, X_t = k_t)$$
$$= \binom{n}{k_1, \cdots, k_t} p_1^{k_1} \cdots p_t^{k_t}$$

where $k_i \in \{0, 1, \cdots, n\}$, $1 \le i \le t$, $\sum_{i=1}^{t} k_i = n$, and $p_1 + \cdots + p_t = 1$.

**Thm** Suppose $(X_1, \cdots, X_t)$ follows the multinomial distribution with parameters $n$ and $(p_1, \cdots, p_t)$ with $p_i \geq 0$ and $\sum_i p_i = 1$. Then

1. $X_i \sim \mathrm{Binomail}(n, p_i)$ and hence

   $$\mathbb{E}[X_i] = np_i$$
   $$\mathrm{Var}(X_i) = np_i(1 - p_i)$$

2. $\mathrm{Cov}(X_i, X_j) = -np_ip_j, \ i \neq j.$            (negative correlated)

3. $M_{X_1, \cdots, X_t}(s_1, \cdots, s_t) = (p_1 e^{s_1} + \cdots + p_t e^{s_t})^n.$

Proof

(3)

$$M_{X_1,\cdots,X_t}(s_1,\cdots,s_t) = \mathbb{E}\left[e^{X_1 s_1 + \cdots + X_t s_t}\right]$$

$$= \sum_{\substack{k_1,\cdots,k_t=0 \\ k_1+\cdots+k_t=n}}^{n} \binom{n}{k_1,\cdots,k_t} p_1^{k_1} \cdots p_t^{k_t} e^{k_1 s_1 + \cdots + k_t s_t}$$

$$= \sum_{\substack{k_1,\cdots,k_t=0 \\ k_1+\cdots+k_t=n}}^{n} \binom{n}{k_1,\cdots,k_t} (p_1 e^{s_1})^{k_1} \cdots (p_t e^{s_t})^{k_t}$$

$$= (p_1 e^{s_1} + \cdots + p_t e^{s_t})^n$$

(1) To find $M_{X_i}(s_i)$, we simply set $s_j \equiv 0$ for $j \neq i$. Hence

$$M_{X_i}(s_i) = (\underbrace{p_1 + \cdots + p_{i-1} + p_{i+1} + \cdots + p_t}_{=1-p_i} + p_i e^{s_i})^n \implies X_i \sim \text{Binomial}(n, p_i)$$

**(2)** Set $M := M_{X_1, \cdots, X_t}(s_1, \cdots, s_t)$. Then for $i \neq j$,

$$\frac{\partial M}{\partial s_i} = n \left( p_1 e^{s_1} + \cdots + p_t e^{s_t} \right)^{n-1} p_i e^{s_i}$$

$$\frac{\partial^2 M}{\partial s_i \partial s_j} = n(n-1) \left( p_1 e^{s_1} + \cdots + p_t e^{s_t} \right)^{n-2} p_i e^{s_i} p_j e^{s_j}$$

$$\Downarrow$$

$$\mathbb{E}[X_i X_j] = \frac{\partial^2 M}{\partial s_i \partial s_j}\bigg|_{s_1 = \cdots = s_t = 0} = n(n-1)(p_1 + \cdots + p_t)^{n-2} p_i p_j = n(n-1) p_i p_j$$

$$\Downarrow$$

$$\begin{aligned} \mathrm{Cov}(X_i, X_j) &= \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j] \\ &= n(n-1) p_i p_j - n p_i \times n p_j \\ &= -n p_i p_j \end{aligned}$$

$\square$

From a continuous pdf to a multinomial distribution:

**E.g.** Let $Y_i$ be a random sample of size $n$ from $f_Y(y) = 6y(1-y)$, $y \in [0,1]$. Define

$$X_i = \begin{cases} 1 & Y_i \in [0, 0.25) \\ 2 & Y_i \in [0.25, 0.5) \\ 3 & Y_i \in [0.5, 0.75) \\ 4 & Y_i \in [0.75, 1) \end{cases}$$

Find the distribution of $(X_1, \cdots, X_n)$.

