

STAT 210  
Applied Statistics and Data Analysis  
Week 7 - Summary

Joaquin Ortega

King Abdullah University of Science and Technology

All groups have to meet with me before submitting the project proposal.

The deadline for submitting the project proposal is October 23.

The first exam will be on Saturday, October 25, 9:00 - 12:00 am, Room 2322.

## Video 24: Tables in R

Tabular summaries of data are a frequent starting point for statistical analysis.

A **contingency table** in Statistics is a table that displays the multivariate frequency distribution of two or more variables. The name was first used in 1904 by Karl Pearson.

The entries in the cells of a two-way table are either the frequency counts or the relative frequencies, and the table is usually presented as a matrix.

Before looking at the usual statistical techniques for analysis of contingency tables, let us review some of the available tools for producing them.

# Functions for Producing Tables

The function `cut()` divides observations according to the values of a continuous variable.

```
cut(data, breaks, labels = NULL, right = TRUE)
```

- `breaks` defines the break points of each level or class
- `labels` specifies the value to use when an observation falls in one class
- `right` specifies the type of interval: `right = F` is for  $[a, b)$  and `right = T` is for  $(a, b]$ .

## Functions for Producing Tables

The function `table()` builds a contingency table of the counts at each combination of factor levels.

```
(table1 <- with(mtcars, table(cyl, gear)))
```

```
##      gear
## cyl   3   4   5
##   4   1   8   2
##   6   2   4   1
##   8  12   0   2
```

The function `prop.table()` produces tables of relative frequencies.

```
prop.table(data, margin)
```

# Functions for Producing Tables

```
(table2 <- round(prop.table(table1,1),3))
```

```
##      gear
## cyl      3      4      5
##   4 0.091 0.727 0.182
##   6 0.286 0.571 0.143
##   8 0.857 0.000 0.143
```

```
rowSums(table2)
```

```
## 4 6 8
## 1 1 1
```

```
(table3 <- round(prop.table(table1,2),3))
```

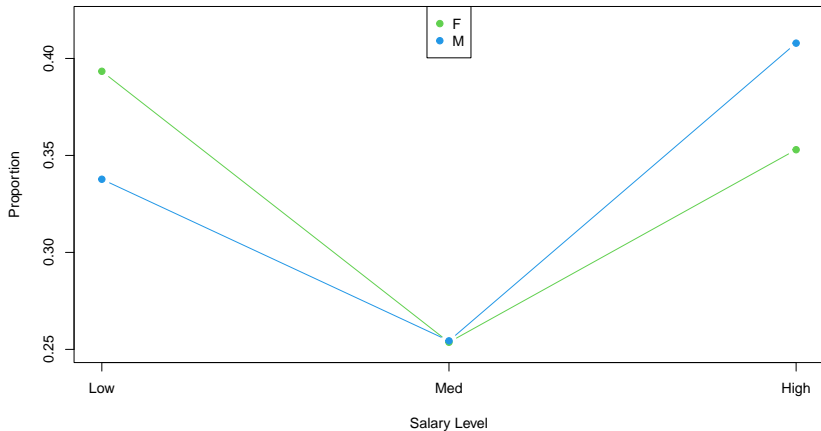
```
##      gear
## cyl      3      4      5
##   4 0.067 0.667 0.400
##   6 0.133 0.333 0.200
##   8 0.800 0.000 0.400
```

```
colSums(table3)
```

```
## 3 4 5
## 1 1 1
```

