Section 9.1 - Data Analysis for Two-way Tables

Statistics 104

Autumn 2004

Data Analysis for Two-way Tables

Want to look at the breakdown of counts for two categorical variables.

Example: Berkeley Admissions Data

Major	Admitted	Rejected	Applied
A	600	333	933
B	370	215	585
C	322	596	918
D	269	523	792
E	148	436	584
F	46	668	714
Total	1755	2771	4526

Example: Aspirin Study

	Stroke	No Stroke	Total
Aspirin	15	63	78
Placebo	34	43	77
Total	49	106	155

Every observation must fit into exactly one cell of the table.

Often want to look at table of percentages (or proportions)

$$\frac{\#\text{in cell}}{\text{Total } \#\text{obs}} \times 100\%$$

For the Berkeley admissions data

Major	Admitted	Rejected
A	13.26	7.36
B	8.17	4.75
C	7.11	13.17
D	5.94	11.56
E	3.27	9.63
F	1.02	14.76

In addition there are a number of summaries of this table (or the table of counts) that people will look at

Marginal distributions

- Looks at only one of the two variables
- Get by adding across rows or down columns

Major	Admitted	Rejected	Total
A	13.26	7.36	20.61
B	8.17	4.75	12.93
C	7.11	13.17	20.28
D	5.94	11.56	17.50
E	3.27	9.63	12.90
F	1.02	14.76	15.78
Total	38.78	61.22	100

These are the data analogues to marginal probabilities.

Conditional distributions

- An approach to looking at each row or column separately

  Lets look at each major (row) separately

$$\frac{\text{\# admitted in program}}{\text{\# applied to program}} \times 100\%$$

$$\frac{\text{\# rejected in program}}{\text{\# applied to program}} \times 100\%$$

Major A

Conditional acceptance percentage

$$= \frac{600}{933} \times 100\% = 64.31\%$$

Conditional rejection percentage

$$= \frac{333}{933} \times 100\% = 35.69\%$$

Major	Admitted	Rejected	Applied	% Admit	% Reject
A	600	333	933	64.31	35.69
B	370	215	585	63.25	36.75
C	322	596	918	35.08	64.92
D	269	523	792	33.96	66.04
E	148	436	584	25.34	74.66
F	46	668	714	6.44	93.56
Total	1755	2771	4526	38.78	61.22

In the earlier analysis of the Aspirin study, we were looking at the conditional distribution of stroke and no stroke, conditional on the treatment given.

An important advantage of looking at conditional distributions is that it allows valid comparisons to be made.

The 600 admitted in major A is not directly comparable to the 370 admitted in major B since many more people applied to major A (933 vs 585).

Note that the rules for joint, conditional and marginal probabilities work with proportions based on two-way tables.

Note that tables can be extended to more than two categorical variables.

Example: Treating a deadly disease.

- 2 treatments: A & B
- Survival after 1 year
- Gender

	Men		Women	
	Trt A	Trt B	Trt A	Trt B
# Survived	48	27	8	42
# Died	72	33	32	126
Total Treated	120	60	40	168
% Survived	40	45	20	25

Now lets ignore gender, collapsing the three-way table to a two-way table

	Treatment A	Treatment B
# Survived	56	69
# Died	104	159
Total Treated	160	228
% Survived	35	30.26

So if we ignore gender, the data suggests that treatment A is better than treatment B, though not statistically significant.

```
. prtesti 160 56 228 69 , count

Two-sample test of proportion                    x: Number of obs = 160
                                                 y: Number of obs = 228

------------------------------------------------------------------------------
Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |        .35   .0377078                      .2760942    .4239058
       y |   .3026316   .0304243                       .243001    .3622621
---------+--------------------------------------------------------------------
    diff |   .0473684   .0484512                     -.0475941    .1423309
         |  under Ho:   .0481936     0.98   0.326
------------------------------------------------------------------------------

               Ho: proportion(x) - proportion(y) = diff = 0

       Ha: diff < 0            Ha: diff != 0              Ha: diff > 0
        z =  0.983               z =  0.983                z =  0.983
    P < z =  0.8372        P > |z| =  0.3257           P > z =  0.1628
```

However if we include gender in our analysis, the data suggests that treatment B is better for both men and women. (not statistically significant either)

The is an example of **Simpson's Paradox**

> A reversal of the direction of a comparison or an association when data from several groups are aggregated (combined) to form a single group.

Why the reversal in direction?

- Men's survival is better than women's
- Many more men get treatment A than treatment B
- Women tend to get treatment B instead of treatment A

- When looking at data when ignoring gender, the apparent superiority of treatment A is an artifact of the men's better survival rate.

- In the analysis ignoring gender, gender is a lurking variable which is confounded with treatment.

Back to the Fall 1973 Berkeley Admissions data

	Men	Women
# Admitted	1193	557
# Rejected	1494	1278
# Applied	2691	1835
% Admitted	44.52	30.35

Here is the three-way table of admittance for men and women for each major

	Men		Women	
Major	# Admitted	# Rejected	# Admitted	# Rejected
A	511	314	89	19
B	353	207	17	8
C	120	205	202	391
D	138	279	131	244
E	54	137	94	299
F	22	351	24	317

The conditional acceptance rates for men and women for each major

	Men		Women	
Major	% Admitted	% Rejected	% Admitted	% Rejected
A	61.94	38.06	82.41	17.59
B	63.04	36.96	68.00	32.00
C	36.92	63.08	34.06	65.94
D	33.09	66.91	34.93	65.07
E	28.27	71.73	23.92	76.08
F	5.90	94.10	7.04	92.96

In the straight comparison of men vs women, major is a lurking variable. Men are much more likely to apply to majors A and B, which are much easier to get into., Women, however, are more likely to apply to other majors, which are much harder to get into.

Conditional distribution of program applied to for each gender

	A	B	C	D	E	F
Men	30.66	20.81	12.08	15.50	7.10	13.86
Women	5.89	1.36	32.32	20.44	21.42	18.58