Introductory Statistics with R
(Chap. 7~8)

Seungyeon Seo

Introduction

7. Analysis of variance and the Kruskal-Wallis test

- One-way analysis of variance
- Kruskal–Wallis test
- Two-way analysis of variance
- The Friedman test
- The ANOVA table in regression analysis

8. Tabular data

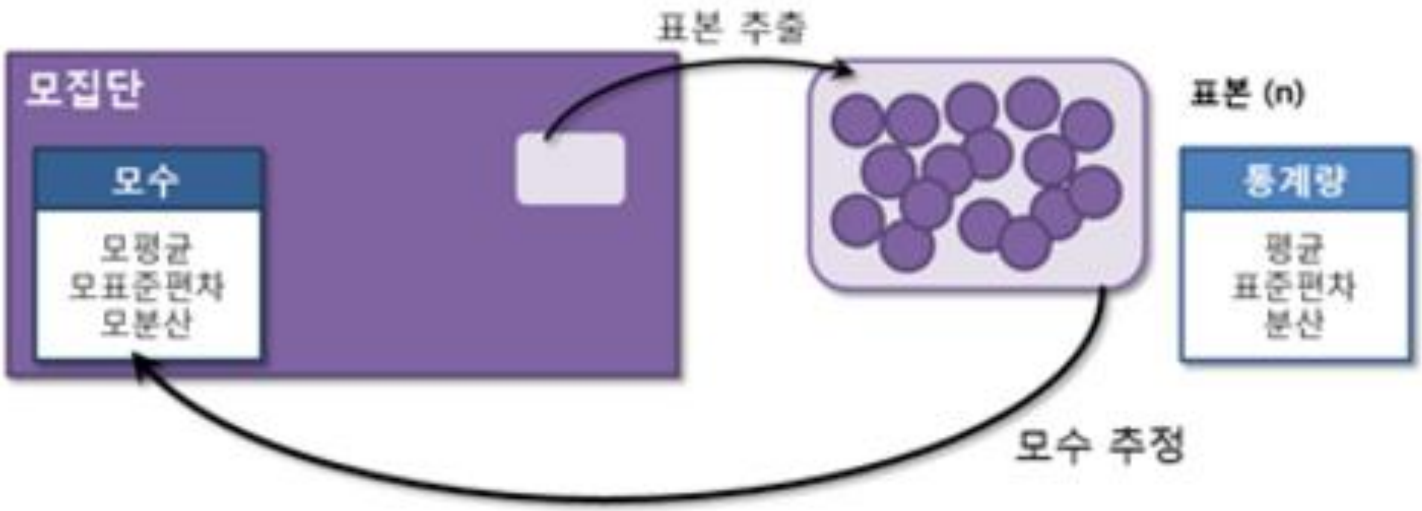
- Single proportions
- Two independent proportions
- k proportions test for trend
- r × c tables