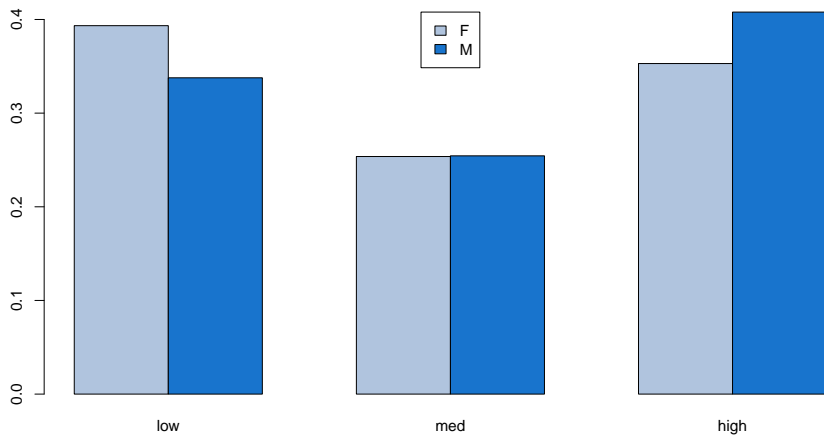
# Graphical Representations

Proportion of individuals in each income level





# Graphical Representations



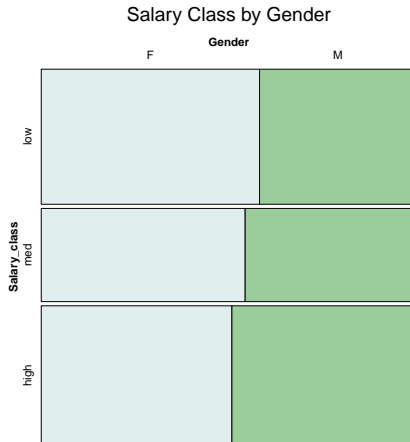
# Mosaic Plots

```
mosaicplot(freq_gender_sal1, xlab='Salary class', ylab='gender',  
            main='Salary Class by Gender', col=c('azure2','darkseagreen3'))
```



# Mosaic Plots

```
library(vcd)
mosaic(~ Salary_class + Gender, data = humans,
       main = 'Salary Class by Gender', highlighting = 'Gender',
       highlighting_fill = c('azure2', 'darkseagreen3'))
```



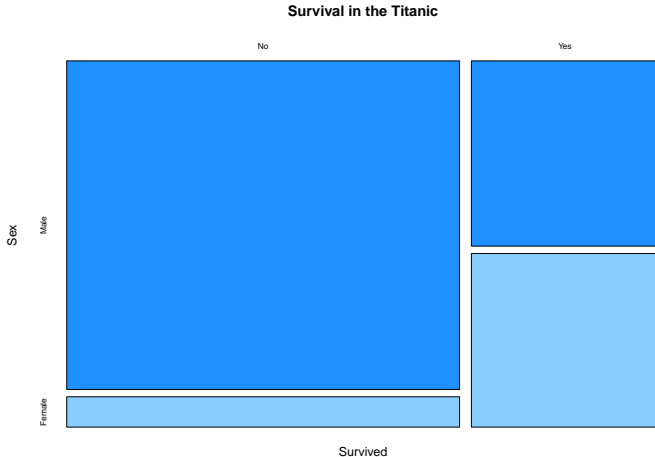
## Video 25: Contingency Tables

```
library(vcd)
data("Titanic")
titanic.table <- apply(Titanic, c(4, 2), sum)
(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')
```



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

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

In general, the simple case of a  $2 \times 2$  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  $P_1$  and  $P_2$  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  $P_1$ ) and  $n_2$  (from  $P_2$ ) 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.



Table 1: 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

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

Use the data to estimate the proportions:

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

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

To estimate  $p$ , pool all the information:

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

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  $\pi$  instead of  $p$  and create a table of expected values.

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

	P1	P2	Total
Success			
Failure			
Total	$n_1$	$n_2$	$n$

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

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

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	$n_1$	$n_2$	$n$

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	$n_1$	$n_2$	$n$

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	$n_1$	$n_2$	$n$

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  $\chi^2$  distribution with  $(r - 1)(c - 1)$  degrees of freedom, where  $r$  and  $c$  stand for the number of rows and columns of the table.



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 correction
```

```
##
```

```
## data:  titanic.table[1:2, 1:2]
```

```
## X-squared = 454.5, df = 1, p-value < 2.2e-16
```

# Contingency Tables: Independence

The  $\chi^2$  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.

```
library(MASS)
stdt.tab <- with(survey, table(Smoke, Exer))
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 size is  $N$ , we would expect

$$Np_i q_j$$

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  $1 \leq i \leq m, 1 \leq j \leq n$ . Introduce the notation:

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

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

## Contingency Tables: Independence

Then

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

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  $\chi^2_\nu$  distribution with

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

degrees of freedom.

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

```
## Warning in chisq.test(stdt.tab): Chi-squared approximation may be in
##
## Pearson's Chi-squared test
##
## data:  stdt.tab
## X-squared = 5.4885, df = 6, p-value = 0.4828
```

# Chi-Square Distribution

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.

```
chisq.test(stdt.tab)$expected
```

```
## Warning in chisq.test(stdt.tab): Chi-squared approximation may be in
```

```
##           Exer
```

## Smoke	Freq	None	Some
## Heavy	5.360169	1.072034	4.567797
## Never	92.097458	18.419492	78.483051
## Occas	9.258475	1.851695	7.889831
## Regul	8.283898	1.656780	7.059322



## Fisher's Exact Test

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  $2 \times 2$  tables and small samples, but in principle can be extended to general contingency tables, although for large tables, the calculation may be complicated.

