Chapter 6 : Two Way Tables

This chapter introduces some ideas to analyze relationships between categorical data. Regression allows one to model (if reasonable data properties hold) relationships through linear regression. Earlier we have seen that we can use bar charts or graphs and pie charts to help with understanding the distribution of a categorical data type.

Example

Use the data from problem 6.2 Edition 4 (This data is not in Edition 5).

Table 1 gives the data by counts for each cause of death. Table 2 gives the proportion of the deaths by cause for each age group. Since pictures are often a convenient way of showing such a summary Figure 1 shows the three bar plots, one for each age group.

How are the proportions calculated? These are done for each age group. For age 15 to 24,the death counts are

$$14966, 171, 1628, 1083, 5148, 3921, 6105$$

and the total number of deaths is 33022. Thus the proportion of deaths by cause are

$$\frac{14966}{33022}, \frac{171}{33022}, \frac{1628}{33022}, \frac{1083}{33022}, \frac{5148}{33022}, \frac{3921}{33022}, \frac{6105}{33022}$$

and these are recorded in Table 2.

Figure 1 and Table 2 show that the cause of death changes quite a bit from one age group to another. We see that proportionately accidents and murder decreases by age group, but cancer increases by age group.

Table 1: Causes of Death by Age Group

cause	15-24	25-44	45-64
Accidents	14966	27844	23699
AIDS	171	6879	5917
Cancer	1628	19041	144936
Heart disease	1083	16283	101713
Murder	5148	7367	2756
Suicide	3921	11251	1057
Other	6105	40259	156980
Total deaths	33022	128924	437058

Table 2: Proportion of Death by Cause and Age Group

cause	age15.24	age25.44	age45.64
Accidents	0.453	0.216	0.054
AIDS	0.005	0.053	0.014
Cancer	0.049	0.148	0.332
Heart	0.033	0.126	0.233
Murder	0.156	0.057	0.006
Suicide	0.119	0.087	0.002
Other	0.185	0.312	0.359

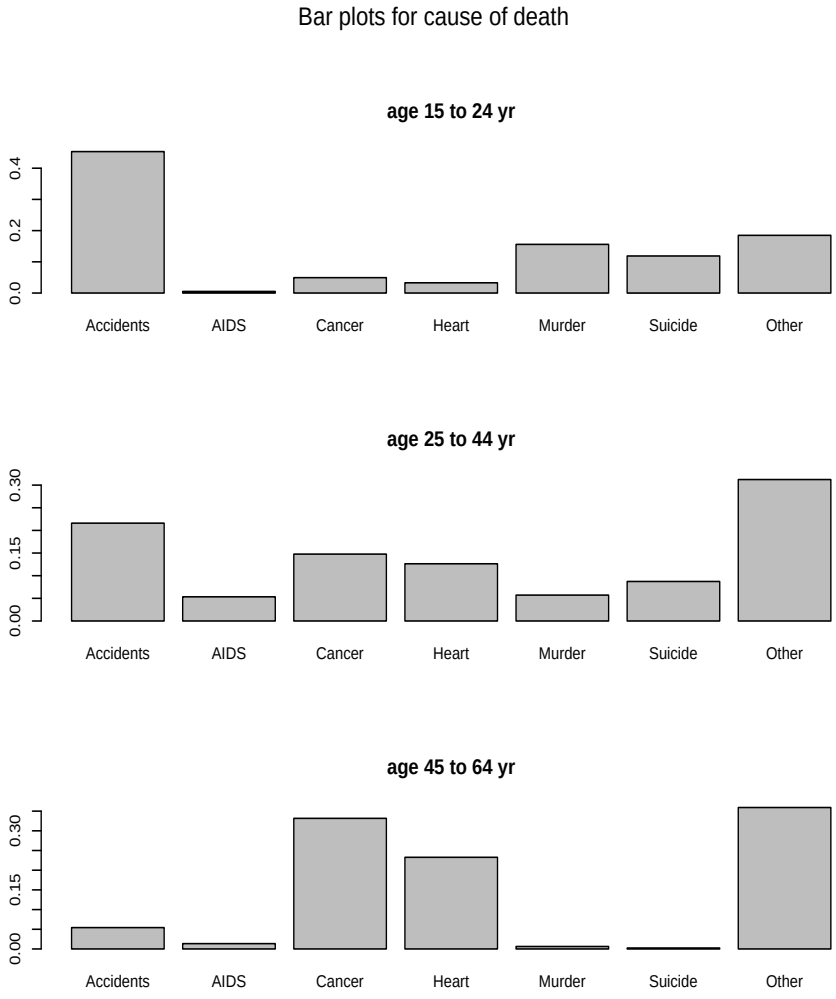

Figure 1: Bar plots of Death by Age Group

Simpson's Paradox

Table 3 gives data for graduate school admissions at the University of California at Berkeley for the 6 largest departments at the time. The totals for these departments are also given. For male applicants 46% were admitted and for female applicants 30% were admitted (columns 3 and 6 respectively). Does this mean that the female applicants were being treated unfairly?