7. Analysis of variance and the Kruskal-Wallis test

Parametric vs. Nonparametric Test

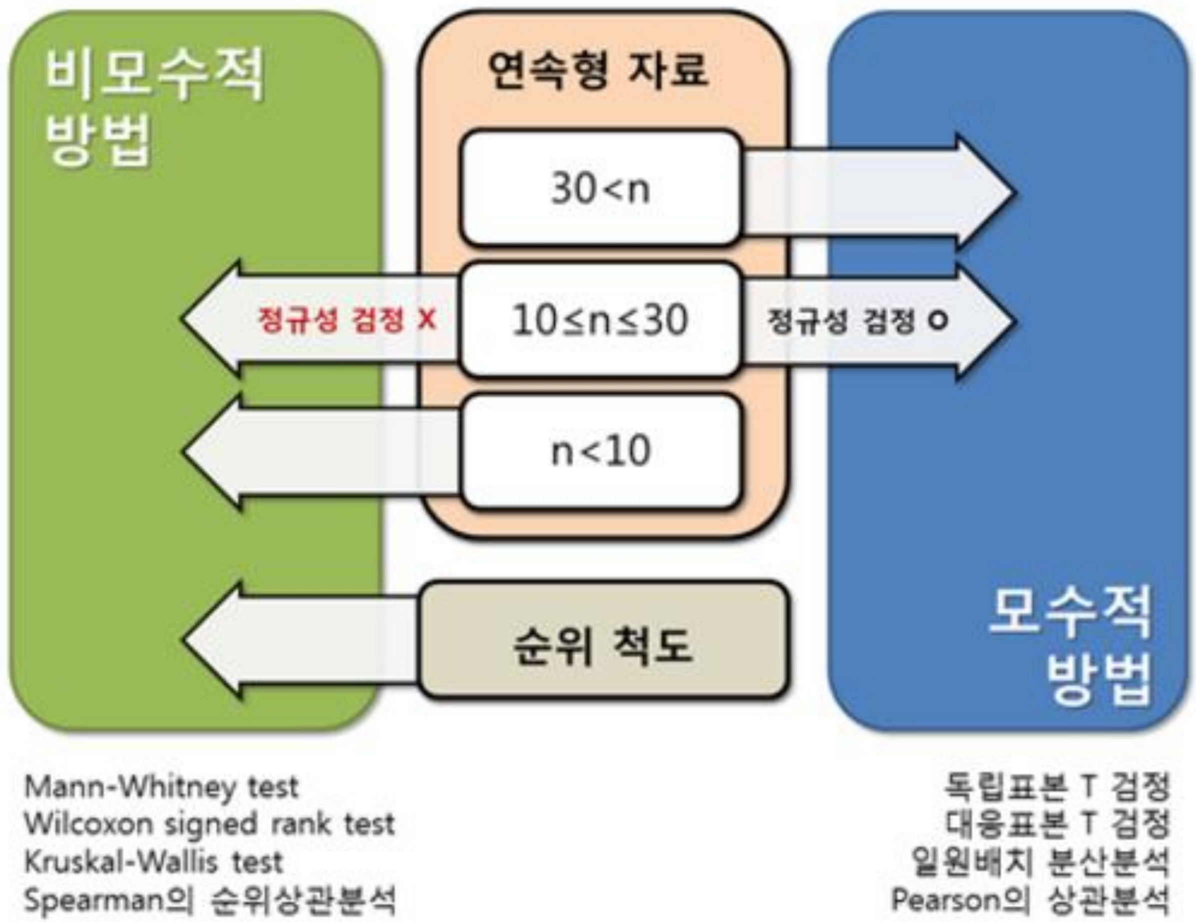
		Nonparametric test	Parametric test
Requirements		Non-normal distribution Un-known distribution Very small sample size or ranked data	Normal distribution
Statistics		Median	Mean
Group	**1**	1 sample Wilcoxon signed rank test	1 sample t-test
	2	Mann-Whitney test	2 sample t-test
		Wilcoxon signed rank test	Paired 2-sample t-test
	More than 2	Kruskal-Wallis test	One-way analysis of variance

Statistics vs. Parameter



http://dermabae.tistory.com/147

Parametric vs. Nonparametric Test



http://dermabae.tistory.com/147

Parametric vs. Nonparametric Test

		Nonparametric test	**Parametric test**
Requirements		Non-normal distribution Un-known distribution Very small sample size or ranked data	Normal distribution
Statistics		Median	Mean
Group	**1**	1 sample Wilcoxon signed rank test	1 sample t-test
	2	Mann-Whitney test	2 sample t-test
		Wilcoxon signed rank test	Paired 2-sample t-test
	More than 2	Kruskal-Wallis test	One-way analysis of variance

Analysis of Variance and the Kruskal-Wallis Test

- ❖ One-way analysis of variance
- ❖ Kruskal–Wallis test
- ❖ Two-way analysis of variance
- ❖ The Friedman test
- ❖ The ANOVA table in regression analysis

Analysis of variance (ANOVA)

- ❖ 목적: 두 개 이상 다수의 집단을 비교
- ❖ 가설 검정 방법: F distribution
  (집단 내의 분산, 총 평균과 각 집단의 평균의 차이에 의해 생긴 집단 간 분산의 비교)

Requirements

- ❖ An independent random sample
- ❖ A normal distribution
- ❖ Being equal between the population variances and responses for the group levels