For  $2 \times 2$  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

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

# Small Samples: Fisher's Exact Test

## Titanic data

```
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 1  
## 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
```

## Video 26: Comparing Proportions

# The Binomial Distribution

The distribution of the number of individuals of type  $A$ ,  $n_A$ , in the sample is binomial with parameters  $n$  and  $p$ :

$$P(n_A = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

The expected value and variance for this distribution are given by

$$E(n_A) = np, \quad \text{Var}(n_A) = np(1 - p).$$

A natural estimator for the (unknown) proportion  $p$  is the observed proportion of individuals of type  $A$  in the sample:

$$\pi = \frac{n_A}{n}.$$

It is unbiased with variance

$$\text{Var}(\pi) = \frac{1}{n^2} \text{Var}(n_A) = \frac{p(1 - p)}{n}.$$

# The Normal Approximation

By the Central Limit Theorem, for  $n$  large, the binomial distribution can be approximated by a normal distribution.

## Rule of Thumb

If  $n$  and  $p$  are such that  $np \geq 5$  and  $n(1 - p) \geq 5$ , the binomial distribution can be approximated by the normal distribution.

Thus, if  $np \geq 5$  and  $n(1 - p) \geq 5$  the sampling density for the (sample) proportion  $\pi$  can be approximated by a normal distribution with parameters

$$p \quad \text{and} \quad \frac{p(1 - p)}{n}.$$

## One-sample problem for proportions.

The following data come from Kaye, D.H., *Statistical Evidence of Discrimination*, JASA (1982).

In a case about discrimination against blacks in grand jury selection in Alabama, the plaintiff argued that of the 1050 individuals called to serve as jurors, only 177 were black.

At the time, 25% of those eligible to serve were blacks.

**Do the data support the claim of discrimination?** We want to test

$$H_0 : p = 0.25 \quad \text{vs} \quad p < 0.25$$

and choose a level  $\alpha = 0.01$ .



```
prop.test(n.A,n,p_0)
```

```
##  
## 1-sample proportions test with continuity correction  
##  
## data:  n.A out of n, null probability p_0  
## X-squared = 36.698, df = 1, p-value = 1.379e-09  
## alternative hypothesis: true p is not equal to 0.25  
## 95 percent confidence interval:  
##  0.1466952 0.1929145  
## sample estimates:  
##           p  
## 0.1685714
```

```
binom.test(n.A,n,p_0)
```

```
##  
## Exact binomial test  
##  
## data:  n.A and n  
## number of successes = 177, number of trials = 1050, p-value = 2.454e-10  
## alternative hypothesis: true probability of success is not equal to 0.25  
## 95 percent confidence interval:  
##  0.1464049 0.1926129  
## sample estimates:  
## probability of success  
##                0.1685714
```

## Two Independent Proportions

## Two Independent Proportions

Assume now that we have two samples of sizes  $n_1$  and  $n_2$ , respectively, with number of successes  $m_1$  and  $m_2$ . The corresponding proportions are

$$\pi_i = \frac{m_i}{n_i}$$

for  $i = 1, 2$ , and we want to compare these two values.

We want to test

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

The normal approximation requires

$$n_i \times \pi_i \geq 5, \quad n_i \times (1 - \pi_i) \geq 5$$

for  $i = 1, 2$ .

The test can be carried out using `prop.test`.

The following matrix corresponds to the number of patients involved in car accidents that survived or died. The use of seat belts is also reported in the data, which were registered at a hospital in North Carolina.

```
car.accidents <- data.frame(survived = c(1781,1443),  
                             died=c(135,47))  
rownames(car.accidents) <- c('nsb','sb')  
car.accidents
```

```
##      survived died  
## nsb      1781  135  
## sb       1443   47
```

To test whether the use of seat belts affected the rates of survival we compare the proportions using the function `prop.test`

```
prop.test(as.matrix(car.accidents))

##
## 2-sample test for equality of proportions with continuity correction
##
## data:  as.matrix(car.accidents)
## X-squared = 24.333, df = 1, p-value = 8.105e-07
## alternative hypothesis: two.sided
## 95 percent confidence interval:
## -0.05400606 -0.02382527
## sample estimates:
##      prop 1      prop 2
## 0.9295407 0.9684564
```

## Problem List 6

- Problem 1: Diana Katherine Condor