University of Washington

STATISTICS

Elements of Statistical Methods
Goodness-of-Fit (Ch 13)

Fritz Scholz

Spring Quarter 2010

April 9, 2010

Partitions of Sample Spaces

So far our inference problems concerned means or medians of populations.

Now we will focus on inference concerning sample space probabilities.

Let $E_1, \ldots, E_k \subset S$ be mutually exclusive events such that their union is S.

Such a collection of sets, $E_1, \ldots, E_k$, is called a partition of S.

Example 1: A single die is rolled. $S = \{1,2,3,4,5,6\}$.

Let $k = 6$ and $E_i = \{i\}, i = 1,2,\ldots,6$.

Example 2: Let X be a discrete random variable with $X(S) = \{0,1,2,3,\ldots\}$.

Let $k = 5$ and $E_i = \{i-1\}$ for $i = 1,2,3,4$ and $E_5 = \{4,5,6,\ldots\}$.

Example 3: Let X be a continuous random variable with $X(S) = (-\infty, \infty)$.

Let $k = 7$ with $E_1 = (-\infty, -5)$, $E_2 = [-5, -3)$, $E_3 = [-3, -1)$, $E_4 = [-1, 1)$, $E_5 = [1, 3)$, $E_6 = [3, 5)$, $E_7 = [5, \infty)$.

Testing Hypotheses

Given a partition of S, our interest centers on the cell probabilities

$p_1 = P(E_1), \ldots, p_k = P(E_k)$ with $\vec{p} = (p_1, \ldots, p_k) \in \Pi$, where

$$\Pi = \left\{ (\pi_1, \ldots, \pi_k) : \pi_i \geq 0, i = 1, \ldots, k, \text{ and } \sum_{i=1}^{k} \pi_i = 1 \right\} \subset R^k$$

The generic testing problem consists of partitioning $\Pi = \Pi_0 \cup \Pi_1$ with $\Pi_0 \cap \Pi_1 = \emptyset$ and then testing $H_0 : \vec{p} \in \Pi_0$ against $H_1 : \vec{p} \in \Pi_1$.

Example: When rolling a die we can test the fairness of the die by specifying $k = 6$

$$\Pi_0 = \left\{ \left(\frac{1}{6}, \frac{1}{6}, \ldots, \frac{1}{6} \right) \right\} \subset \Pi \subset R^6 \quad \text{and} \quad \Pi_1 = \{ \vec{\pi} \in \Pi : \vec{\pi} \notin \Pi_0 \}$$

Example

In 1882, R.Wolf (with time on his hands) tossed a die $n = 20000$ times, observing

j	1	2	3	4	5	6
o_j	3407	3631	3176	2916	3448	3422

One way to become "famous." Was Wolf tossing a fair die?

For each cell the expected count is $e_j = np_j = 20000/6 = 3333\frac{1}{3}$.

Are the discrepancies explainable by pure chance, even with a fair die?

There are many ways to measure discrepancies between the o_j and e_j, $j = 1, \ldots, 6$.

best known is Pearson's chi-squared statistic $$X^2 = \sum_{j=1}^{k} \frac{(o_j - e_j)^2}{e_j}$$

Maximum Likelihood Estimates (MLEs)

Given the cell probabilities $p_1, \ldots, p_k$ we can ask:

What is the probability or likelihood of the observed counts $o_1, \ldots, o_k$. It is

$$L(p_1, \ldots, p_k) = P(O_1 = o_1, \ldots, O_k = o_k) = C p_1^{o_1} \cdots p_k^{o_k}$$

where C counts the number of ways of how the cell occurrences in the n trials can result in counts of $o_1, \ldots, o_k$. $C = n!/(o_1! \cdots o_k!)$, similar to binomial case.

The maximum likelihood estimates (MLEs) of $p_1, \ldots, p_k$ are those values $p_1, \ldots, p_k$ that maximize $L(p_1, \ldots, p_k)$ for the given observed counts $o_1, \ldots, o_k$.

These estimates are those values of $p_1, \ldots, p_k$ that would make most probable what we observed. This is a powerful and useful estimation principle.

Without further restrictions beyond $p_i \geq 0$ and $p_1 + \ldots + p_k = 1$

the MLEs of $p_1, \ldots, p_k$ are $\hat{p}_1 = o_1/n, \ldots, \hat{p}_k = o_k/n$ (the plug-in estimates).

Likelihood Ratio Discrepancy Measure

The maximum value of the likelihood thus is

$$L(\hat{p}_1,\ldots,\hat{p}_k) = C\hat{p}_1^{o_1}\cdots\hat{p}_k^{o_k} = C\left(\frac{o_1}{n}\right)^{o_1}\ldots\left(\frac{o_k}{n}\right)^{o_k}$$

Under our previous null hypothesis we have $H_0 : p_1 = \ldots = p_k = 1/6$.

Under that restriction, the MLEs of $p_1,\ldots,p_k$ are $\check{p}_1 = 1/6,\ldots,\check{p}_k = 1/6$

with likelihood $$L(\check{p}_1,\ldots,\check{p}_k) = C\,\check{p}_1^{o_1}\cdots\check{p}_k^{o_k} = C\left(\frac{1}{6}\right)^{o_1}\ldots\left(\frac{1}{6}\right)^{o_k}$$

$$\Longrightarrow \qquad L(\check{p}_1,\ldots,\check{p}_k) \le L(\hat{p}_1,\ldots,\hat{p}_k) \quad \text{or} \quad \lambda = \frac{L(\check{p}_1,\ldots,\check{p}_k)}{L(\hat{p}_1,\ldots,\hat{p}_k)} \in [0,1]$$

since $(\check{p}_1,\ldots,\check{p}_k)$ maximizes over the much more restricted set $\Pi_0 = \left\{\left(\frac{1}{6},\frac{1}{6},\ldots,\frac{1}{6}\right)\right\}$.

