Tufts UNIVERSITY

Goodness of Fit Tests

Parameters Unknown

Bruce M. Boghosian

Tufts University School of Arts and Sciences

Department of Mathematics

Tufts University

Outline

Bruce M. Boghosian

Revisiting Benford's Law

Review and example of GoF from last time

GoF tests with parameters unknown

Summary

Revisiting Benford's Law

- First digits of mantissas of data in scientific notation
- First digits can be $1, \ldots, 9$
- From an argument worked out last time, mantissa pdf is

$$P(x) = \frac{c}{x} \quad \text{for} \quad 1 \le x < 10.$$

- By normalization, $c = (\ln 10 - \ln 1)^{-1} = (\ln 10)^{-1}$
- Probability that mantissa first digit is d is then

$$\begin{aligned} p_d^{(1)} &= \frac{1}{\ln 10} \int_d^{d+1} \frac{dx}{x} = \frac{\ln(d+1)}{\ln 10} - \frac{\ln d}{\ln 10} \\ &= \log_{10}(d+1) - \log_{10} d = \log_{10}\left(1 + \frac{1}{d}\right). \end{aligned}$$

Mantissa first-digit probabilities

- If mantissas were uniformly distributed, first-digit probabilities $p_d^{(1)}$ would all be equal to $1/9$
- Instead $p_d^{(1)} = \log_{10}(1 + 1/d)$ for $d = 1, \ldots, 9$, yielding

d	$p_d^{(1)}$
1	0.30103
2	0.176091
3	0.124939
4	0.09691
5	0.0791812
6	0.0669468
7	0.0579919
8	0.0511525
9	0.0457575

What about second-digit probabilities?

- Note that second digits can be $0, \ldots, 9$
- If mantissas were uniformly distributed, second-digit probabilities $p_d^{(2)}$ would all be equal to $1/10$
- Actual 2nd digit probability is $p_d^{(2)} = \sum_{j=1}^{9} \log_{10}\left(\frac{j+\frac{d+1}{10}}{j+\frac{d}{10}}\right)$

d	$p_d^{(2)}$
0	0.119679
1	0.11389
2	0.108821
3	0.10433
4	0.100308
5	0.0966772
6	0.0933747
7	0.090352
8	0.0875701
9	0.0849974

And third-digit probabilities?

Bruce M. Boghosian

Revisiting Benford's Law

Review and example of GoF from last time

GoF tests with parameters unknown

Summary

- Note that third digits can be $0, \ldots, 9$
- If mantissas were uniformly distributed, third-digit probabilities $p_d^{(3)}$ would all be equal to $1/10$
- 3nd digit probs.: $p_d^{(3)} = \sum_{j=1}^{9} \sum_{k=0}^{9} \log_{10}\left(\frac{j+\frac{k}{10}+\frac{d+1}{100}}{j+\frac{k}{10}+\frac{d}{100}}\right)$

d	$p_d^{(3)}$
0	0.101784
1	0.101376
2	0.100972
3	0.100573
4	0.100178
5	0.0997876
6	0.0994013
7	0.0990192
8	0.0986412
9	0.0982672

Generalized Benford's Law

- Note that $p_d^{(n)} \to {}^1\!/10$ as n increases
- Results below are for $n = 4$ are given below

$$p_d^{(4)} = \sum_{j=1}^{9} \sum_{k=0}^{9} \sum_{l=0}^{9} \log_{10}\left(\frac{j + \frac{k}{10} + \frac{l}{100} + \frac{d+1}{1000}}{j + \frac{k}{10} + \frac{l}{100} + \frac{d}{1000}}\right)$$

d	$p_d^{(4)}$
0	0.100176
1	0.100137
2	0.100098
3	0.100059
4	0.100019
5	0.0999803
6	0.0999412
7	0.0999022
8	0.0998633
9	0.0998244

Pearson's Goodness of Fit Test

- **Key idea:** Bin your possible outcomes into t categories.
- **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 statistic
$$D = \sum_{i=1}^{t} \frac{(X_i - np_i)^2}{np_i}$$
has approximately a χ^2 distribution with $t-1$ df.
- 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

- **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 version of 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 $i = 1, \ldots, t$.

Example 1

- Over a number of years, the distribution of prison sentences for people convicted of grand theft has been approximately

$$f_{Y_0}(y) = \frac{y^2}{9} \quad \text{for } 0 < y \leq 3$$

- Recent review of sentences of 50 individuals showed
  - Eight served less than one year in jail.
  - Sixteen served between one and two years.
  - Twenty-six served between two and three years.
- Are these data consistent with $f_Y(y)$ using the $\alpha = 0.05$ level of significance?

