Math 362: Mathmatical Statistics II

Le Chen
le.chen@emory.edu

Emory University
Atlanta, GA

Last updated on March 24, 2020

2020 Spring

Chapter 10. Goodness-of-fit Tests

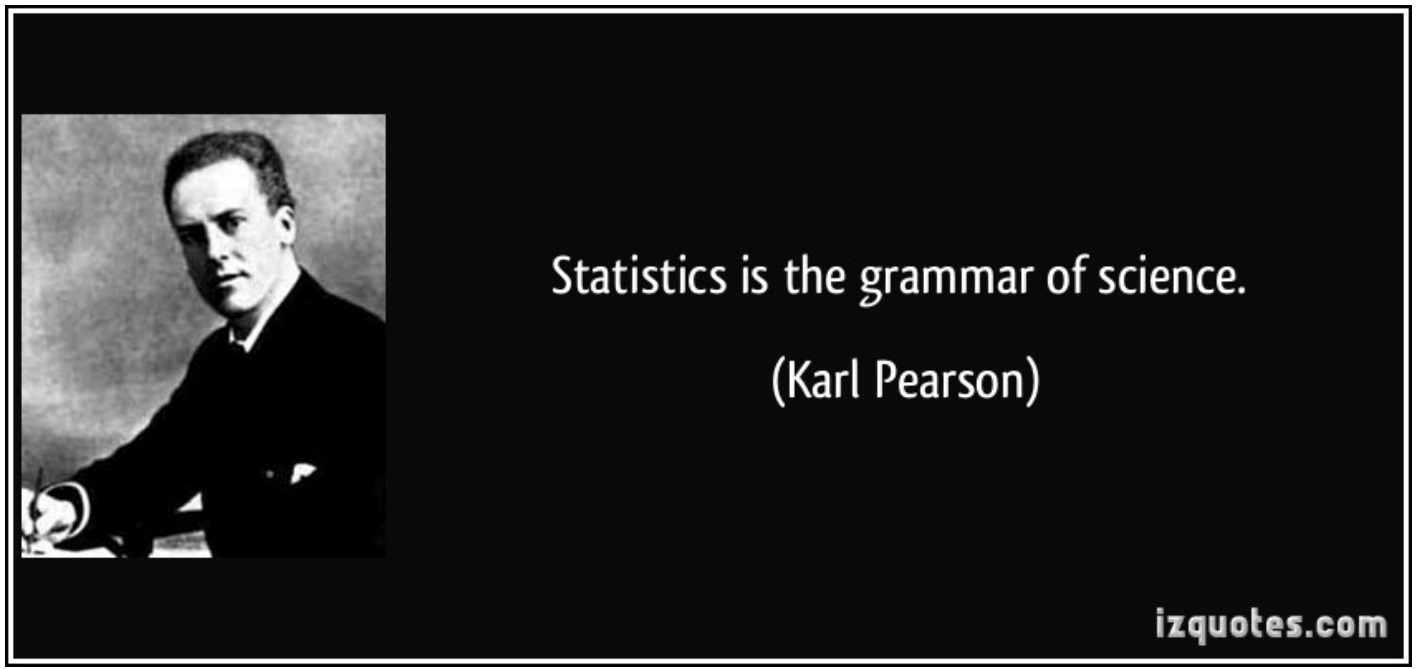
Chapter 10. Goodness-of-fit Tests

§ 10.1 Introduction

§ 10.2 The Multinomial Distribution

§ 10.3 Goodness-of-Fit Tests: All Parameters Known

§ 10.1 Introduction



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

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

$$df = \text{numer of classes} - \text{number of estimated parameters} - 1$$

$$\text{All expected} \geq 5$$

Chapter 10. Goodness-of-fit Tests

§ 10.1 Introduction

§ 10.2 The Multinomial Distribution

§ 10.3 Goodness-of-Fit Tests: All Parameters Known

§ 10.2 The Multinomial Distribution

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

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

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

Properties of multinomial distribution

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

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

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

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

$$
\begin{aligned}
M_{X_1,\cdots,X_t}(s_1,\cdots,s_t) &= \mathbb{E}\left[e^{X_1 s_1+\cdots+X_t s_t}\right] \\
&= \sum_{\substack{k_1,\cdots,k_t=0 \\ k_1+\cdots+k_t=n}}^{n} \binom{n}{k_1,\cdots,k_t} p_1^{k_1}\cdots p_t^{k_t} e^{X_1 s_1+\cdots+X_t s_t} \\
&= \sum_{\substack{k_1,\cdots,k_t=0 \\ k_1+\cdots+k_t=n}}^{n} \binom{n}{k_1,\cdots,k_t} (p_1 e^{s_1})^{k_1}\cdots(p_t e^{s_t})^{k_t} \\