If $L(\check{p}_1,\ldots,\check{p}_k) \approx L(\hat{p}_1,\ldots,\hat{p}_k)$ or $\lambda \approx 1$, then H_0 would be plausible, because $(\check{p}_1,\ldots,\check{p}_k) = (1/6,\ldots,1/6)$ is almost as good as $(\hat{p}_1,\ldots,\hat{p}_k)$ in giving highest probability to $o_1,\ldots,o_k$. A small λ is evidence against H_0.

A Second Null Hypothesis

Consider the hypothesis that opposing faces on the die have same probability,

then $p_1 = p_6, p_2 = p_5, p_3 = p_4$ and our null hypothesis takes the form

$H_0 : \vec{p} \in \Pi_0 = \{(p_1, p_2, p_3, p_3, p_2, p_1) : 2p_1 + 2p_2 + 2p_3 = 2(p_1 + p_2 + p_3) = 1\}$

Using calculus, the maximum likelihood estimates restricted to this H_0 are

$$\check{p}_1 = \check{p}_6 = \frac{(o_1 + o_6)/2}{n}, \quad \check{p}_2 = \check{p}_5 = \frac{(o_2 + o_5)/2}{n}, \quad \check{p}_3 = \check{p}_4 = \frac{(o_3 + o_4)/2}{n}$$

i.e., under H_0 the estimates of $2p_1 = p_1 + p_6$, $2p_2 = p_2 + p_5$ and $2p_3 = p_3 + p_4$

are again just the plug-in estimates

$$2\check{p}_1 = \frac{o_1 + o_6}{n}, \quad 2\check{p}_2 = \frac{o_2 + o_5}{n}, \quad 2\check{p}_3 = \frac{o_3 + o_4}{n}$$

$\lambda = L(\check{p}_1, \ldots, \check{p}_k)/L(\hat{p}_1, \ldots, \hat{p}_k) \approx 1$ again supports the current hypothesis H_0

(equal opposing face probabilities), for the same reason as before.

Small λ would present evidence against that null hypothesis.

A Third Null Hypothesis

To round out the possible hypotheses we also consider

$H_0 : p_1 + p_6 = p_2 + p_5 = p_3 + p_4 = 1/3,$

i.e., the combined probabilities of the three opposing face pairs are the same.

Under H_0 we can view p_1 and p_6 as two-stage probabilities, namely

$$\begin{aligned} p_1 &= P_{H_0}(\{1\}) = P_{H_0}(\{1\} \cap (\{1\} \cup \{6\})) = P_{H_0}(\{1\}|\{1\} \cup \{6\}) \cdot P_{H_0}(\{1\} \cup \{6\}) \\ &= P(\{1\}|\{1\} \cup \{6\}) \cdot P_{H_0}(\{1\} \cup \{6\}) = \frac{P(\{1\} \cap (\{1\} \cup \{6\}))}{P(\{1\} \cup \{6\})} \cdot \frac{1}{3} = \frac{p_1}{p_1 + p_6} \cdot \frac{1}{3} \end{aligned}$$

$$p_6 = \frac{p_6}{p_1 + p_6} \cdot \frac{1}{3} = \frac{1}{3} - p_1, \quad p_2 = \frac{p_2}{p_2 + p_5} \cdot \frac{1}{3}, \quad p_5 = \frac{p_5}{p_2 + p_5} \cdot \frac{1}{3}$$

$$p_3 = \frac{p_3}{p_3 + p_4} \cdot \frac{1}{3}, \quad p_4 = \frac{p_4}{p_3 + p_4} \cdot \frac{1}{3}$$

Within each pair (p_1, p_6) (p_2, p_5) and (p_3, p_4) only one can be freely chosen under H_0, since within each pair they have to add to $1/3$.

MLEs Under Third Null Hypothesis

With calculus one can again find $\vec{p} = (p_1, \ldots, p_6)$ that maximizes $L(p_1, \ldots, p_6)$

subject to $\vec{p} \in \Pi_0 = \{(p_1, \ldots, p_6) : p_1 + p_6 = p_2 + p_5 = p_3 + p_4 = 1/3\}$

That maximizing $\check{\vec{p}} = (\check{p}_1, \ldots, \check{p}_6)$ is given by

$$\check{p}_1 = \frac{\hat{p}_1}{\hat{p}_1 + \hat{p}_6} \cdot \frac{1}{3} = \frac{o_1/n}{o_1/n + o_6/n} \cdot \frac{1}{3} = \frac{o_1}{o_1 + o_6} \cdot \frac{1}{3}, \quad \check{p}_6 = \frac{o_6}{o_1 + o_6} \cdot \frac{1}{3}$$

$$\check{p}_2 = \frac{o_2}{o_2 + o_5} \cdot \frac{1}{3}, \quad \check{p}_5 = \frac{o_5}{o_2 + o_5} \cdot \frac{1}{3}, \quad \check{p}_3 = \frac{o_3}{o_3 + o_4} \cdot \frac{1}{3}, \quad \check{p}_4 = \frac{o_4}{o_3 + o_4} \cdot \frac{1}{3}$$

basically using the plug-in estimates $\hat{p}_i$ in the H_0 representation of p_i.

$\lambda = L(\check{p}_1, \ldots, \check{p}_k)/L(\hat{p}_1, \ldots, \hat{p}_k) \approx 1$ again supports the current hypothesis H_0

(equal opposing face probabilities), for the same reason as before.

Small λ would present evidence against that null hypothesis.