Example 1 (continued)

- Note that $t = 3$ and all parameters are known.
- Theoretical bin probabilities
  - $p_{1_0} = \int_0^1 dy\ f_Y(y) = \int_0^1 dy\ \frac{y^2}{9} = \frac{1}{27}$
  - $p_{2_0} = \int_1^2 dy\ f_Y(y) = \int_1^2 dy\ \frac{y^2}{9} = \frac{7}{27}$
  - $p_{3_0} = \int_2^3 dy\ f_Y(y) = \int_2^3 dy\ \frac{y^2}{9} = \frac{19}{27}$
- Pearson statistic

$$d = \frac{\left(8 - 50 \cdot \frac{1}{27}\right)^2}{50 \cdot \frac{1}{27}} + \frac{\left(16 - 50 \cdot \frac{7}{27}\right)^2}{50 \cdot \frac{7}{27}} + \frac{\left(26 - 50 \cdot \frac{19}{27}\right)^2}{50 \cdot \frac{19}{27}} = 23.5212$$

- We have $\chi^2_{1-\alpha,t-1} = \chi^2_{0.95,2} = 5.991$.
- Since $d > \chi^2_{1-\alpha,t-1}$, we reject null hypothesis $H_0 :\ f_Y(y) = f_{Y_0}(y)$.

Example 1 (continued)

- In the way of explanation, note that
  - $50\left(\frac{1}{27}\right) = 1.85 < 8$
  - $50\left(\frac{7}{27}\right) = 12.96 < 16$
  - $50\left(\frac{19}{27}\right) = 35.19 > 26$
- There are more shorter sentences and fewer longer ones than the pdf would suggest.
- This is possibly due to recent judicial reforms that aim to correct the excesses of the Sentencing Reform Act of 1984, including [1]
  - US has 5% of world's population, but 25% of its prisoners.
  - US has 33% of world's female prisoners.
  - Federal prison population increased 800% since 1980.
  - Spending on federal prisons increased 1700% since 1980.
  - Federal prisons operating at 131% capacity.
  - Half of federal prisoners serving for nonviolent drug crimes.

[1] Statistics from ABA web page.

Example 2

- According to the M&M/Mars Company, the frequencies associated with M&M colors are
  - 30% brown
  - 20% each yellow and red
  - 10% each orange, blue and green
- Observed numbers in three pounds ($n = 1527$) of M&Ms

Color	k_i
Brown	455
Yellow	343
Red	318
Orange	152
Blue	130
Green	129

- Note $t = 6$ and all parameters known.
- Examine H_0 at confidence level $\alpha = 0.05$.

Example 2 (continued)

Bruce M. Boghosian

- Pearson statistic is

$$\begin{aligned} d =& \frac{(455 - 1527 \cdot 0.30)^2}{1527 \cdot 0.30} + \frac{(343 - 1527 \cdot 0.20)^2}{1527 \cdot 0.20} + \frac{(318 - 1527 \cdot 0.20)^2}{1527 \cdot 0.20} \\ &+ \frac{(152 - 1527 \cdot 0.10)^2}{1527 \cdot 0.10} + \frac{(130 - 1527 \cdot 0.10)^2}{1527 \cdot 0.10} + \frac{(129 - 1527 \cdot 0.10)^2}{1527 \cdot 0.10} \\ =& 12.2262 \end{aligned}$$

- We have $\chi^2_{1-\alpha,t-1} = \chi^2_{0.95,4} = 11.070$
- Since $d > \chi^2_{1-\alpha,t-1}$, we reject null hypothesis H_0.

What happens if you don't know the parameters?

- Suppose that you have s distribution parameters.
- If you don't know the s distribution parameters, use, e.g., MLE to estimate them.
- Substitute the MLE parameters into the distribution, and analyze as though the parameters are known.
- In this way, obtain $\hat{p}_i$ for $i = 1, \ldots, t$ for each of the t bins.
- Form the analog of Pearson's statistic

$$D_1 = \sum_{i=1}^{t} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

- This has approximate χ^2 distribution with $t - 1 - s$ df.
- Note one df lost for each parameter estimated.

