STAT 513 fa 2020 Lec 10 slides

Contingency Table Analysis

Karl B. Gregory

University of South Carolina

These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.

Exercise: Let Y_1 and Y_2 be independent rvs such that

$$
\begin{aligned}
Y_1 &\sim \text{Binomial}(n_1, p_1) \\
Y_2 &\sim \text{Binomial}(n_2, p_2).
\end{aligned}
$$

Derive the asymptotic LRT for

$$H_0\colon\ p_1 = p_2 \text{ versus } H_1\colon\ p_1 \neq p_2.$$

Exercise (cont): Formulate the asymptotic LRT as

$$\text{Reject } H_0 \text{ iff } 2\sum_{i=1}^{2}\sum_{j=1}^{2} O_{ij} \log\left(\frac{O_{ij}}{E_{ij}}\right) > \chi^2_{1,\alpha}.$$

Exercise: Random assignment of 75 subjects to treatment (surgery) or control (incision only). "Success" if subject experienced reduction in migraine pain.

	Successes	Failures	Total
Treatment	41	8	49
Control	15	11	26
Total	56	19	75

Use asymp. LRT to test whether treatment has any effect at $\alpha = 0.05$.

Another kind of test: *Fisher's exact test*.

			Total
	X_{obs}		R_1
			R_2
Total	C_1	C_2	N,

- Draw C_1 marbles from a bag with N marbles, R_1 of which are red.
- Let $X = \#$ red marbles drawn.
- Then $X \sim \text{Hypergeometric}(N, R_1, C_1)$.

Fisher's exact test rejects H_0 when this p-value is less than α:

$$p_{\text{Fisher}} = \sum_{x=\max\{0, C_1+R_1-N)\}}^{\min\{C_1, R_1\}} P(X = x) \cdot \mathbf{1}\left(P(X = x) \leq P(X = X_{\text{obs}})\right)$$

Exercise: Compute the p-value of Fisher's exact test based on the data

	Successes	Failures	Total
Treatment	41	8	49
Control	15	11	26
Total	56	19	75

Example: Counts of eye and hair color from 6800 people:

	Brown	Black	Fair	Red	Total
Brown	438	288	115	16	857
Grey or Green	1387	746	946	53	3132
Blue	807	189	1768	47	2811
Total	2632	1223	2829	116	6800

How do we test for an association between eye and hair color?

Multinoulli trial

An experiment with $M \geq 1$ possible outcomes having probabilities

$$p_1, \ldots, p_M \quad \text{such that} \quad p_1 + \cdots + p_M = 1$$

is called a *multinoulli trial*.

Extension of the *Bernoulli trial*, which has two outcomes with probs p and $1 - p$.

Multinoulli random vector

Let $X = (X_1, \dots, X_M)^T$ be a rvec based on a Multinoulli trial w/ prbs $p_1, \dots, p_M$ such that

$$X_j = \begin{cases} 1 & \text{if outcome } j \text{ occurs} \\ 0 & \text{otherwise} \end{cases} \quad \text{for } j = 1, \dots, M.$$

Then X is a *Multinoulli random vector* and has pmf

$$P((X_1, \dots, X_M)^T = (x_1, \dots, x_M)^T) = p_1^{x_1} \cdot \dots \cdot p_M^{x_M}$$

for all $(x_1, \dots, x_M)^T$ having one nonzero entry which is equal to 1.

We write $X \sim \text{Multinoulli}(p_1, \dots, p_M)$.

Multinomial distribution

Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Multinoulli}(p_1, \dots, p_M)$ and let $Y = \sum_{i=1}^n X_i$.

Then $Y = (Y_1, \dots, Y_M)^T$ is a *Multinomial rvec* and has pmf

$$P((Y_1, \dots, Y_M)^T = (y_1, \dots, y_M)^T) = \left(\frac{n!}{y_1! \cdots y_M!}\right) p_1^{y_1} \cdots p_M^{y_M}$$

for $(y_1, \dots, y_M) \in \{0, 1, \dots, n\}^M$ such that $\sum_{j=1}^M y_j = n$.

We write $Y \sim \text{Multinomial}(p_1, \dots, p_M, n)$.

Note that $\text{Multinomial}(p, 1-p, n)$, which has $M = 2$, is same as $\text{Binomial}(n, p)$.

Exercise: Find the MLEs of $p_1, \dots, p_M$ based on $Y \sim \text{Multinomial}(p_1, \dots, p_M, n)$.

Exercise: Let $Y_1, \dots, Y_K$ be independent rvecs such that

$$Y_k = (Y_{k1}, \dots, Y_{kM})^T \sim \text{Multinomial}(p_{k1}, \dots, p_{kM}, n_k), \quad \text{for } k = 1, \dots, K.$$

Derive the asymptotic LRT for

$$H_0\colon (p_{11}, \dots, p_{1M}) = \dots = (p_{K1}, \dots, p_{KM})$$
$$\text{versus } H_1\colon (p_{j1}, \dots, p_{jM}) \neq (p_{i1}, \dots, p_{iM}) \text{ for some } i \neq j.$$

Note: For $K = 2$ and $M = 2$, these hypotheses become those in Slide 2.

Exercise: Formulate the asymptotic LRT as

$$\text{Reject } H_0 \text{ iff} \quad 2\sum_{j=1}^{M}\sum_{k=1}^{K} O_{kj}\log\left(\frac{O_{kj}}{E_{kj}}\right) > \chi^2_{(K-1)(M-1),\alpha}.$$

Exercise: Counts of eye and hair color from 6800 people:

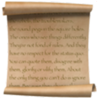
	Brown	Black	Fair	Red	Total
Brown	438	288	115	16	857
Grey or Green	1387	746	946	53	3132
Blue	807	189	1768	47	2811
Total	2632	1223	2829	116	6800

Use asymptotic LRT to test for an association at $\alpha = 0.01$.

Pearson's chi-squared test

A classical test of no association is *Pearson's chi-squared test*, which is

$$\text{Reject } H_0 \text{ iff } \quad \sum_{j=1}^{M}\sum_{k=1}^{K}\frac{(O_{kj}-E_{kj})^2}{E_{kj}} > \chi^2_{(K-1)(M-1),\alpha}.$$



Predates the development of the likelihood ratio approach.

Fairly close to the asymptotic LRT.

Exercise: Compute Pearson's test statistic on the migraine surgery data:

	Successes	Failures	Total
Treatment	41	8	49
Control	15	11	26
Total	56	19	75