The Likelihood Ratio Chi-Squared Statistic

Let $\check{e}_j = n\check{p}_j$ be the expected cell count when $p_j = \check{p}_j$, as estimated under H_0.

$$\lambda = \frac{L(\check{p}_1,\ldots,\check{p}_k)}{L(\hat{p}_1,\ldots,\hat{p}_k)} = \frac{C\ \check{p}_1^{o_1}\cdots\check{p}_k^{o_k}}{C\ \left(\frac{o_1}{n}\right)^{o_1}\cdots\left(\frac{o_k}{n}\right)^{o_k}} = \frac{C\ \left(\frac{\check{e}_1}{n}\right)^{o_1}\cdots\left(\frac{\check{e}_k}{n}\right)^{o_k}}{C\ \left(\frac{o_1}{n}\right)^{o_1}\cdots\left(\frac{o_k}{n}\right)^{o_k}} = \left(\frac{\check{e}_1}{o_1}\right)^{o_1}\cdots\left(\frac{\check{e}_k}{o_k}\right)^{o_k}$$

$$G^2 = -2\log\lambda = 2\sum_{j=1}^{k} o_j \log(o_j/\check{e}_j)$$

Since $\lambda \in [0,1]$, we have $G^2 \geq 0$. Large values of G_2 are evidence against H_0.

The null distribution of G^2 can usually be well approximated by a chi-squared distribution with appropriate degrees of freedom.

Under H_0 both Pearson's X^2 and the G^2 statistic are fairly close to each other and the approximate null distribution applies to X^2 as well.

Appropriate Degrees of Freedom

The appropriate degrees of freedom for the approximating chi-squared distribution is obtained as the difference of the full dimension of Π, i.e., $k-1$, and the dimension of the space in which Π_0 is embedded.

In our first example, where $\Pi_0 = \left\{ \left(\frac{1}{6}, \frac{1}{6}, \ldots, \frac{1}{6} \right) \right\}$, that dimension is zero, so the degrees of freedom for the approximating chi-squared distribution are $(6-1)-0 = 5$.

In the second example the dimension of Π_0 is 2. Of the parameters p_1, p_2, p_3 only 2 can vary freely, since $p_1 + p_2 + p_3 = 1/2$. The degrees of freedom for the approximating chi-squared distribution are $(6-1)-2 = 3$.

In the third example the dimension of Π_0 is 3, as alluded to previuosly. The degrees of freedom for the approximating chi-squared distribution are $(6-1)-3 = 2$.

Analysis of the Wolf Dice Data

Testing $H_0 : p_1 = \ldots = p_6 = 1/6$ we find $\check{e}_j = 20000\check{p}_j = 20000/6$

$$G^2 = 2\sum_{j=1}^{6} o_j \log(o_j/\check{e}_j) = 95.8023 \quad \text{and} \quad X^2 = \sum_{j=1}^{6}(o_j - \check{e}_j)^2/\check{e}_j = 94.189$$

`1-pchisq(95.8023,df=5)=0` and `1-pchisq(94.189,df=5)=0`,

the evidence against H_0 (the die is fair) is overwhelming.

Testing $H_0 : p_1 = p_6,\ p_2 = p_5,\ p_3 = p_4$ we find with $\check{e}_j = n\check{p}_j$

$$\begin{aligned} \check{e}_1 = \check{e}_6 &= (3407 + 3422)/2 = 3414.5 \\ \check{e}_2 = \check{e}_5 &= (3631 + 3448)/2 = 3539.5 \\ \check{e}_3 = \check{e}_4 &= (3176 + 2916)/2 = 3046.0 \end{aligned}$$

and obtain $G^2 = 15.8641$ and $X^2 = 15.1971$ with respective p-values

`1-pchisq(15.8641,df=3)=.00121` and `1-pchisq(15.1971,df=3)=.00166`.

Analysis of the Wolf Dice Data (continued)

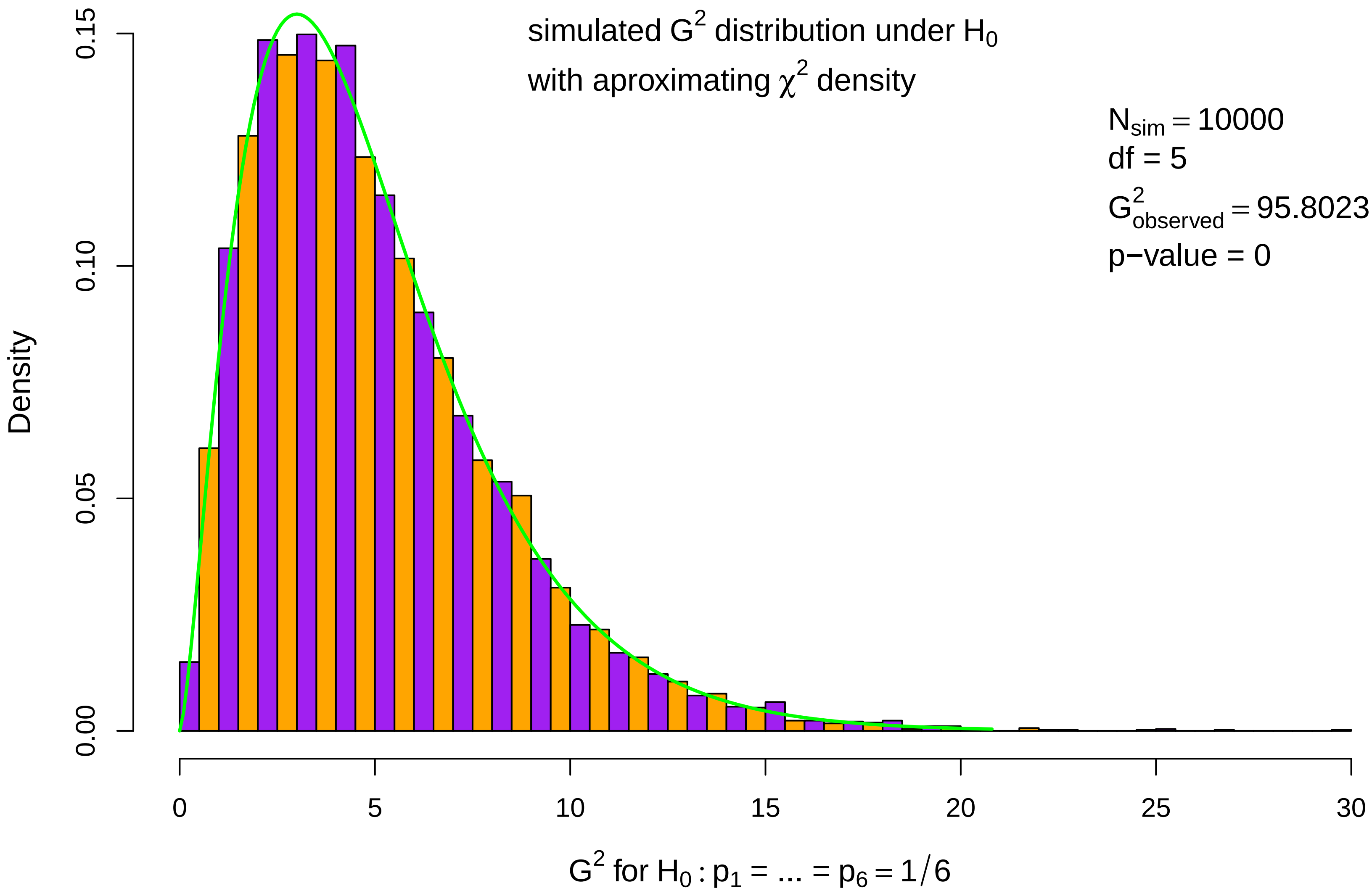
Testing $H_0 : p_1 + p_6 = p_2 + p_5 = p_3 + p_4$ we find with $\check{e}_j = n\check{p}_j$