Theorem on GoF with unknown parameters

- **Thm.:** Suppose n observations taken from the s-parameter distribution $f_Y(y)$ (or $p_X(k)$).
- Let $r_1, \ldots, r_t$ be mutually exclusive outcomes (bins, ranges) for each of the n observations.
- Let $\hat{p}_i$ be the estimated Prob$(Y \in r_i)$, as calculated from $f_Y(y)$ (or $p_X(k)$) after the parameters have been replaced by the MLE parameters.
- Let X_i denote the number of times that $Y \in r_i$ for $i = 1, \ldots, t$.
- Form the statistic

$$D_1 = \sum_{i=1}^{t} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

Theorem (continued)

- **Thm. (continued):** Then

$$D_1 = \sum_{i=1}^{t} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

is approximately distributed as a χ^2 distribution with $t - 1 - s$ degrees of freedom.

- Define the r_i so that $n\hat{p}_{i_0} \geq 5$ for all i.
- To test $H_0 : \ f_Y(y) = f_0(y)$ (or $H_0 : \ p_X(k) = p_0(k)$) at the α level of significance, calculate

$$d_1 = \sum_{i=1}^{t} \frac{\left(k_i - n\hat{p}_{i_0}\right)^2}{n\hat{p}_{i_0}}$$

where $k_1, \ldots, k_t$ are observed results, and $n\hat{p}_{i_0}$ are expected estimated frequencies based on H_0.

- Reject H_0 if $d_1 \geq \chi^2_{1-\alpha,t-1-s}$.

Example 1: Mortality for women over 80 yrs. old

- Data from the *Times of London* over three year period
- H_0 asserts this can be modeled by a Poisson distribution.
- See data below with $t = 11$ bins.

Number deaths (i)	Observed frequency (f_i)
0	162
1	267
2	271
3	185
4	111
5	61
6	27
7	8
8	3
9	1
10+	0

- Total days $= \sum_{j=1}^{11} f_j = 1096$
- Total deaths $= \sum_{j=1}^{11} j\, f_j = 2364$

Example 1 (continued)

- Poisson distribution is the one-parameter distribution

$$\hat{p}_{j_0} = \text{Prob}\left(j \text{ deaths in a day}\right) = \frac{\lambda^j}{j!} e^{-\lambda}.$$

- Estimate for λ is

$$\lambda_e = \frac{\text{Total deaths}}{\text{Total days}} = \frac{2364}{1096} = 2.15693$$

- With this estimate, estimated expected bin population is

$$n\hat{p}_{j_0} = n\frac{\lambda_e^j}{j!} e^{-\lambda_e}.$$

Example 1 (continued)

- Revised table, showing est. exp. Poisson distribution
- Note that we have less than five observations in bins 9-11.
- Fix this by creating a "7+" bin (next slide).

Number deaths (i)	Obs. freq. (f_i)	Est. exp. freq. ($n\hat{p}_{j_0}$)
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

Example 1 (continued)

- Merging bins to create a "7+" bin yields new table.
- Hence we now have $t = 8$.

Number deaths (i)	Obs. freq. (f_i)	Est. exp. freq. ($n\hat{p}_{j_0}$)
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+	12	7.3

- The d_1 statistic is then

$$d_1 = \frac{(162-126.8)^2}{126.8} + \frac{(267-273.5)^2}{273.5} + \cdots + \frac{(12-7.3)^2}{7.3} = 25.98$$

Example 1 (continued)

Bruce M. Boghosian

- The d_1 statistic is then

$$d_1 = \frac{(162 - 126.8)^2}{126.8} + \frac{(267 - 273.5)^2}{273.5} + \cdots + \frac{(12 - 7.3)^2}{7.3} = 25.98$$

- Since we have $t = 8$ classes and $s = 1$ estimated parameter, we examine

$$\chi^2_{1-\alpha,t-1-s} = \chi^2_{1-0.05,8-1-1} = \chi^2_{0.95,6} = 12.592$$

- Since $d_1 > \chi^2_{1-\alpha,t-1-s}$, we reject null hypothesis that data is Poisson distributed.
- Note that there is particular discrepancy in the $j = 0$ case. It may be that deaths occur in clusters, such as flu epidemics, invalidating the Poisson assumption (effectively making λ depend on time).

Summary

- We have generalized Benford's Law to include less significant digits, finding that less significant digits are more uniformly distributed.
- We have reviewed GoF tests with all parameters known and given two examples.
- We have described GoF tests with unknown parameters and given an example.