**Sol.** $(X_1, X_2, X_3, X_4)$ follows multinomial distribution with parameters $(p_1, p_2, p_3, p_4)$ where

$$p_1 = \int_0^{\frac{1}{4}} 6y(1-y)dy = \cdots = \frac{5}{32},$$
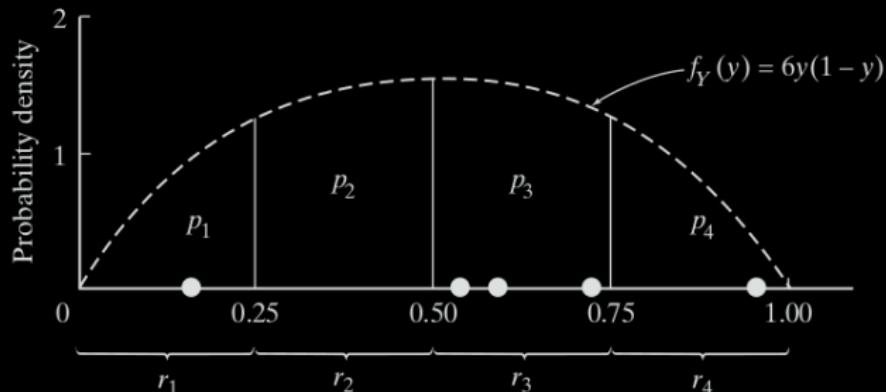
and by symmetry,

$$p_4 = p_1 = \frac{5}{32} \quad \text{and} \quad p_2 = p_3 = \frac{1}{2}\left(1 - p_1 - p_4\right) = \frac{11}{32}.$$

$\square$

Remark In this way, we transform the outcomes, any values between $[0, 1]$, into **categorical data**. This chapter is about

**Analysis of Categorical Data**

## 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*.

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

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

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

$$\|$$

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

$$\|$$

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

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

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

$$\Downarrow$$

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

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

**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 $\pi_i$ are

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

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

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

$$\Updownarrow$$

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

$$\Updownarrow$$

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

$\Rightarrow$

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

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

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

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

Def.

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

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

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

**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

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

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

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.

```
 1 │ > # EX 10.3.2
 2 │ > library(data.table)
 3 │ > mydat <- fread('http://math.emory.edu/~lchen41/teaching/2020_Spring/Case_
   │       10-3-2.data')
 4 │ trying URL 'http://math.emory.edu/~lchen41/teaching/2020_Spring/Case_10-3-2.
   │       data'
 5 │ Content type 'unknown' length 153 bytes
 6 │ ==================================================
 7 │ downloaded 153 bytes
 8 │
 9 │ > head(mydat)
10 │    Digit Oi    Pi
11 │ 1:     1 111 0.301
12 │ 2:     2  60 0.176
13 │ 3:     3  46 0.125
14 │ 4:     4  29 0.097
15 │ > pi = mydat[,3]
16 │ > oi = mydat[,2]
17 │ > n = sum(oi)
18 │ > ei = n*pi
19 │ > di = (ei-oi)^2/ei
20 │ > gi = 2*oi*log(oi/ei)
21 │ > print(paste("Using Pearson's test, D value is equal to ", round(sum(di),3)))
22 │ [1] "Using Pearson's test, D value is equal to 2.491"
23 │ > print(paste("Using the G-test, G value is equal to ", round(sum(gi),3)))
24 │ [1] "Using the G-test, G value is equal to 2.484"
```

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 → reduce to a set of classes:

**Table 10.3.5**

| Class | Observed Frequency, $k_i$ | $P_{i_o}$ | $40 p_{i_o}$ |
|---|---|---|---|
| $0 \le y < 0.20$ | 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, $k_i$ | $P_{i_o}$ | $40 p_{i_o}$ |
|---|---|---|---|
| $0 \le y < 0.40$ | 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: $\left( \chi^2_{.95,2}, \infty \right) = (5.992, \infty).$

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>
```

**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 |

$$d = ... = 0.47$$

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

**Chi Square distribution of freedom 3**

