Handout 8

Three-way tables

A $I \times J \times K$ table of counts is called a three-way contingency table. Here is an example.

Example 1. From the past year's record in a city, we have the counts of cured and not-cured for two surgical procedures for a certain disease. There are three hospitals in this city.

	Hospital 1		Hospital 2		Hospital 3		
Surgical Procedure	Cured	Not-Cured	Cured	Not-Cured	Cured	Not-Cured	Total
A	52	38	147	65	42	30	374
B	61	26	153	94	102	56	492
Total	113	64	300	159	144	86	866

X =surgical procedure, Y =outcome of surgery, Z=hospital.

This an example of a three-way table with $I = 2, J = 2$ and $K = 3$.

Joint Probability: $\pi_{ijk} = P(X = i, Y = j, Z = k)$.

Conditional probability:

$$P(X = i, Y = j|Z = k) = \frac{P(X = i, Y = j, Z = k)}{P(Z = k)} = \frac{\pi_{ijk}}{\pi_{++k}}.$$

For instance

$$\begin{aligned} \pi_{121} &= 38/866 = 0.0439, \\ P(X &= 1, Y = 2|Z = 3) = \frac{30}{144 + 86} = \frac{30}{230} = 0.1304. \end{aligned}$$

Conditional odds

$$\begin{aligned} \text{hospital 1} &: \quad \theta_{XY(1)} = \frac{(52)(26)}{(61)(38)} = 0.583, \\ \text{hospital 2} &: \quad \theta_{XY(2)} = \frac{(147)(94)}{(153)(65)} = 0.961, \\ \text{hospital 2} &: \quad \theta_{XY(2)} = \frac{(42)(56)}{(102)(30)} = 0.769. \end{aligned}$$

Independence and Conditional Independence

X, Y and Z are independent if $\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$ for all i, j, k.

X and Y are marginally independent if for all j and k

$$P(X = i, Y = j) = P(X = i)P(Y = j), \text{ ie, } \pi_{ij+} = \pi_{i++}\pi_{+j+}.$$

X and Y are conditionally independent (given Z) if for all i, j, k

$$\begin{aligned} P(X &= i, Y = j|Z = k) = P(X = i|Z = k)P(Y = j|Z = k), \ ie \\ \frac{\pi_{ijk}}{\pi_{++k}} &= \frac{\pi_{i+k}}{\pi_{++k}}\frac{\pi_{+jk}}{\pi_{++k}}, \text{ ie, } \pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}. \end{aligned}$$

Conditional independence of X and Y (given Z) does not imply marginal independence of X and Y. The textbook has an example of this in Section 2.7.5.

We will deal with all these issues later in detail when we examine loglinear models.

Simpson's Paradox.

Suppose we are observing several groups, and establish a relationship or correlation for each of these groups. Simpson's paradox says that when we combine all of the groups together and look at the data in aggregate form, the correlation that we noticed before may reverse itself. This is most often due to lurking variables that have not been considered, but sometimes it is due to the numerical values of the data

We begin with an example.

Example 2. Comparison of TB deaths in 1910 in New York City versus Richmond, Virginia reveal that the mortality was lower in New York City. However, the mortality among whites was higher in New York City, and the mortality among blacks was also higher in New York City.

How could this happen? Here are some fictitous numbers.

Mortality (percentage of death among TB patients):

NYC, 20% among whites and 40% among nonwhites,

RM (Richmond), 10% among whites and 30% among nonwhites.

Proportion of whites in NYC and RM:

NYC: 90% white, RM: 20% white.

Clearly, mortality among whites in NYC is higher than mortality among whites in RM, and mortality among nonwhites in NYC is higher than mortality among nonwhites in RM.

However, we can calculate overall mortality

$$\begin{aligned} \text{NYC} &: \quad (0.2)(0.9) + (0.4)(0.1) = 0.22, \\ \text{RM} &: \quad (0.1)(0.2) + (0.3)(0.8) = 0.26. \end{aligned}$$

Thus TB mortality in NYC is 22% which is lower than TB mortality in RM.

Now we consider a second hypothetical example that is taken from the pages of a course posted on the web (Math 425, Univ. of Michigan, Fall 2015).

Example 3. Suppose we are treating a life-threatening disease. We have to choose between an old method of treatment and a new one. To assess whether the new treatment is better than the old, we take 80 patients, and treat 40 of them in the new way, and 40 of them in the old. We find that among the patients treated in the new way, 20 are cured and 20 die, while among the patients treated in the old way, 24 are cured and 16 die.

	Cured	Died	Total	Cure Rate
New	20	20	40	$20/40 = 0.5$
Old	24	16	40	$24/40 = 0.6$

Here the old cure-rate, 60%, was better than the new, 50%. But now we make a finer analysis of the same data by examining separately male and female patients. Suppose that when this is done, the following numbers emerge:

	Males					Females		
	Cured	Died	Cure Rate			Cured	Died	Cure Rate
New	8	2	$8/10 = 0.8$			12	18	$12/30 = 0.4$
Old	21	9	$21/30 = 0.7$			3	7	$3/10 = 0.3$

Among male patients the new cure rate, 80%, is better than the old cure rate, 70%. Also, among female patients, the new cure rate, 40%, is better than the old, 30%. Thus a man would prefer the new treatment, and so would a woman, but the overall cure-rate is higher for the old method of treatment..