

Statistics 305/605: Introduction to Biostatistical Methods for Health Sciences

Chapter 15, part 1: Contingency Tables

Jinko Graham

2018-10-03

Contingency Tables

- ▶ Contingency tables display the joint frequency distribution of two categorical variables.
- ▶ E.G.: Let's consider the data of Mungan et al. 2000 click on 21,737 bladder cancer patients
 - ▶ Two categorical variables: **gender**, which has 2 levels, and **cancer stage**, which has 4 levels.
 - ▶ The first few lines of the data file are as follows:

##	Gender	Cancer.Stage
## 1	Male	I
## 2	Male	I
## 3	Male	I
## 4	Male	I
## 5	Male	I
## 6	Male	I
## 7	Male	I
## 8	Male	I

- ▶ The contingency table made by cross-tabulating the gender and cancer stage variables of the Mungan data is as follows:

##		Cancer.Stage			
##	Gender	I	II	III	IV
##	Female	3926	402	356	852
##	Male	12418	995	883	1905

Terminology: Cells of a Table

- ▶ The cells of the table are its entries.
 - ▶ In the table cross-tabulating the gender and cancer stage variables of the Mungan data, the first cell of the table is 3926

##		Cancer.Stage			
##	Gender	I	II	III	IV
##	Female	3926	402	356	852
##	Male	12418	995	883	1905

Terminology: Row and Column Variables

- ▶ The row variable in a table defines the rows, the column variable the columns.
 - ▶ In the table below, the row variable is Gender and the column variable is Cancer.Stage.

##		Cancer.Stage			
##	Gender	I	II	III	IV
##	Female	3926	402	356	852
##	Male	12418	995	883	1905

Terminology: Row and Column Margins

- ▶ The **row margin** is the tabulation of the row variable and the **column margin** is the tabulation of the column variable.
- ▶ For the Mungan data,
 - ▶ the row margin (tabulation of Gender) is 5536 and 16201 Females and Males, respectively
 - ▶ The column margin (tabulation of Cancer Stage) is 16344, 1397, 1239, 2757 for cancer stages I through IV, respectively.
- ▶ Exercise: verify these table margins yourself.

Adding Margins to a Table

- ▶ It is common practice to add margins to a contingency table.
- ▶ In the following, the row margins (first table) and column margins (second table) have been added:

##		I	II	III	IV	Total
##	Female	3926	402	356	852	5536
##	Male	12418	995	883	1905	16201

##		I	II	III	IV
##	Female	3926	402	356	852
##	Male	12418	995	883	1905
##	Total	16344	1397	1239	2757

Cancer stage distribution given gender

```
##           Cancer.Stage
## Gender           I      II      III      IV
##   Female    3926    402    356    852
##   Male     12418    995    883   1905
```

- For each gender category, we can divide the counts in each row by the row total to get proportions.

```
##           Cancer.Stage
## Gender           I           II           III           IV
##   Female  0.70917630 0.07261561 0.06430636 0.15390173
##   Male    0.76649590 0.06141596 0.05450281 0.11758533
```

- This gives an estimate of the conditional distributions of cancer stage in each gender.

Gender distribution given cancer stage

- Likewise, for each cancer stage category we can divide the counts in each column by the column total to get proportions.

```
##           Cancer.Stage
## Gender           I           II           III           IV
##   Female 0.2402105 0.2877595 0.2873285 0.3090316
##   Male   0.7597895 0.7122405 0.7126715 0.6909684
```

- This gives an estimate of the conditional distributions of gender in each cancer stage.

Independence of Row and Column Variables.

- ▶ If the conditional distributions of the row variable given the column variable are all the same, they will also be the same as the overall distribution of the row variable.
 - ▶ E.G., if the conditional gender distribution is 25% female and 75% male in each cancer stage, we will have 25% females and 75% males overall too.
- ▶ When this happens we say that the column and row variables are *independent* because:
 - ▶ Knowing the value of the column variable tell us nothing about the row variable;
 - ▶ E.G. Knowing cancer stage tells us nothing about gender; so $P(\text{Gender} = \text{Female} \mid \text{Stage} = \text{I}) = P(\text{Gender} = \text{Female})$

- ▶ One can use the definition of conditional probability to show that independence of row and column variables is equivalent to the following two statements:
 1. The conditional distributions given the different levels of the row variable are all equal
 2. The conditional distributions given the different levels of the column variable are all equal.
- ▶ The opposite of independence is dependence, or an *association*.
- ▶ We next discuss how to test for association.