$$
\begin{aligned}
\check{e}_1 &= 20000[3407/(3407+3422)]/3 = 3326.012 \\
\check{e}_6 &= 20000[3422/(3407+3422)]/3 = 3340.655 \\
\check{e}_2 &= 20000[3631/(3631+3448)]/3 = 3419.504 \\
\check{e}_5 &= 20000[3448/(3631+3448)]/3 = 3247.163 \\
\check{e}_3 &= 20000[3176/(3176+2916)]/3 = 3475.596 \\
\check{e}_4 &= 20000[2916/(3176+2916)]/3 = 3191.070
\end{aligned}
$$

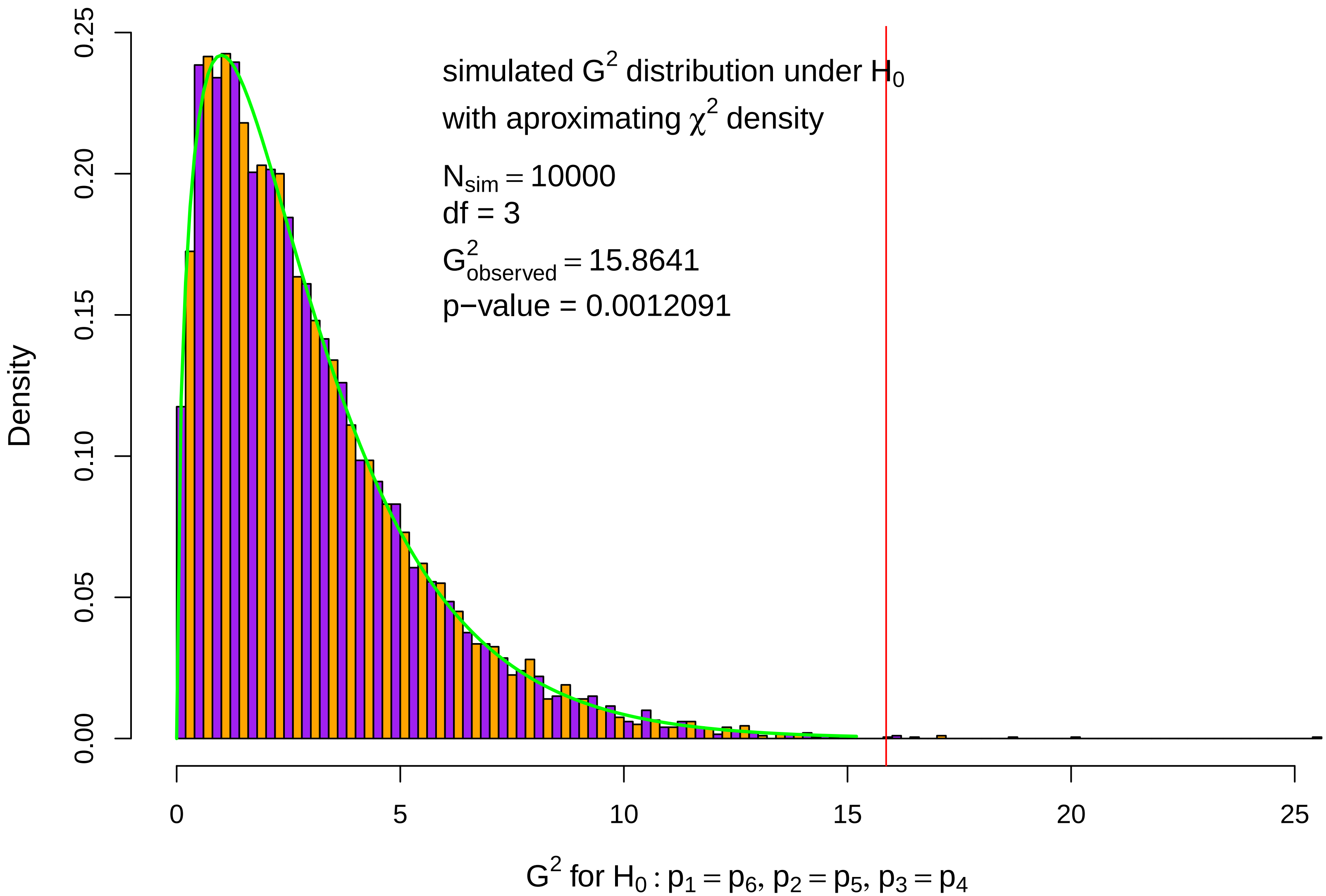
and obtain $G^2 = 79.9382$ and $X^2 = 79.0992$ with respective p-values

`1-pchisq(79.9382,df=2)=0` and `1-pchisq(79.0992,df=2)=0`.