&= \left(p_1 e^{s_1}+\cdots+p_t e^{s_t}\right)^n
\end{aligned}
$$

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

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

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

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

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

$$\Downarrow$$

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

$$\Downarrow$$

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

□

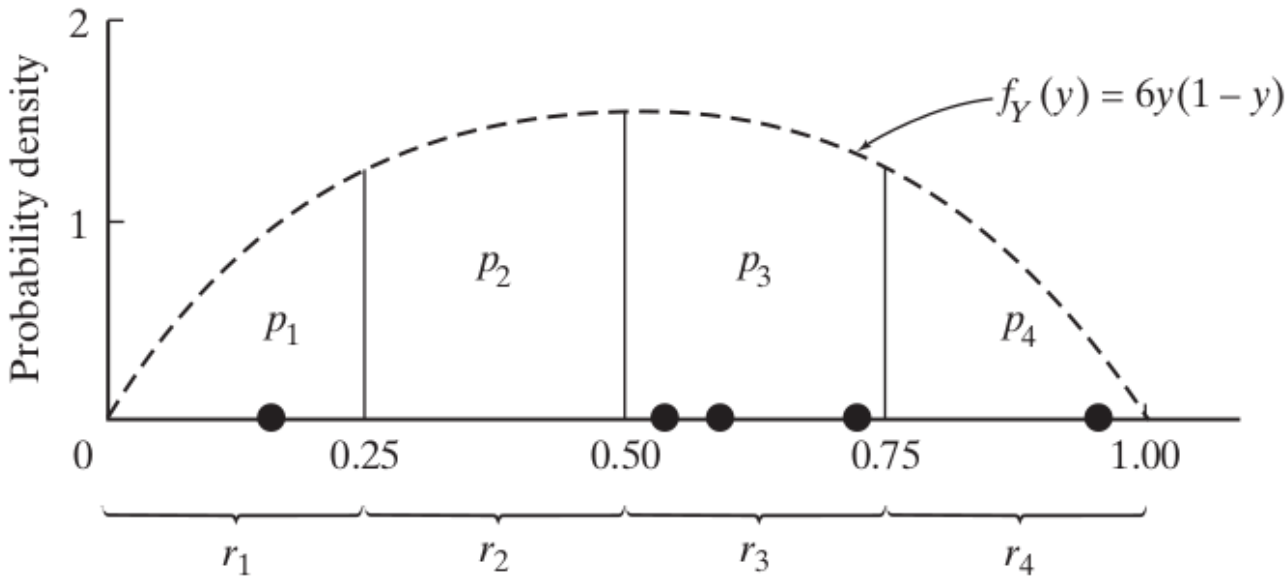
From a continuous pdf to a multinomial distribution:

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. ... □



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

Analysis of Categorical Data

Chapter 10. Goodness-of-fit Tests

§ 10.1 Introduction

§ 10.2 The Multinomial Distribution

§ 10.3 Goodness-of-Fit Tests: All Parameters Known

§ 10.3 Goodness-of-Fit Tests: All Parameters Known

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 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.,

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

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 $n \to \infty$,

$$\frac{O_1 - n\pi_1}{\sqrt{n\pi_1(1-\pi_1)}} \xrightarrow{d} N(0,1) \quad \Longrightarrow \quad \frac{(O_1 - n\pi_1)^2}{n\pi_1(1-\pi_1)} \xrightarrow{d} \chi^2_1$$

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

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

Hence,

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

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.

