Bruce M. Boghosian

Introduction and motivation

Pearson's Goodness of Fit Test

Benford's Law

Summary

Goodness of Fit Tests

All Parameters Known

Bruce M. Boghosian

Tufts University School of Arts and Sciences

Department of Mathematics

Tufts University

Outline

1 Introduction and motivation

2 Pearson's Goodness of Fit Test

3 Benford's Law

4 Summary

Introduction and motivation

- Suppose that you can specify both of
  - the model distribution for your data, and
  - all parameters of that distribution.
- You would like to check GoF for your data to
  - the known model distribution,
  - with the known parameters.

Null and alternative hypotheses, continuous r.v.s

- Let $f_Y(y)$ be the true pdf.
- Let $f_0(y)$ be the presumed pdf.
- Null and alternative hypotheses:

$$H_0 : f_Y(y) = f_0(y)$$
$$H_1 : f_Y(y) \neq f_0(y)$$

Null and alternative hypotheses, discrete r.v.s

- Let $p_X(k)$ be the true probability distribution.
- Let $p_0(k)$ be the presumed probability distribution.
- Null and alternative hypotheses:

$$H_0 : p_X(k) = p_0(k)$$
$$H_1 : p_X(k) \neq p_0(k)$$

- Another way to describe H_0 and H_1 for discrete r.v.s:

$$H_0 : p_1 = p_{1_0},\ p_2 = p_{2_0}, \ldots, p_t = p_{t_0}$$
$$H_1 : p_j \neq p_{j_0} \text{ for at least one } j \in \{1, \ldots, t\}$$

Pearson's Goodness of Fit Test

Bruce M. Boghosian

- **Thm.:** Let $r_1, \ldots, r_t$ be the set of possible outcomes associated with each of n independent trials, where $P(r_i) = p_i$ for $i = 1, \ldots, t$. Let the r.v. X_i be the number of times r_i occurs for $i = 1, \ldots, t$.
- The r.v.
$$D = \sum_{i=1}^{t} \frac{(X_i - np_i)^2}{np_i}$$
has approximately a χ^2 distribution with $t-1$ degrees of freedom. (For the approximation to be adequate, the t classes should be defined so that $np_i \geq 5$, for all i.

Pearson's Goodness of Fit Test

Bruce M. Boghosian

Introduction and motivation

Pearson's Goodness of Fit Test

Benford's Law

Summary

- **Thm. (continued):** Let $k_1, \ldots, k_t$ be the observed frequencies for outcomes $r_1, \ldots, r_t$, respectively, and let $np_{1_0}, \ldots, np_{t_0}$ be the corresponding expected frequencies, based on the null hypothesis.
- At the α level of significance, $H_0 : \; f_Y(y) = f_0(y)$ (or similar discrete H_0) is rejected if

$$d = \sum_{i=1}^{t} \frac{(k_i - np_{i_0})^2}{np_{i_0}} \geq \chi^2_{1-\alpha, t-1},$$

where, again, $np_{i_0} \geq 5$ for all $i = 1, \ldots, t$.

Plausibility of Pearson's Goodness of Fit Test

- **Pf.:** A full proof is beyond the scope of this course.
- We can, however, motivate the case where $t = 2$.

$$
\begin{aligned}
D &= \frac{(X_1 - np_1)}{np_1} + \frac{(X_2 - np_2)}{np_2} \\
&= \frac{(X_1 - np_1)}{np_1} + \frac{[n - X_1 - n(1 - p_1)]}{n(1 - p_1)} \\
&= \frac{(X_1 - np_1)^2}{np_1(1 - p_1)} \\
&= \left[\frac{X_1 - E(X_1)}{\sqrt{\mathrm{Var}(X_1)}}\right]^2
\end{aligned}
$$

- Note D is the square of a variable that is asymptotically a standard normal, so it is χ^2 distributed with $2 - 1 = 1$ degrees of freedom.

Plausibility of Pearson's Goodness of Fit Test

Bruce M. Boghosian

Introduction and motivation

Pearson's Goodness of Fit Test

Benford's Law

Summary

- **Comment (given without proof):** A decision rule based on D is asymptotically equivalent to the GLRT of

$$H_0 : \; p_1 = p_{1_0}, \ldots, p_t = p_{t_0}.$$

- Likelihood function

$$L(p_1, \ldots, p_t) = \prod_{j=1}^{n} \frac{n!}{k_{1j}! \cdots k_{tj}!} p_1^{k_{1j}} \cdots p_t^{k_{tj}}$$

$$\begin{aligned} \therefore \ln L(p_1, \ldots, p_t) =& \sum_{j=1}^{n} \left[\ln \left(\frac{n!}{k_{1j}! \cdots k_{tj}!} \right) + k_{1j} \ln p_1 + \cdots + k_{tj} \ln p_t \right] \\ =& K_1 \ln p_1 + \cdots + K_t \ln p_t + \sum_{j=1}^{n} \ln \left(\frac{n!}{k_{1j}! \cdots k_{tj}!} \right) \end{aligned}$$

where $K_i := \sum_{j=1}^{n} k_{ij}$.