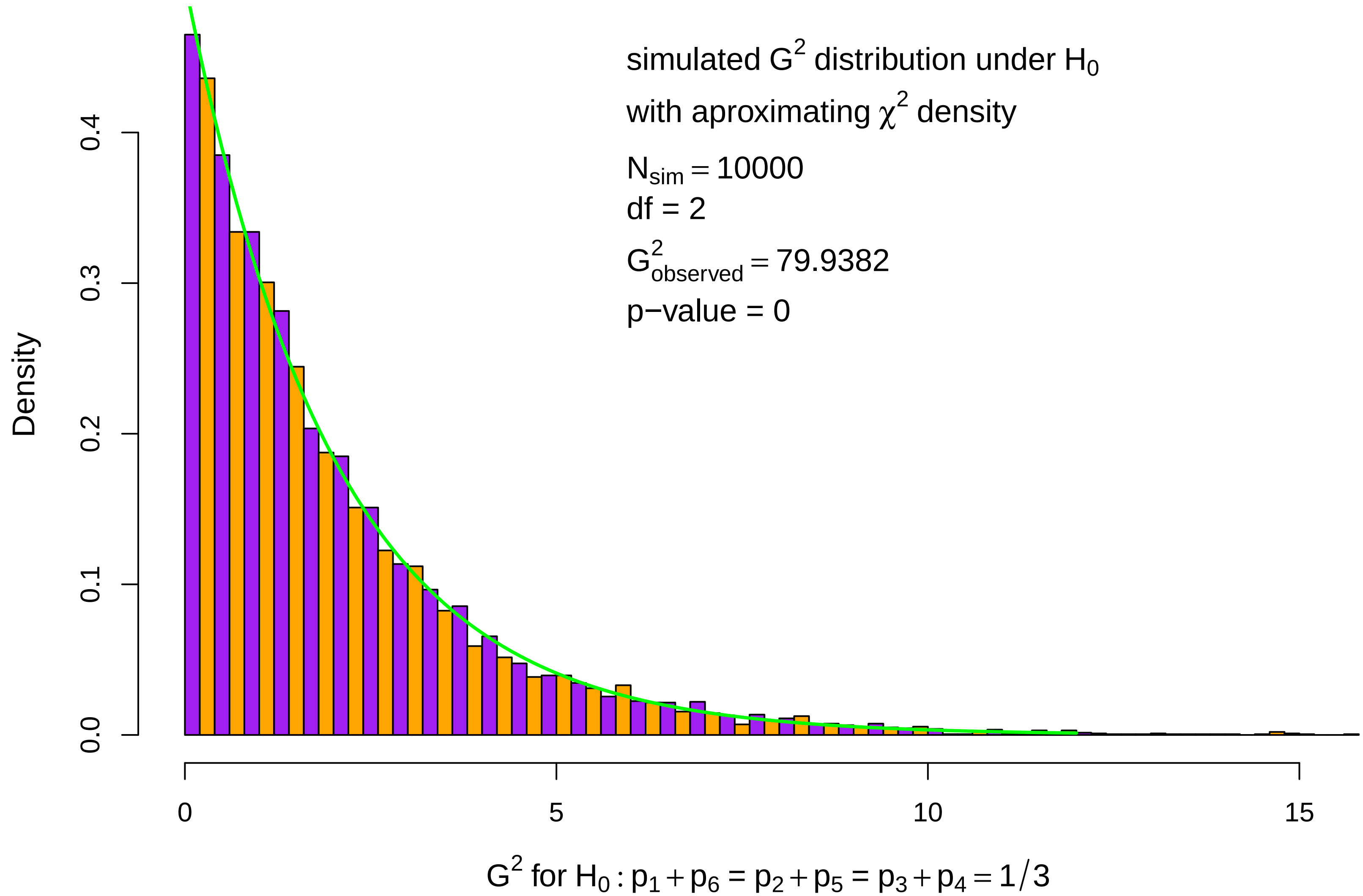
Testing $H_0 : p_1 = \ldots = p_6 = 1/6$ Using G^2