```
1 │ > # Case Study 10.3.3
2 │ > x=seq(0,10,0.1)
3 │ > plot(x,dchisq(x,3),type = "l")
4 │ > abline(v=0.47,col = "gray60")
5 │ > text(0.47,0,"0.47")
6 │ > title("Chi Square distribution
7 │ +        of freedom 3")
8 │ > pchisq(0.47,3)
9 │ [1] 0.07456892
```
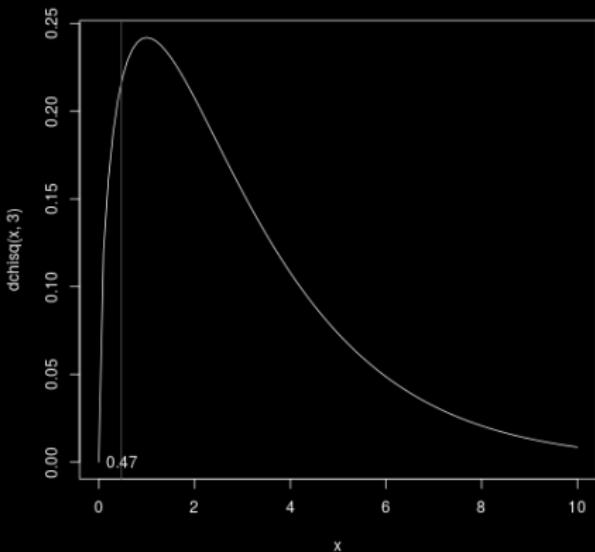
**E.g. 2'** A second look at the random generator in E.g. 2.

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

**Chi Square distribution of freedom 2**

