STAT 210
Applied Statistics and Data Analysis
Contingency Tables

Joaquin Ortega

Fall 2020

Statistical analysis of contingency tables

Contingency tables

Contingency tables are often the starting point of statistical analysis.

In this section, we will consider the problem of determining whether the distribution of a certain variable A is related to the value of another variable B, or in a more technical language, whether the conditional distribution of A given the value of B is the same for all values of B.

The following table has the information we will be analyzing that relates gender to survival in the Titanic disaster.

The data are in the package `vcd`.

Contingency tables

```r
library(vcd)
data("Titanic")
str(Titanic)
```

```
##  'table' num [1:4, 1:2, 1:2, 1:2] 0 0 35 0 0 0 17 0 118
##  - attr(*, "dimnames")=List of 4
##   ..$ Class   : chr [1:4] "1st" "2nd" "3rd" "Crew"
##   ..$ Sex     : chr [1:2] "Male" "Female"
##   ..$ Age     : chr [1:2] "Child" "Adult"
##   ..$ Survived: chr [1:2] "No" "Yes"
```

Contingency tables

```
(titanic.table <- apply(Titanic, c(4, 2), sum))
```

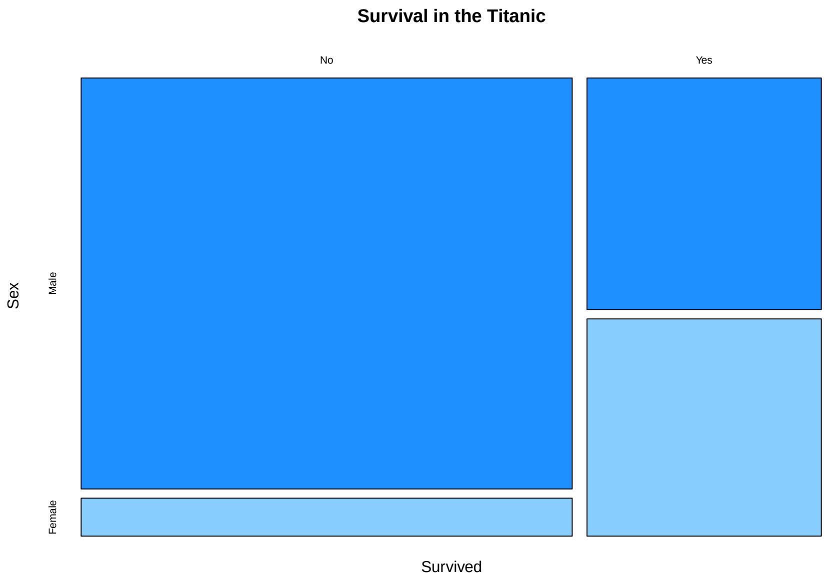
```
##          Sex
## Survived Male Female
##      No  1364    126
##      Yes  367    344
```

```
(titanic.table <- addmargins(titanic.table))
```

```
##          Sex
## Survived Male Female  Sum
##      No  1364    126 1490
##      Yes  367    344  711
##      Sum 1731    470 2201
```

Contingency tables

```
mosaicplot(titanic.table[1:2,1:2],
           col = c('dodgerblue','skyblue1'),
           main = 'Survival in the Titanic')
```



Contingency tables

```r
kable(titanic.table, 'latex', booktabs = T,
      caption = 'Observed values') %>%
  kable_styling(bootstrap_options = "hover",
                full_width = F, position = "center",
                latex_options ='striped')
```

Table 1: Observed values

	Male	Female	Sum
No	1364	126	1490
Yes	367	344	711
Sum	1731	470	2201

Contingency tables

With this information, we want to explore if survival is related to gender.

The proportion of surviving individuals in the male population is

$$\pi_1 = \frac{367}{1731} = 0.212$$

while for the female population, it is

$$\pi_2 = \frac{344}{470} = 0.732$$

We want to test

$$H_0 : \pi_1 = \pi_2 \quad \text{vs} \quad H_A : \pi_1 \neq \pi_2.$$

Contingency tables

In general, the simple case of a 2×2 contingency table can be described as follows: We have two populations or groups, and we want to study whether the presence of some characteristic occurs in the same proportion.

Let us call `P1` and `P2` the two populations and p_1 and p_2 the proportions of the given trait in each of them.

We take samples of sizes n_1 (from `P1`) and n_2 (from `P2`) and $s_i, i = 1, 2$ represent how many trials in each sample were successful, i.e., the individuals have the characteristic.

With these results, we build a contingency table.

Contingency tables

Table 2: Observed values

	P1	P2	Total
Success	s_1	s_2	s
Failure	$n_1 - s_1$	$n_2 - s_2$	$n - s$
Total	n_1	n_2	n

Here $s = s_1 + s_2$ is the total number of successes.