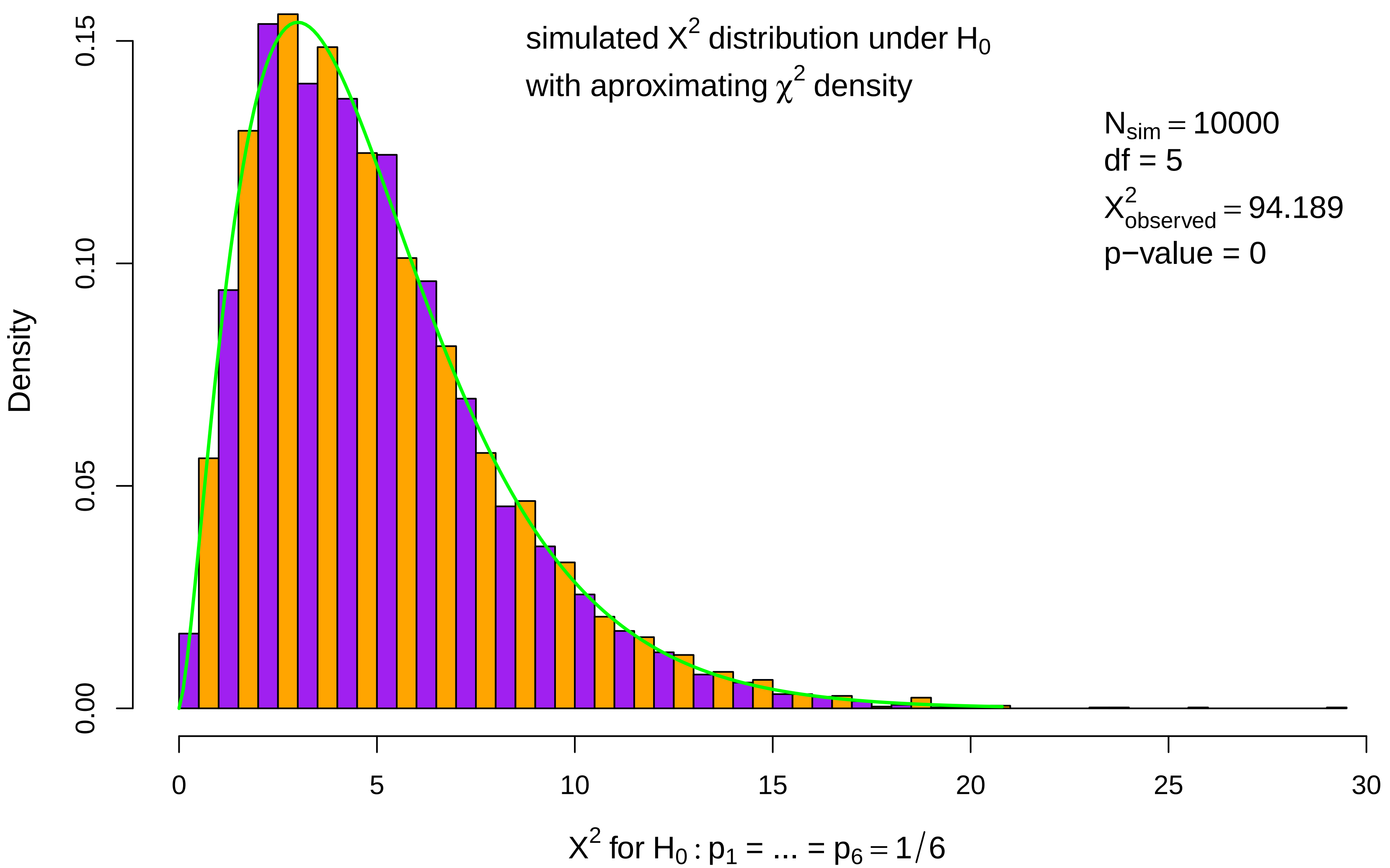
Testing $H_0 : p_1 = p_6, p_2 = p_5, p_3 = p_4$ Using G^2



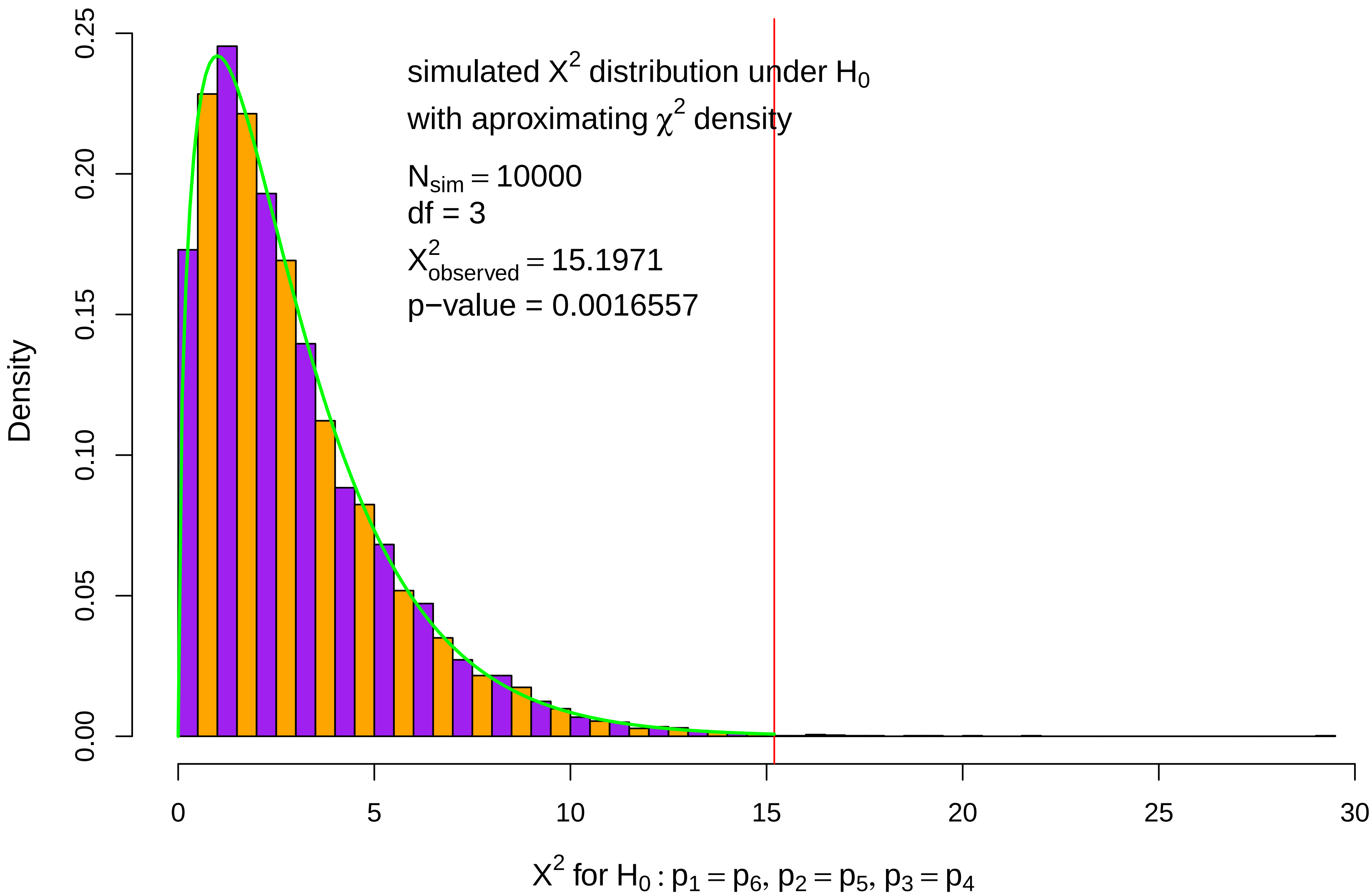
Testing $H_0 : p_1 + p_6 = p_2 + p_5 = p_3 + p_4 = 1/3$ Using G^2