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

For approximation accuracy, one should require that $n\pi_i \geq 5$ for all i.

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: $\lambda \in (0, \lambda_*)$.

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,

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

$\therefore$ Pearson's Chi-square test is an approximation of Pearson's χ^2 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

$$H_0 : p_1 = p_{10}, \cdots, p_9 = p_{90}$$

$$v.s.$$

$$H_1 : p_i \neq p_{i0} \quad \text{for at least one } i = 1, \cdots, 9.$$

Table 10.3.2

Digit	Observed, k_i	Benford p_{i_o}	Expected $(= 355 \cdot p_{i_o})$	$(k_i - 355p_{i_o})^2/355p_{i_o}$
1	111	0.301	106.9	0.16
2	60	0.176	62.5	0.10
3	46	0.125	44.4	0.06
4	29	0.097	34.4	0.86
5	26	0.079	28.0	0.15
6	22	0.067	23.8	0.13
7	21	0.058	20.6	0.01
8	20	0.051	18.1	0.20
9	20	0.046	16.3	0.82
	355	1.000	355.0	2.49

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

Conclusion: Fail to reject.

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:

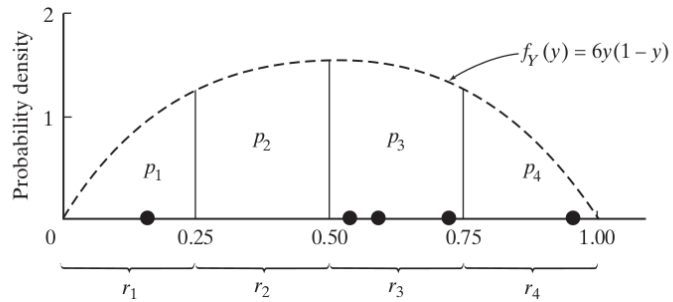


Table 10.3.5

Class	Observed Frequency, k_i	P_{i_o}	$40\,p_{i_o}$
$0 \le y < 0.20$	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.

