Independence Models, Likelihood Ratio Tests, and a Side of Bacon

Aubrey Komorowski
Department of Mathematics and Computer Science
Dr. John C. Kern, Advisor
Duquesne University
April 21, 2006

Outline

Pass the Pigs®

- Two or more players compete to earn 100 points
- Roll two pig-shaped dice
- Configuration determines point-value:

Position					
1	2	3	4	5	6

Configuration Categories:

1. Positive-Scoring Roll
2. Zero-Scoring Roll
3. Pigs at rest **and** in physical contact

The Pigs are Tired: Collecting our Data

- Data was collected from 6000 rolls generated by two people, each rolling a total of 3000 times.
- An imposed source of variation is that of rolling height: 3000 rolls were from 5 inches high, while the other 3000 rolls were from 8 inches high.
- In order to distinguish the two pigs, one from each of the two sets was randomly selected, and marked with a black dot on its snout.

The Data

Observed frequencies for the black–pink pig positions (6000 rolls):

Pink Pig Position

Black Pig Position		1	2	3	4	5	6
	1	573	656	139	360	56	12
	2	623	731	185	449	58	17
	3	155	180	45	149	17	5
	4	396	473	124	308	45	8
	5	54	67	13	47	2	1
	6	10	10	0	7	1	1

Note: 23 of the 6000 rolls resulted in physical contact between the two pigs.

Five Inch and Eight Inch Data Tables

5-Inch

		Pink Pig Position					
		1	2	3	4	5	6
Black Pig Position	1	328	349	79	202	30	6
	2	301	339	90	222	29	11
	3	85	99	27	84	13	1
	4	161	209	59	131	22	2
	5	30	29	9	25	2	1
	6	4	6	0	3	0	1

8-Inch

		Pink Pig Position					
		1	2	3	4	5	6
Black Pig Position	1	245	307	60	158	26	6
	2	322	392	95	227	29	6
	3	70	81	18	65	4	4
	4	235	264	65	177	23	6
	5	24	38	4	22	0	0
	6	6	4	0	4	1	0

Likelihood Function Introduction

Let X_1 and X_2 be i.i.d. Poisson r.v.'s.

$$p(X_1 = x_1 \; and \; X_2 = x_2) \; = \; \left(\frac{\lambda^{x_1} e^{-\lambda}}{x_1!}\right) \cdot \left(\frac{\lambda^{x_2} e^{-\lambda}}{x_2!}\right)$$

The Likelihood Function, L(λ), treats x_1 and x_2 as fixed; λ unknown:

$$L(\lambda) \; = \; \frac{\lambda^{x_1} e^{-\lambda} \lambda^{x_2} e^{-\lambda}}{x_1! x_2!} \; \propto \; \lambda^{x_1 + x_2} e^{-2\lambda}$$

Theorem: Let X_1, …, X_n be i.i.d. r.v.'s from distribution $f(x \mid \theta)$. For H_o: $\theta \in \Omega_0$,

$$-2 \cdot \ln \left[\frac{\sup_{\theta \in \Omega_0} L(\theta)}{\sup L(\theta)} \right]$$

is approximately χ^2_r when H_o is true and n is large.

Likelihood Ratio Test (LRT) Example

Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. Then

$$L(\mu, \sigma^2) \;=\; \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Testing $H_o : \mu = 0$ versus $H_a : \mu \neq 0$ gives

$$-2 \cdot \ln \left[\frac{\sup_{\mu=0} L(\mu, \sigma^2)}{\sup L(\mu, \sigma^2)}\right] \;=\; -n \cdot \left[1 - \frac{\sum x_i^2}{\sum (x_i - \bar{x})^2}\right]$$

$$Note : \hat{\mu} \;=\; \bar{x} \qquad\qquad \hat{\sigma^2} \;=\; \frac{\sum (x_i - \bar{x})^2}{n}$$

- Histogram of 5000 LRT statistic realizations, where n=250 (The χ^2_1 density is superposed):

Multinomial Model for Pig Data

- Let n_{ijk} denote the observed frequency and p_{ijk} denote the unknown probability in the i^{th} row, j^{th} column, and k^{th} layer.