Plausibility of Pearson's Goodness of Fit Test

- We have
  - $\Omega = \{(p_1, \ldots, p_t) \mid p_1 + \cdots + p_t = 1\}$
  - $\omega = \{(p_1, \ldots, p_t) \mid p_1 = p_{1_0}, \ldots, p_t = p_{t_0}\}$
- To find $\max_\Omega L(p_1, \ldots, p_t)$, use Lagrange multiplier μ

$$0 = \frac{\partial}{\partial p_i}\left[\ln L(p_1, \ldots, p_t) - \mu\left(p_1 + \cdots + p_t\right)\right] = \frac{K_i}{p_i} - \mu,$$

so $p_i = \frac{K_i}{\mu}$, whence the normalization condition gives

$$p_i = \kappa_i := \frac{K_i}{K_1 + \cdots + K_t}.$$

- It follows that

$$\max_\Omega L\left(p_1, \ldots, p_t\right) = L\left(\kappa_1, \ldots, \kappa_t\right).$$

Plausibility of Pearson's Goodness of Fit Test

- We also have

$$\max_{\omega} L\left(p_1, \ldots, p_t\right) = L\left(p_{1_0}, \ldots, p_{t_0}\right)$$

- So the GLR is

$$\lambda = \frac{\max_{\omega} L\left(p_1, \ldots, p_t\right)}{\max_{\Omega} L\left(p_1, \ldots, p_t\right)} = \frac{L\left(p_{1_0}, \ldots, p_{t_0}\right)}{L\left(\kappa_1, \ldots, \kappa_t\right)}.$$

where

$$\kappa_i := \frac{K_i}{K_1 + \cdots + K_t}.$$

- Relating this to D in the asymptotic limit of large n is not straightforward.

Example: Benford's Law

- Many years ago, the astronomer Simon Newcomb noticed that the first pages of tables of logarithms are more smudged from use than later pages.
- Leading digit of number in scientific notation is 1 to 9.
- In tables of numbers, including seemingly random data and statistics of various sorts, you might think that each leading digit would appear with probability $1/9 \approx 11.1\%$.
- Instead numbers are observed to lead off with the digit 1 about 30% of the time, with digit 2 about 17.6% of the time, etc. They lead off with 9 only 4.6% of the time.
- This claim was checked by Frank Benford in the 1930s. It is called *Benford's Law*, or the *Newcomb-Benford Law*.

Explanation of Benford's Law

- T.P. Hill (1998)
- Benford's Law applies to dimensioned data – with units.
- The probability distribution of real numbers in scientific notation surely can not depend on these units, so it must be invariant under scaling, whence

$$P(cx) = f(c)P(x)$$

- If $\int dx\ P(x) = 1$, then $\int dx\ P(cx) = {}^1/_c$, so $f(c) = {}^1/_c$,

$$P(cx) = \frac{1}{c}P(x).$$

- Differentiate with respect to c: $xP'(cx) = -P(x)/c^2$.
- Set $c = 1$ to obtain $xP'(x) = -P(x)$.
- This has solution $P(x) = {}^1/_x$.

Explanation of Benford's Law

- $P(x) = {}^{1}\!/_{x}$ is not normalizable for all x, but we need to use it only in finite intervals.
- The probability that the leading digit is k is then

$$p_D(k) = \frac{\int_k^{k+1} dx\; P(x)}{\int_1^{10} dx\; P(x)} = \frac{\log(k+1) - \log k}{\log 10 - \log 1} = \log_{10}\left(1 + \frac{1}{k}\right)$$

- Results are consistent with what is often observed

d	$p_D(k)$
1	0.301030
2	0.176091
3	0.124939
4	0.0969100
5	0.0791812
6	0.0669468
7	0.0579919
8	0.0511525
9	0.0457575

Testing for Benford's Law

- Benford's Law is sometimes used in financial audits
- In first 355 digits of university's operating budget, observed digit frequency is

d	k_d
1	111
2	60
3	46
4	29
5	26
6	22
7	21
8	20
9	20

- Form the statistic

$$d = \frac{[111 - 355(0.301)]^2}{355(0.301)} + \cdots + \frac{[20 - 355(0.046)]^2}{355(0.046)} = 2.49$$

- Note $d < \chi^2_{0.95,8} = 15.507$ so we fail to reject H_0.

Summary

Bruce M. Boghosian

Introduction and motivation
Pearson's Goodness of Fit Test
Benford's Law
Summary

- We have motivated GoF tests with known parameters.
- We have stated and justified Pearson's GoF test.
- We have derived Benford's Law.
- We have studied GoF of data to Benford's Law.