```
1 > # Example 10.3.1
2 > x=seq(0,10,0.1)
3 > plot(x,dchisq(x,2),type = "l")
4 > abline(v=1.84,col = "gray60")
5 > text(1.84,0,"1.84")
6 > title("Chi Square distribution
7 +      of freedom 2")
8 > pchisq(1.84,2)
9 [1] 0.601481
```
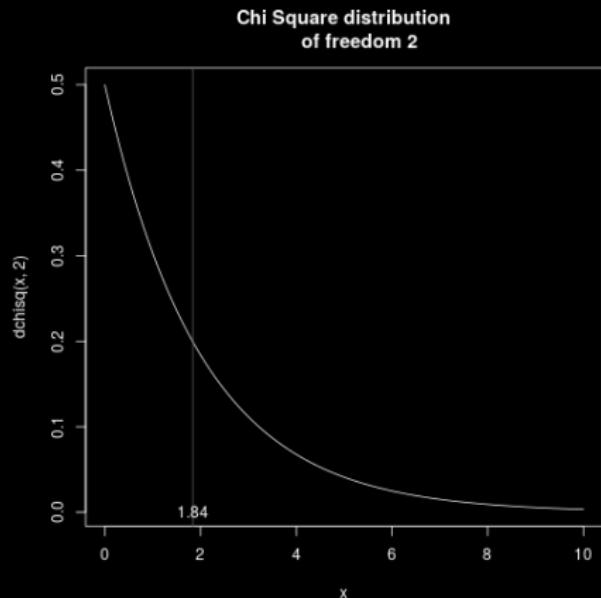


*P*-value $= 0.601$  $\implies$  No.

# Chapter 10. Goodness-of-fit Tests

| $p_i$ are known | $p_i$ are unknown |
|---|---|
| $D = \sum_{i=1}^{t} \frac{(X_i - np_i)^2}{np_i}$ | $D_1 = \sum_{i=1}^{t} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i}$ |
| $\chi^2$ with f.d. $t-1$ | $\chi^2$ with f.d. $t-1-s$ |
| $d = \sum_{i=1}^{t} \frac{(k_i - np_{i0})^2}{np_{i0}}$ | $d_1 = \sum_{i=1}^{t} \frac{(k_i - n\hat{p}_{i0})^2}{n\hat{p}_{i0}}$ |
| $np_{i0} \geq 5$ | $n\hat{p}_{i0} \geq 5$ |
| $d > \chi^2_{1-\alpha, t-1}$ | $d_1 > \chi^2_{1-\alpha, t-1-s}$ |

† $s$ is the number of unknown parameters.

   df = <u>number of classes</u> − 1 − <u>number of unknown parameters.</u>

**E.g. 1** Binomial data: 4096 students, each shots basketball 4 times. Let $X_i$ be the number of hits for the $i$th student.



| Number of Hits, $i$ | Obs. Freq., $k_i$ |
|---|---|
| $r_i's$ $\{$ 0 | 1280 |
| 1 | 1717 |
| 2 | 915 |
| 3 | 167 |
| 4 | 17 |

People believe that $X_i$ should following binomial$(4, p)$, that is, shotting basketball should be something like trying to get red chocolate beans from a jar of beans of two colors.

Find the MLE for $p$. Use the data to make a conclusion.

**Sol.** 1) $H_0 : X_i \sim \text{binomal}(4, p)$.

2) Under $H_0$, the MLE for $p$ is $p_e = ... = 0.251$

3) Compute the expected frequenies:

**Table 10.4.1**

| Number of Hits, $i$ | Obs. Freq., $k_i$ | Estimated Exp. Freq., $n\hat{p}_{i_o}$ |
|---|---|---|
| $r_i's$ $\begin{cases} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{cases}$ | 1280 <br> 1717 <br> 915 <br> 167 <br> 17 | 1289.1 <br> 1728.0 <br> 868.6 <br> 194.0 <br> 16.3 |

$$\implies \quad d_1 = \cdots = 6.401.$$

4) Critical region: $(\chi^2_{.95,5-1-1}, +\infty) = (7.815, +\infty)$

5) Conclusion: Fail to reject.

6) Alternatively, $P$-value $= \mathbb{P}(\chi^2_3 \geq 6.401) = 0.094,$ ... discuss... $\qquad\square$

**E.g. 2** Does the number of death per day follow the Poisson distribution?

| Number of Deaths, $i$ | Obs. Freq., $k_i$ |
|---|---|
| 0 | 162 |
| 1 | 267 |
| 2 | 271 |
| 3 | 185 |
| 4 | 111 |
| 5 | 61 |
| 6 | 27 |
| 7 | 8 |
| 8 | 3 |
| 9 | 1 |
| 10+ | 0 |
| | 1096 |

**Sol.** 1) Let $X_i$ be the number of death in $i$th day, $1 \le i \le 1096$.

2) $H_0 : X_i$ follow Poisson($\lambda$).

3) The MLE for $\lambda$ is: $\lambda_e = \cdots = 2.157$.

4) Compute the expected frequencies:

**Table 10.4.2**

| Number of Deaths, $i$ | Obs. Freq., $k_i$ | Est. Exp. Freq., $n\hat{p}_{i_o}$ |
|---|---|---|
| 0 | 162 | 126.8 |
| 1 | 267 | 273.5 |
| 2 | 271 | 294.9 |
| 3 | 185 | 212.1 |
| 4 | 111 | 114.3 |
| 5 | 61 | 49.3 |
| 6 | 27 | 17.8 |
| 7 | 8 | 5.5 |
| 8 | 3 | 1.4 |
| 9 | 1 | 0.3 |
| 10+ | 0 | 0.1 |
| | 1096 | 1096 |

**Table 10.4.3**

| | Number of Deaths, $i$ | Obs. Freq., $k_i$ | Est. Exp. Freq., $n\hat{p}_{i_o}$ |
|---|---|---|---|
| | 0 | 162 | 126.8 |
| | 1 | 267 | 273.5 |
| | 2 | 271 | 294.9 |
| $r_1, r_2, \ldots, r_8$ | 3 | 185 | 212.1 |
| | 4 | 111 | 114.3 |
| | 5 | 61 | 49.3 |
| | 6 | 27 | 17.8 |
| | 7+ | 12 | 7.3 |
| | | 1096 | 1096 |

$$\implies \quad d_1 = \cdots = 25.98.$$

5) $P$-value $= \mathbb{P}(\chi^2_{1,8-1-1} \ge 25.98) = 0.00022.$ Reject! $\square$

**E.g. 1** Whether are the two ratings independent?

**Table 10.5.5**

|  |  | Ebert Ratings | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | Down | Sideways | Up | Total |