Let $d = p_1 - p_2$. We want to use the information in the table to test

$$H_0 : d = 0 \qquad \text{vs} \qquad H_A : d \neq 0$$

Contingency tables

Use the data to estimate the proportions:

$$\pi_1 = \frac{s_1}{n_1}, \qquad \pi_2 = \frac{s_2}{n_2}.$$

Under the null hypothesis $p_1 = p_2 = p$.

To estimate p, pool all the information:

$$\pi = \frac{n_1}{n}\pi_1 + \frac{n_2}{n}\pi_2 \Big(= \frac{s_1 + s_2}{n} \Big)$$

If p is the true proportion for both samples, we would expect to have $n_i \times p$ successes and $n_i \times (1 - p)$ failures in sample $i = 1, 2$.

Use π instead of p and create a table of expected values.

Contingency tables

How many successes do we **expect** in each population?

	P1	P2	Total
Success			
Failure			
Total	n_1	n_2	n

Contingency tables

How many successes do we **expect** in each population?

	P1	P2	Total
Success	$\pi \times n_1$	$\pi \times n_2$	
Failure			
Total	n_1	n_2	n

Contingency tables

How many successes do we **expect** in each population?

	P1	P2	Total
Success	$\pi \times n_1$	$\pi \times n_2$	$\pi \times n$
Failure			
Total	n_1	n_2	n

Contingency tables

How many failures do we **expect** in each population?

	P1	P2	Total
Success	$\pi \times n_1$	$\pi \times n_2$	$\pi \times n$
Failure	$(1 - \pi) \times n_1$	$(1 - \pi) \times n_2$	
Total	n_1	n_2	n

Contingency tables

How many failures do we **expect** in each population?

	P1	P2	Total
Success	$\pi \times n_1$	$\pi \times n_2$	$\pi \times n$
Failure	$(1 - \pi)\times n_1$	$(1 - \pi)\times n_2$	$(1 - \pi)\times n$
Total	n_1	n_2	n

Contingency tables

Compare expected values with observed.

If the difference is large, we will question the null hypothesis.

The statistic for Pearson's test is

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where O stands for observed, and E for expected and the sum runs over all cases.

Under the null hypothesis, this statistic has approximately a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom, where r and c stand for the number of rows and columns of the table.

Contingency tables

Since the χ^2 distribution is continuous and our data discrete, there is a continuity correction for this statistic, due to Yates,

$$\chi^2 = \sum \frac{(|O - E| - 0.5)^2}{E}$$

The pooled estimate for p in our example is

$$\frac{711}{2201} = 0.323$$

and the table of expected values is

Contingency tables

	Male	Female	Total
No			
Yes			
Total	1731	470	2201

Contingency tables

	Male	Female	Total
No	1731×0.677	470×0.677	1490
Yes	1731×0.323	470×0.323	711
Total	1731	470	2201

Contingency tables

	Male	Female	Total
No	1171.89	318.19	1490
Yes	559.11	151.81	711
Total	1731	470	2201

Contingency tables

Expected values

	Male	Female	Total
No	1171.89	318.19	1490
Yes	559.11	151.81	711
Total	1731.00	470.00	2201

Observed values

	Male	Female	Sum
No	1364	126	1490
Yes	367	344	711
Sum	1731	470	2201

Contingency tables

To calculate the test statistic we can take advantage of vectorial calculations in R:

```
(chi.st <- sum((tab.exp[1:2,1:2]
       -titanic.table[1:2,1:2])^2/tab.exp[1:2,1:2]))
```

```
## [1] 456.8973
```

The associated p-value is given by

```
1-pchisq(chi.st,1)
```

```
## [1] 0
```

Contingency tables

In R:

```
chisq.test(titanic.table[1:2,1:2],correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  titanic.table[1:2, 1:2]
## X-squared = 456.87, df = 1, p-value < 2.2e-16
```

```
chisq.test(titanic.table[1:2,1:2])
```

```
##
##  Pearson's Chi-squared test with Yates' continuity corre
##
## data:  titanic.table[1:2, 1:2]
## X-squared = 454.5, df = 1, p-value < 2.2e-16
```

Contingency Tables: Independence

Contingency Tables: Independence

The χ^2 test can also be used to test for independence of categorical variables in contingency tables.

Consider as an example the set `survey` in the `MASS` package that has the responses of 237 Statistics students to a series of questions.

We consider

- `Smoke`,

a factor with four levels: `Heavy`, `Regul` (regularly), `Occas` (occasionally), `Never`, and

- `Exer`,

how frequently the student exercises, with levels `Freq` (frequently), `Some`, `None`.