On the other hand when we look department by department, the percent admitted is approximately the same for all departments (differing by 3 percent or less) except for two departments : (i) A - where 62% males are admitted versus 89% females admitted and (ii) B - where 63% males are admitted versus 68% females admitted. Now it appears that females are admitted at the same or higher rates, exactly the opposite conclusion using only totals.

What is happening? Here applicants do not choose their departments for application at random. Upon examining Table 3 one sees that relatively few females apply to department A, with its higher admission rate, and relatively more apply to departments C and E with their lower admission rates. These assignments to groups can seriously affect conclusions if one does not take them into account.

Major	Male Applicants	Male admitted	Male percent	Female Applicants	Female admitted	Female percent
A	825	512	62	108	89	82
B	560	353	63	25	17	68
C	325	120	37	593	202	34
D	417	138	33	375	131	35
E	191	53	28	393	94	24
F	272	16	6	341	24	7
total	2590	1192	46	1835	557	30

Table 3: Berkeley Graduate School Admissions Data

See the text for a hypothetical and specially constructed example about road versus helicopter ambulances. In that example when one lumps together serious and less serious accidents, it appears that helicopter ambulance patients die at a much higher rate than those transported in road ambulances. Does this mean that it is more dangerous for a patient in terms of survival to use a helicopter ambulance or that there is some systemic problem with air helicopter ambulances?

The problem is that helicopter ambulance deliver a relatively higher proportion of the most serious patients and relatively a smaller proportion of the less serious patients as compared with road ambulances. When these two types of patients are separated for individual type analyzes one finds that 52% of the serious patients survive using the helicopter ambulance versus 40% survival for serious patients using road ambulances. This is the opposite conclusion one would reach without considering how the data is collected and put into a group.

In both these examples pooling data can be very misleading. This make reading official reports and newspaper or internet new reports very difficult to interpret if one does not know the source or data types used to construct these. Scientific experimentation is done in a manner to avoid this. Observational studies, which occur in many social sciences and epidemiology, need to be very careful in the data collection methods to try to avoid these issues.

Simpson's Paradox : *An association or comparison that holds for all of several groups can reverse direction when the data are combined or pooled into a single group. The reversal is called Simpson's Paradox.*

Simpson's paradox is named after E. H. Simpson (1951) *Journal of the Royal Statistical Society (Series B)*, 13: 238241. It was known earlier for some examples (G. H. Yule (1903) *Biometrika*, 2: 121134), the more general understanding is attached to Simpson.

Conditional and Marginal Distributions

We now return to Table 2. Consider just the first column. This gives the (observed) distribution amongst the causes of death for the age group 15 to 24. Since this distribution is conditional on the person being in this particular age group, it is also known as a *conditional distribution*. Similarly columns 2 and 3 give the conditional distribution of causes of death for the two age groups 25-44 and 45-64 respectively.

There are two other distributions that are obtained from this table. These will be known as a *marginal distribution*. To motivate this concept we can ask

- amongst all deaths what proportion (what is the probability) that a death falls into each of the possible age groups? For example amongst the deaths what is the proportion that the death is in the 15-24 age group?
- given there is a death what is the proportion (or probability) that the death falls into a particular cause? For example amongst the deaths what is the proportion that the cause of death is cancer?

Below we view again the contents of Table 2. The first numerical column gives the *conditional*, upon being in age group 15-24, proportion (or probability) of the various causes of death within this age group. Thus we see that the probability of death by accident, conditional upon the death occurring in the 15-24 age group, is .454. The second column gives the conditional probability of the various causes of death, given the death occurs in the 25

to 44 age group.

cause	age 15-24	age 25-44	age 45-64
Accidents	0.453	0.216	0.054
AIDS	0.005	0.053	0.014
Cancer	0.049	0.148	0.332
Heart	0.033	0.126	0.233
Murder	0.156	0.057	0.006
Suicide	0.119	0.087	0.002
Other	0.185	0.312	0.359

Table 4 is the same as Table 1 except for an extra column at the end. It gives the totals for each row (actually the totals by row for the first 3 numerical columns), and the last element being the total number of observed deaths. Using this data we can calculate or estimate the *marginal distribution* of the cause of death. Notice the numbers in the bottom row of the table below add up to 0.999, not 1 as they should, again being due to round off error. The exact fractions in the second row do add up to 1,as there is no round off error at that stage of the calculations.

Accidents	AIDS	Cancer	Heart Disease	Murder	Suicide	Other
$\frac{66509}{599004}$	$\frac{12967}{599004}$	$\frac{165605}{599004}$	$\frac{119079}{599004}$	$\frac{15271}{599004}$	$\frac{16229}{599004}$	$\frac{203344}{599004}$
$= 0.111$	$= 0.022$	$= 0.276$	$= 0.199$	$= 0.025$	$= 0.027$	$= 0.339$

This type of information is useful for social resource planning and expenditure. It is also useful information for insurance companies.

Table 4: Causes of Death by Age Group

cause	15-24	25-44	45-64	deaths by cause
Accidents	14966	27844	23699	66509
AIDS	171	6879	5917	12967
Cancer	1628	19041	144936	165605
Heart disease	1083	16283	101713	119079
Murder	5148	7367	2756	15271
Suicide	3921	11251	1057	16229
Other	6105	40259	156980	203344
Total deaths	33022	128924	437058	599004