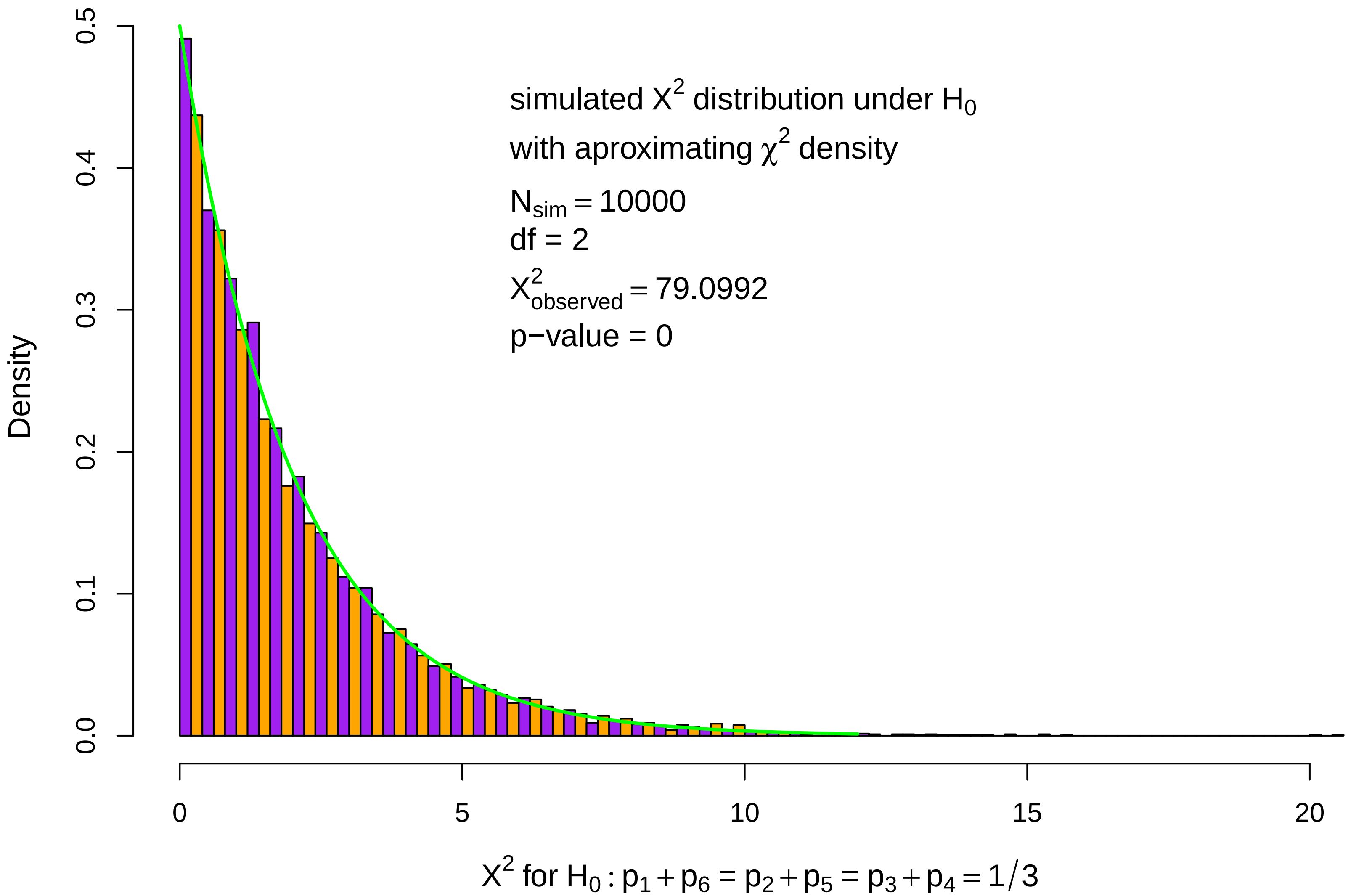
Testing $H_0 : p_1 = \ldots = p_6 = 1/6$ Using X^2



Testing $H_0 : p_1 = p_6, p_2 = p_5, p_3 = p_4$ Using X^2



Testing $H_0 : p_1 + p_6 = p_2 + p_5 = p_3 + p_4 = 1/3$ Using X^2



Some Comments

The 10000 simulated cell counts in 20000 rolls of a die with cell probabilities $\vec{p} = \mathrm{p}$ were generated via `rmultinom(10000,20000,p)` for $\vec{p} = (1/6,\ldots,1/6)$ and for the estimated value of $\vec{p}$ under the other two hypotheses.

The simulated distributions of G^2 and X^2 are well approximated by the respective chi-squared distributions.

At conventional significance levels all three null hypotheses should be rejected.

Compared with the other two, the null hypothesis $H_0 : p_1 = p_6, p_2 = p_5, p_3 = p_4$ seems to fall within the realm of possibilities.

The three G^2 discrepancy criteria add up $95.8023 = 15.8641 + 79.9382$, i.e., we have a decomposition of 95.8023. This suggests that the main reason for rejecting the fair die hypothesis is the rejection of the third hypothesis.