- Model n_{ijk} as multinomial, with parameters p_{ijk}. The corresponding likelihood for the p_{ijk}'s is then:

$$L(\vec{p}) \propto \prod_{i}^{I} \prod_{j}^{J} \prod_{k}^{K} p_{ijk}^{n_{ijk}}$$

$$\ln(L(\vec{p})) \; = \; \sum_{i}^{I} \sum_{j}^{J} \sum_{k}^{K} n_{ijk} \ln(p_{ijk})$$

- $L(\vec{p})$ is maximized when $p_{ijk} \; = \; \frac{n_{ijk}}{n_{\ldots}}$

The LRT statistic is ln(restricted)− ln(unrestricted),

$$G^2 \; = \; 2 \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} n_{ijk} \ln \left(\frac{n_{ijk}}{\hat{m}_{ijk}^{(l)}} \right)$$

where $\hat{m}_{ijk}^{(l)}$ is the Maximum Likelihood Estimate (MLE) under the restricted model l.

MLE's for Models 0 and 1

- **Model 0:** Complete Independence

$$
\begin{aligned}
M_0 : p_{ijk} &= p_{i..}p_{.j.}p_{..k} \\
\hat{m}_{ijk}^{(0)} &= \frac{n_{i..}n_{.j.}n_{..k}}{n_{...}^2}
\end{aligned}
$$

2 x 2 Example:

1	2
2	4

3	6
6	12

- **Model 1:** Rows are independent of columns and layers.

$$
\begin{aligned}
M_1 : p_{ijk} &= p_{i..}p_{.jk} \\
\hat{m}_{ijk}^{(1)} &= \frac{n_{i..}n_{.jk}}{n_{...}}
\end{aligned}
$$

2 x 2 Example:

1	5
2	10

4	3
8	6

MLE's for Models 2 and 3

- **Model 2:** Columns are independent of rows and layers.

$$\begin{aligned} M_2 : p_{ijk} &= p_{.j.}p_{i.k} \\ \hat{m}^{(2)}_{ijk} &= \frac{n_{.j.}n_{i.k}}{n_{...}} \end{aligned}$$

2 x 2 Example:

1	2
5	10

4	8
3	6

- **Model 3:** Layers are independent of rows and columns.

$$\begin{aligned} M_3 : p_{ijk} &= p_{..k}p_{ij.} \\ \hat{m}^{(3)}_{ijk} &= \frac{n_{..k}n_{ij.}}{n_{...}} \end{aligned}$$

2 x 2 Example:

3	7
6	2

8	2
1	7

MLE's for Models 4 and 5

- **Model 4:** Rows are independent of columns **given** layers.

$$\begin{aligned} M_4 : p_{ijk} &= \frac{p_{i.k}p_{.jk}}{p_{..k}} \\ \hat{m}^{(4)}_{ijk} &= \frac{n_{i.k}n_{.jk}}{n_{..k}} \end{aligned}$$

2 x 2 Example:

1	5
2	10

4	3
8	6

- **Model 5:** Rows are independent of layers **given** columns.

$$\begin{aligned} M_5 : p_{ijk} &= \frac{p_{ij.}p_{.jk}}{p_{.j.}} \\ \hat{m}^{(5)}_{ijk} &= \frac{n_{ij.}n_{.jk}}{n_{.j.}} \end{aligned}$$

2 x 2 Example:

1	2
2	4

6	1
12	2

MLE for Model 6

- **Model 6:** Columns are independent of layers **given** rows.

$$\begin{aligned} M_6 : p_{ijk} &= \frac{p_{ij.}p_{i.k}}{p_{i..}} \\ \hat{m}^{(6)}_{ijk} &= \frac{n_{ij.}n_{i.k}}{n_{i..}} \end{aligned}$$

2 x 2 Example:

1	3
2	5

2	6
4	10

The Pig Data from the 5 and 8 inch tables was applied to **Models 0-6**, in pursuit of finding a best fit.