Contingency Tables: Independence

We use `table` to produce the contingency table for these two variables.

```r
library(MASS)
(stdt.tab <- with(survey,table(Smoke,Exer)))
```

```
##        Exer
## Smoke   Freq None Some
##   Heavy    7    1    3
##   Never   87   18   84
##   Occas   12    3    4
##   Regul    9    1    7
```

Contingency Tables: Independence

Add totals

```
stdt.tot <- cbind(stdt.tab,
                  Total = apply(stdt.tab, 1, sum))
(stdt.tot <- rbind(stdt.tot,
                  Total = apply(stdt.tot, 2, sum)))
```

```
##       Freq None Some Total
## Heavy    7    1    3    11
## Never   87   18   84   189
## Occas   12    3    4    19
## Regul    9    1    7    17
## Total  115   23   98   236
```

Contingency Tables: Independence

We want to compare (categorical) variables X and Y with values

$$x_1, \dots, x_m \quad \text{and} \quad y_1, \dots, y_n$$

and probability functions

$$p_1, \dots, p_m \quad \text{and} \quad q_1, \dots, q_n.$$

If the variables are independent

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j) = p_i q_j$$

for any $1 \leq i \leq m, 1 \leq j \leq n$.

If the total sample is of size N, we would expect

$$N p_i q_j$$

individuals to be in the ij-th cell of the contingency table.

Contingency Tables: Independence

Since p_i and q_j are unknown, we estimate them by the corresponding proportions.

Use the row totals divided by N to estimate the p_i's and the column totals divided by N to estimate the q_j's.

Let n_{ij} be the number in the ij-th cell for $1 \le i \le m, 1 \le j \le n$. Introduce the notation:

$$n_{\bullet j} = \sum_{i=1}^{m} n_{ij} \qquad n_{i\bullet} = \sum_{j=1}^{n} n_{ij}$$

$$n_{\bullet\bullet} = \sum_{i=1}^{m} \sum_{j=1}^{n} n_{ij} = N.$$

Contingency Tables: Independence

Then

$$\hat{p}_i = \frac{n_{i\bullet}}{n_{\bullet\bullet}}, \qquad \hat{q}_j = \frac{n_{\bullet j}}{n_{\bullet\bullet}}$$

The expected value for the number in the ij-th cell is

$$E_{ij} = N\hat{p}_i\hat{q}_j = n_{\bullet\bullet}\frac{n_{i\bullet}}{n_{\bullet\bullet}}\frac{n_{\bullet j}}{n_{\bullet\bullet}} = \frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}.$$

We use the same statistic as before

$$\chi^2 = \sum_{ij}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O stands for observed, E for expected, and the sum runs over all cases.

Contingency Tables: Independence

This statistic has a χ^2_ν distribution with

$$\nu = (m-1)(n-1)$$

degrees of freedom.

```r
chisq.test(stdt.tab)
```

```
## Warning in chisq.test(stdt.tab): Chi-squared approximat:

##
##  Pearson's Chi-squared test
##
## data:  stdt.tab
## X-squared = 5.4885, df = 6, p-value = 0.4828
```

Fisher's exact test

Small samples: Fisher's exact test

The Chi-square distribution approximation requires that the expected value for each cell be at least 5. When this is not satisfied, results can be incorrect.

Under the assumption that the margins (totals) in the contingency table are fixed, it is possible to calculate an exact value for the significance of the deviation from the null hypothesis.

Fisher's exact test is mostly used for 2×2 tables and small samples, but in principle can be extended to general contingency tables, although for large tables, the calculation may be complicated.

For 2×2 tables, the calculation uses the hypergeometric distribution.

Hypergeometric distribution

Consider a population of size N with K individuals of type A.

The probability that in a sample of size $n \leq N$ there are precisely $k \leq K$ individuals of type A when sampling **without replacement** is given by the **hypergeometric distribution**

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

for $1 \leq n \leq N$ and $0 \leq k \leq K \leq N$. Recall that

$$\binom{a}{b} = \frac{a!}{b!(a-b)!}.$$

Hypergeometric distribution

	Pop1	Pop2	Total
Type A			a+b
Not Type A			c+d
Total	a+c	b+d	n=a+b+c+d

Hypergeometric distribution

	Pop1	Pop2	Total
Type A	a		a+b
Not Type A			c+d
Total	a+c	b+d	n=a+b+c+d

Hypergeometric distribution

	Pop1	Pop2	Total
Type A	a	b	a+b
Not Type A	c	d	c+d
Total	a+c	b+d	n=a+b+c+d

Hypergeometric distribution

	Pop1	Pop2	Total
Type A	a	b	a+b
Not Type A	c	d	c+d
Total	a+c	b+d	n=a+b+c+d

Population has size `n`

Type A in population `a+b`

Sample has size `a+c`

Hypergeometric distribution

Probability:

$$\frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!}$$

Small samples: Fisher's exact test

Left- and right-handedness data

```r
fisher.test(titanic.table[1:2,1:2])
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  titanic.table[1:2, 1:2]
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to
## 95 percent confidence interval:
##   7.97665 12.92916
## sample estimates:
## odds ratio
##    10.1319
```

Small samples: Fisher's exact test

Student data

```
fisher.test(stdt.tab)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  stdt.tab
## p-value = 0.4138
## alternative hypothesis: two.sided
```