|  | Down | 24 | 8 | 13 | 45 |
| Siskel | Sideways | 8 | 13 | 11 | 32 |
| Ratings | Up | 10 | 9 | 64 | 83 |
|  | Total | 42 | 30 | 88 | 160 |

**E.g. 2** Whether is the suicide rate independent of the mobility factor?

**Table 10.5.7**

| City | Suicides per 100,000, $x_i$ | Mobility Index, $y_i$ | City | Suicides per 100,000, $x_i$ | Mobility Index, $y_i$ |
|------|------|------|------|------|------|
| New York | 19.3 | 54.3 | Washington | 22.5 | 37.1 |
| Chicago | 17.0 | 51.5 | Minneapolis | 23.8 | 56.3 |
| Philadelphia | 17.5 | 64.6 | New Orleans | 17.2 | 82.9 |
| Detroit | 16.5 | 42.5 | Cincinnati | 23.9 | 62.2 |
| Los Angeles | 23.8 | 20.3 | Newark | 21.4 | 51.9 |
| Cleveland | 20.1 | 52.2 | Kansas City | 24.5 | 49.4 |
| St. Louis | 24.8 | 62.4 | Seattle | 31.7 | 30.7 |
| Baltimore | 18.0 | 72.0 | Indianapolis | 21.0 | 66.1 |
| Boston | 14.8 | 59.4 | Rochester | 17.2 | 68.0 |
| Pittsburgh | 14.9 | 70.0 | Jersey City | 10.1 | 56.5 |
| San Francisco | 40.0 | 43.8 | Louisville | 16.6 | 78.7 |
| Milwaukee | 19.3 | 66.2 | Portland | 29.3 | 33.2 |
| Buffalo | 13.8 | 67.6 | | | |

$$\bar{x} = 20.8 \quad \text{and} \quad \bar{y} = 56.0$$

**Table 10.5.8**

| | | Mobility Index | |
|------|------|------|------|
| | | Low (<56.0) | High (≥56.0) |
| Suicide | High (≥20.8) | 7 | 4 |
| Rate | Low (<20.8) | 3 | 11 |

**Thm 10.4.1** Suppose that *n* observations are taken on a sample space partitioned by the events $A_1, \cdots, A_r$ and $B_1, \cdots, B_c$.

Let $p_i = \mathbb{P}(A_i)$, $q_j = \mathbb{P}(B_j)$, $p_{ij} = \mathbb{P}(A_i \cap B_j)$.

Let $X_{ij}$ be the number of observations belonging to $A_i \cap B_j$.

**a)** Provided that $np_{ij} \geq 5$ for all $i, j$, the r.v.

$$D_2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(X_{ij} - np_{ij})^2}{np_{ij}} \sim \text{Chi square of f.d. } rc - 1$$

**b)** To test $H_0 : A_i$'s are independent of $B_i$'s, calculate the test statistic

$$d_2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(k_{ij} - n\hat{p}_i \hat{q}_j)^2}{n\hat{p}_i \hat{q}_j}$$

where $\hat{p}_i$ and $\hat{q}_j$ are MLE's for $p_i$ and $q_j$, respectively.

Provided that $n\hat{p}_i \hat{q}_j \geq 5$ for all $i, j$, the critical region is

$$(\chi^2_{1-\alpha, (r-1)(c-1)}, +\infty)$$

**E.g. 1** Sol: Compute the expected frequencies:

| Table 10.5.6 | | Ebert Ratings | | | |
|---|---|---|---|---|---|
| | | Down | Sideways | Up | Total |
| Siskel Ratings | Down | 24 (11.8) | 8 (8.4) | 13 (24.8) | 45 |
| | Sideways | 8 (8.4) | 13 (6.0) | 11 (17.6) | 32 |
| | Up | 10 (21.8) | 9 (15.6) | 64 (45.6) | 83 |
| | Total | 42 | 30 | 88 | 160 |

$$\implies \quad d_2 = \cdots = 45.37$$

Critical region is

$$\left(\chi^2_{0.99,(3-1)\times(3-1)}, +\infty\right) = (13.277, +\infty)$$

Alternatively $P$-value $= \mathbb{P}(\chi^2_4 \geq 45.37) = 3.33 \times 10^{-9}$.

Rejection at $\alpha = 0.01$. $\qquad \qquad \square$

**E.g. 2** Sol: Compute the expected frequencies:

**Table 10.5.9**

|  |  | Mobility Index | |
| --- | --- | --- | --- |
|  |  | Low (<56.0) | High (≥56.0) |
| Suicide | High (≥20.8) | 4.4* | 6.6 |
| Rate | Low (<20.8) | 5.6 | 8.4 |

*$\hat{E}(X_{11}) = 4.4$ does not quite satisfy the "$n\hat{p}_i\hat{q}_j \geq 5$" restriction stated in Theorem 10.5.1, but 4.4 is close enough to 5 to maintain the integrity of the $\chi^2$ approximation.

$$\implies \quad d_2 = \cdots = 4.57$$

Critical region is

$$\left(\chi^2_{0.95,(2-1)\times(2-1)}, +\infty\right) = (3.41, +\infty)$$

Alternatively $P$-value $= \mathbb{P}(\chi^2_1 \geq 4.57) = 0.033$

Rejection at $\alpha = 0.05$. $\qquad\square$