Testing Independence

Suppose the sample space S is partitioned two ways:

$$S = A_1 \cup \ldots \cup A_r\,, \quad \text{with } A_i \cap A_j = \emptyset \text{ for } i \neq j$$

$$S = B_1 \cup \ldots \cup B_c\,, \quad \text{with } B_i \cap B_j = \emptyset \text{ for } i \neq j$$

We can construct a third partition made of all intersections $E_{ij} = A_i \cap B_j = A_i B_j$.

$$S = \bigcup_{i=1}^{r} \bigcup_{j=1}^{c} E_{ij} \quad \text{with } A_i B_j \cap A_{i'} B_{j'} = \emptyset \text{ for } (i,j) \neq (i',j')$$

With respect to the A and B partitions it is often of interest

whether they are independent of each other, i.e., do we have

$$p_{ij} = P(E_{ij}) = P(A_i B_j) = P(A_i \cap B_j) = P(A_i) \cdot P(B_j) \quad \text{for all } (i,j)?$$

Karl Pearson's Crime Example

Karl Pearson studied the relationship between the type of crime

and the drinking habits of the involved criminal.

Are these two categorizations or partitions independent of each other?

	$B_1 =$ drink	$B_2 =$ abstain
$A_1 =$ arson	50	43
$A_2 =$ rape	88	62
$A_3 =$ violence	155	110
$A_4 =$ stealing	379	300
$A_5 =$ coining	18	14
$A_6 =$ fraud	63	144

Of course, one might argue that some of these classifications overlap and we

assume that such cases are resolved in a consistent manner, e.g.,

violence $=$ violence without rape.

Some Notation and Estimation

With $p_{ij} = P(E_{ij})$ we have

$$\begin{aligned} p_{i+} &= p_{i1} + p_{i2} + \ldots + p_{ic} = P(A_i B_1 \cup A_i B_2 \cup \ldots \cup A_i B_c) = P(A_i S) = P(A_i) \\ p_{+j} &= p_{1j} + p_{2j} + \ldots + p_{rj} = P(A_1 B_j \cup A_2 B_j \cup \ldots \cup A_r B_j) = P(S B_j) = P(B_j) \end{aligned}$$

The hypothesis of interest is $H_0 : p_{ij} = p_{i+} \cdot p_{+j}$ for $i = 1, \ldots, r, \;\; j = 1, \ldots, c$.

Let $o_{ij} =$ the count of observing E_{ij}, $o_{i+} =$ the count of observing $A_i = A_i B_1 \cup \ldots \cup A_i B_c$, and $o_{+j} =$ the count of observing $B_i = A_1 B_j \cup \ldots \cup A_r B_j$.

Then the unrestricted MLEs of p_{ij}, p_{i+} and p_{+j} are

$$\hat{p}_{ij} = \frac{o_{ij}}{n}\,, \quad \hat{p}_{i+} = \sum_{j=1}^{c} \hat{p}_{ij} = \frac{o_{i+}}{n}\,, \quad \text{and} \quad \hat{p}_{+j} = \sum_{i=1}^{r} \hat{p}_{ij} = \frac{o_{+j}}{n}$$

which are basically the plug-in estimates. Under H_0: mutual independence, the restricted MLEs are

$$\check{p}_{i+} = \hat{p}_{i+} = \frac{o_{i+}}{n}\,, \quad \check{p}_{+j} = \hat{p}_{+j} = \frac{o_{+j}}{n}\,, \quad \text{and} \quad \check{p}_{ij} = \check{p}_{i+} \cdot \check{p}_{+j}$$

Test Statistics G^2 and X^2

Under H_0 the estimated expected counts are

$$\check{e}_{ij} = n\check{p}_{ij} = n\frac{o_{i+}}{n} \cdot \frac{o_{+j}}{n} = \frac{o_{i+} \cdot o_{+j}}{n} = \frac{o_{i+} \cdot o_{+j}}{o_{++}} \qquad \text{since} \quad n = o_{++}$$

and as our G^2 and X^2 test statistics we get

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c} o_{ij} \log\left(\frac{o_{ij}}{\check{e}_{ij}}\right) \qquad \text{and} \qquad X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(o_{ij} - \check{e}_{ij})^2}{\check{e}_{ij}}$$

The null distribution of either statistic is well approximated by a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom. Here

$$(r-1)(c-1) = rc - r - c + 1 = (rc-1) - (r-1) - (c-1)$$

$rc - 1$ of the p_{ij} are free to vary in the unrestricted model, since $\sum_{ij} p_{ij} = 1$, and under H_0 the $p_{ij} = p_{i+} \cdot p_{+j}$ are restricted to $r - 1 + c - 1$ free parameters $p_{1+}, \ldots, p_{r+}$ and $p_{+1}, \ldots, p_{+c}$ since $\sum_i p_{i+} = 1$ and $\sum_j p_{+j} = 1$.

Analysis of Crime Data

```
PearsonCrime <- function(){
tab <- cbind(c(50,88,155,379,18,63),c(43,62,110,300,14,144))
rows <- apply(tab,1,sum); cols <- apply(tab,2,sum)
r <- length(rows); c <- length(cols); n <- sum(rows)
e0 <- outer(rows,cols,"*")/n
G2 <- 2*sum(tab*log(tab/e0)); X2 <- sum((tab-e0)^2/e0)
t.st <- c(G2,X2);names(t.st) <- c("G2","X2")
pG2=1-pchisq(G2,(r-1)*(c-1)); pX2=1-pchisq(X2,(r-1)*(c-1))
p.tst <- c(pG2,pX2)
list(test.statistics=t.st,p.values=p.tst)
}
> PearsonCrime()
$test.statistics
      G2       X2
50.51729 49.73061

$p.values
[1] 1.085962e-09 1.573317e-09   # highly significant
```