Multinomial Simulations: Why We Can (and Should) Use Them Instead of the Chi-Square Goodness-of-Fit Test

Peter E. Freeman - Department of Statistics & Data Science - Carnegie Mellon University

Motivation

Experiment: we toss a six-sided die *k* times

- The data $\mathbf{X} = \{X_1, \ldots, X_6\}$ are sampled according to a multinomial distribution:

$$\mathbf{X} \sim \text{Multinomial}(k, \mathbf{p})$$

where the sum of the data is $\Sigma_i X_i = k$ and the sum of the probabilities is $\Sigma_i p_i = 1$.

Question: is the die fair? Is $\mathbf{p_1} = \cdots = \mathbf{p_6} = \mathbf{1/6}$?

- A standard approach to testing this hypothesis is to use the (approximate) chi-square goodness-of-fit (GoF) test, first proposed by Karl Pearson in 1900. But…
  - in the low-k limit, *this test yields increasingly biased p-value estimates.*

To answer the question, we should use multinomial simulations!

The Old Approach: Chi-Square GoF Test

- In the late 19th century, determining whether a die was fair by working with the multinomial probability mass function directly was computationally infeasible.
- Knowing that a multinomial random variable converges in distribution to a multivariate normal random variable, Pearson (1900) proposed the following test statistic:

$$W = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{m} \frac{(O_i - kp_i)^2}{kp_i}.$$

  - O_i represents the number of observed counts in bin i (out of m bins overall)
  - p_i is the probability of recording a count in bin i under the null
  - $E_i = kp_i$ is the number of expected counts in bin i under the null
- Under the null hypothesis,

$$W \xrightarrow{d} Y \sim \text{ChiSquare}(m-1),$$

i.e., the statistic W converges in distribution to a chi-square random variable.

- A limitation when using the chi-square GoF test is the typically stated rule of thumb that E_i must be ≥ 5 in each bin (although variations on this rule exist).

The Better Approach: Multinomial Simulations

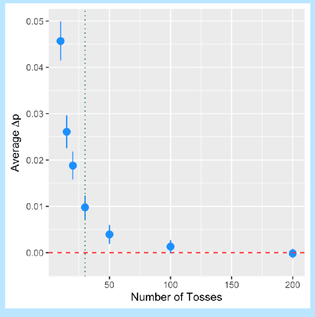
- The goal: to determine the proportion of datasets simulated under the null whose probability mass function values are equal to or smaller than the value we observe. **This is easily done!**

```
> set.seed(236)
> x.obs   <- c(2, 1, 4, 4, 3, 6)
> m       <- length(x.obs)
> k       <- sum(x.obs)
> p       <- rep(1/m,m)
> num.sim <- 100000
> pmf.obs <- dmultinom(x.obs,prob=p)
> X       <- rmultinom(num.sim,k,p)
> pmf.sim <- apply(X,2,function(x,p){dmultinom(x,prob=p)},p=p)
> sum(pmf.sim <= pmf.obs)/num.sim
[1] 0.47492
```

`x.obs`: observed data for *k* = 20 tosses
`m`: the number of faces (or bins)
`k`: number of multinomial trials
`p`: multinomial probabilities under the null hypothesis
`num.sim`: the number of simulations
`pmf.obs`: the multinomial pmf value for the observed data
`X`: matrix of datasets simulated under the null
`pmf.sim`: pmf values for simulated data
the empirically estimated *p*-value

To achieve greater precision, simply increase `num.sim`.

- The p-value is 0.475 (95% CI 0.472-0.478), in contrast to 0.467 for the chi-square GoF test.
- The simulation above runs for ~ 1 CPU second on a typical desktop/laptop computer.



- In the figure at left, $\Delta p = p_{\text{mult}} - p_{\text{chi}}$. ($\Delta p = 0.008$ for the simulation above.)
- The vertical green dashed line: the expected number of counts for each face is 5.
- For numbers of expected counts $\lesssim 20$, use of the chi-square GoF test leads to *biased* estimates of the true p-value.
  - $\overline{\Delta p} > 0 \;\Rightarrow$ the Type I error rate is *larger* on average for the chi-square GoF test

The take-home message: in the age of computers, there is no reason to continue to use the chi-square GoF test, since exact multinomial tests are easy to code *and* yield unbiased *p*-value estimates (for any value of *k*)!

At https://github.com/pefreeman/USCOTS-2025, you will find the R Shiny app shown below along with R Markdown-based materials that you can freely adapt for your own classroom use.

Looking for a free text for your mathematical statistics course?

Reference

Pearson, K. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, v. 50, pp. 157-175.