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

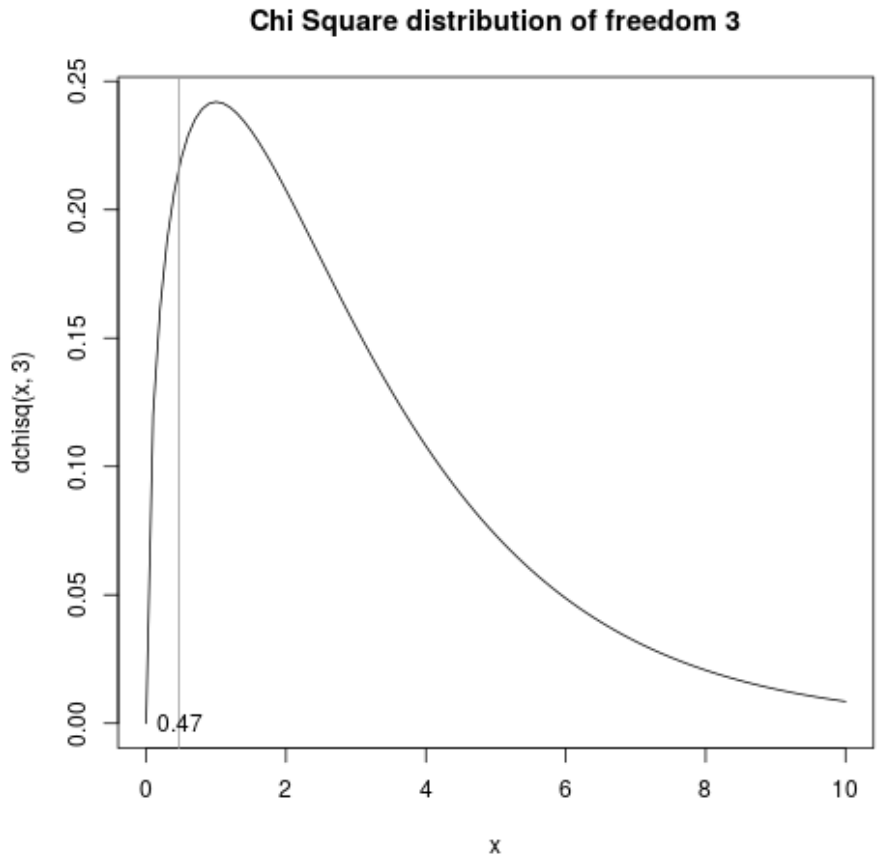
Table 10.3.7

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

$$d = \ldots = 0.47$$

$$P\text{-value} = \mathbb{P}(\chi^2_3 \leq 0.47) = 0.0746.$$

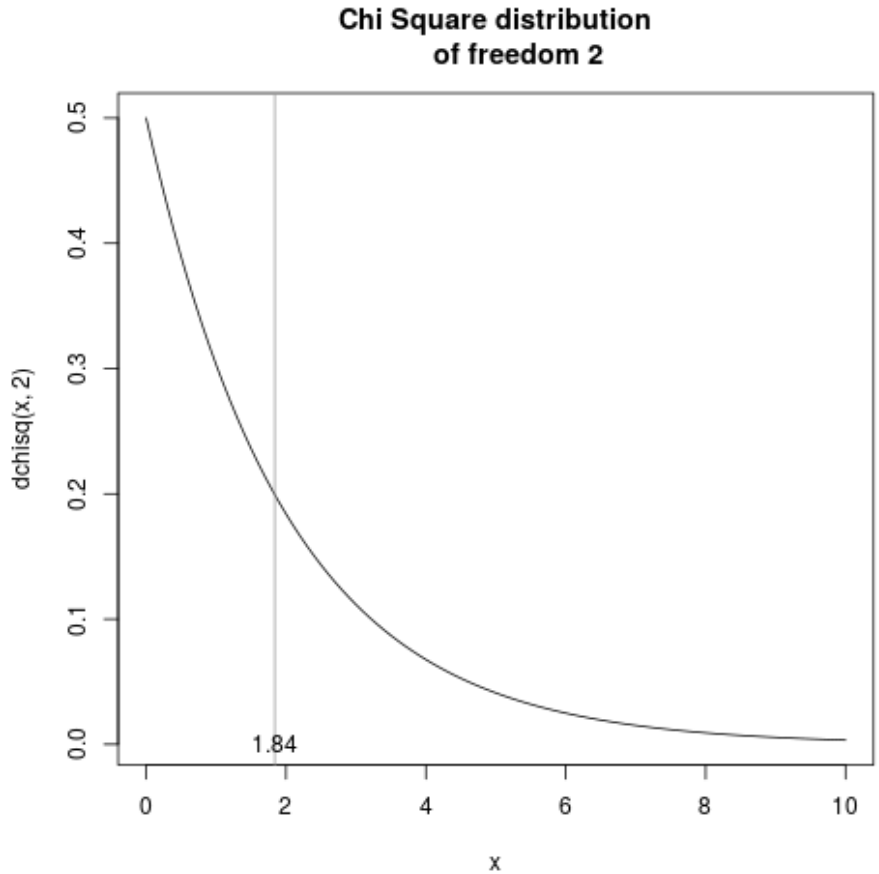
```
> # Case Study 10.3.3
> x=seq(0,10,0.1)
> plot(x,dchisq(x,3),type = "l")
> abline(v=0.47,col = "gray60")
> text(0.47,0,"0.47")
>  title ("Chi Square distribution
+          of freedom 3")
> pchisq(0.47,3)
 [1]  0.07456892
```



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-value $= 0.601 \implies$ No.