Hypothesis Test:

H_o : Model l fits the data ($L(\vec{p})$ maximized at $\hat{m}^{(l)}_{ijk}$)

H_a : Unrestricted model fits data ($L(\vec{p})$ maximized at $\frac{n_{ijk}}{n_{...}}$)

G^2 for Models 0-6 applied to Pig Data

	Model	$d.f.$	G^2	$p - value$
most restrictive	0	60	115.1995	0.000024
	1	55	111.6932	0.000006
	2	55	57.427	0.38533
	3	35	85.265	0.000004
	4	50	53.9209	0.32686
	5	30	81.7596	0.000001
least restrictive	6	30	27.4936	0.59725

- Test Model 2 vs. Model 4:

$$\begin{aligned} G^2 &= 57.427 - 53.9209 = 3.5061 \\ d.f. &= 55 - 50 = 5 \\ p - value &= 0.622 \end{aligned}$$

- Test Model 2 vs. Model 6:

$$\begin{aligned} G^2 &= 57.427 - 27.4936 = 29.9334 \\ d.f. &= 55 - 30 = 25 \\ p - value &= 0.22679 \end{aligned}$$

Model 2 (columns independent of rows and layers) is the best fit for the Pig Data.

G^2 applied to Pig Data *transposed*

Model	$d.f.$	G^2	$p - value$
0	60	94.87904	0.00275
1	55	77.39603	0.0249
2	55	71.404	0.067691
3	35	66.646	0.00099
4	50	53.9209	0.32686
5	30	49.163	0.01512
6	30	43.171	0.05659

- Test Model 2 vs. Model 4:

$$\begin{aligned} G^2 &= 71.404 - 53.9209 = 17.4831 \\ d.f. &= 55 - 50 = 5 \\ p - value &= 0.003 \end{aligned}$$

- Test Model 2 vs. Model 6:

$$\begin{aligned} G^2 &= 71.404 - 43.171 = 28.233 \\ d.f. &= 55 - 30 = 25 \\ p - value &= 0.2972 \end{aligned}$$

Model 4 (rows are independent of columns **given** layers) is the best overall fit.

One Final Test

Scoring Table for all positive-scoring and zero-scoring configurations

Pig 1 Position

Pig 2 Position

	1	2	3	4	5	6
1	1	0	5	5	10	15
2	0	1	5	5	10	15
3	5	5	20	10	15	20
4	5	5	10	20	15	20
5	10	10	15	15	40	25
6	15	15	20	20	25	60

Note: Different position combinations yield the same point value.

- What happens when we test the independence of *point-value* and roll-height?

Point Values of the Pig Data Rolls

	Observed		Expected	
Point Value	5 in.	8 in.	5 in.	8 in.
0	650	629	639.609	639.393
1	667	637	652.109	651.891
5	1147	1190	1168.695	1168.305
10	261	247	254.042	253.957
15	96	75	85.514	85.486
20	164	209	186.531	186.469
25	1	1	1.00	0.999
40	2	0	1.00	0.999
60	1	0	0.500	0.499

H_0: Point-value and Roll-height are independent

$$G^2 \;=\; 2\sum_{i=1}^{I}\sum_{j=1}^{J} observed \cdot \;\ln\left(\frac{observed}{expected}\right)$$

$$= 14.39855 \sim \chi^2_8$$

$$p - value \;=\; 0.072$$

Conclusion

- The computed LRT statistics revealed that Model 2 was the best fit for the Pig Data (i.e. columns independent of rows and layers).
- The best fit Model for the *transposed* Pig Data was Model 4 (i.e. rows independent of columns **given** layers).

Different pig assignments yield a different "best fit" model

- Our final test yielded a non-significant $p - value$ of 0.07; so there is no clear evidence of any relationship between point-value and roll-height.
- **Pig-landing surface as a confounding variable:** Hardwood floor vs. Formica table.
- From the player's perspective, the only results to consider are those from the final test.

References

1. Christensen, Ronald (1990). *Log-Linear Models.* New York: Springer-Verlag, 61-90.
2. Mood, Alexander M., Graybill, Franklin A., and Boes, Duane C. (1963). *Introduction to the Theory of Statistics.* New York: McGraw-Hill, Inc, 440-460.
3. Hogg, Robert V. and Tanis, Elliot A. (2001). *Probability and Statistical Inference.* Upper Saddle River: Prentice-Hall, Inc, 577-582.
4. Bishop, Y., Fienberg, S., and Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge: MIT Press, 56-121.