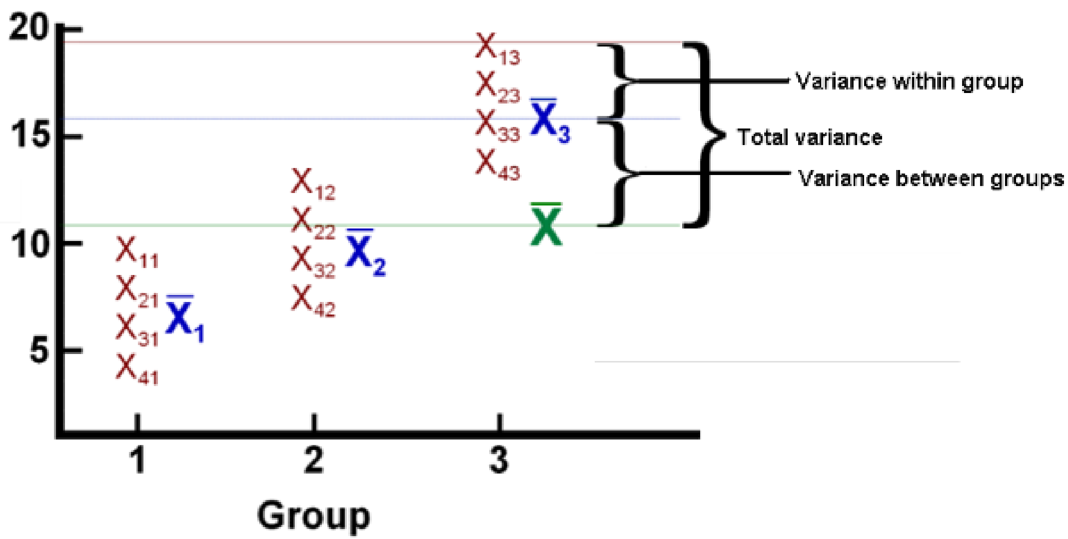
$$x_{ij} = \bar{x}_{.} + \underbrace{(\bar{x}_i - \bar{x}_{.})}_{\substack{\text{deviation of} \\ \text{group mean from} \\ \text{grand mean}}} + \underbrace{(x_{ij} - \bar{x}_i)}_{\substack{\text{deviation of} \\ \text{observation from} \\ \text{group mean}}}$$

One-way Analysis of Variance

$\mathbf{x}_{ij}$	
$\boldsymbol{i}$	Group $\boldsymbol{i}$
$\boldsymbol{j}$	Observation number $\boldsymbol{j}$
$\bar{\boldsymbol{x}}_i$	The mean of group $\boldsymbol{i}$
$\bar{\boldsymbol{x}}.$	Grand mean (average of all observations)

$$x_{ij} = \bar{x}_. + \underbrace{(\bar{x}_i - \bar{x}_.)}_{\text{deviation of group mean from grand mean}} + \underbrace{(x_{ij} - \bar{x}_i)}_{\text{deviation of observation from group mean}}$$

One-way Analysis of Variance



Variation within groups

$$SSD_W = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2$$

Variation between groups

$$SSD_B = \sum_i \sum_j (\bar{x}_i - \bar{x}_.)^2 = \sum_i n_i (\bar{x}_i - \bar{x}_.)^2$$

Total variation

$$SSD_B + SSD_W = SSD_{\text{total}} = \sum_i \sum_j (x_{ij} - \bar{x}_.)^2$$

http://blog.naver.com/ssacstat/220538745824

One-way Analysis of Variance

Limitation

The sums of squares –> only positive

▼

A completely irrelevant grouping –> ”explain”

Partition

- ❖ Normalizing the sums of squares
- ❖ According to the degrees of freedom

SSD_B <- k – 1
SSD_W <- N – k

N: the total number of observations
k: the number of groups

One-way Analysis of Variance

Limitation

The sums of squares –> only positive

▼

A completely irrelevant grouping –> ”explain”

Partition

❖ Normalizing the sums of squares
❖ According to the degrees of freedom

SSD_B <- $k - 1$
SSD_W <- $N - k$

N: the total number of observations
k: the number of groups

$$\mathrm{MS}_W = \mathrm{SSD}_W/(N-k)$$
$$\mathrm{MS}_B = \mathrm{SSD}_B/(k-1)$$

MS_W
❖ The pooled variance
❖ Combining the individual group variances
❖ An estimate of σ^2

MS_B
❖ An estimate of σ^2
❖ Existing a group effect
❖ Being larger

One-way Analysis of Variance

$$F = \mathrm{MS}_B / \mathrm{MS}_W$$

F distribution

- Ideally 1

- The distribution of F under the null hypothesis
- An F distribution with k −1 and N − k degrees of freedom

- Rejecting the hypothesis of identical means
  (if F is larger than the 95% quantile or if the significance level is 5%)

The function lm()

- Used for regression analysis
- aov ()/lme () for more elaborate analyses

- The data values in one vector and a factor variable
- (the division into groups)

- Ventilation (Variance between groups)
- Residual (variance within groups)

Summary

Source	SS	df	MS	F	p
Between Groups (Factor)	$\sum_k n_k(\bar{x}_k - \bar{x}.)^2$	k-1	$\frac{SS_{Between}}{df_{Between}}$	$\frac{MS_{Between}}{MS_{Within}}$	See Minitab Express or F table
Within Groups (Error)	$\sum_k \sum_i (x_{ik} - \bar{x}_k)^2$	n-k	$\frac{SS_{Within}}{df_{Within}}$		
Total	$\sum_k \sum_i (x_{ik} - \bar{x}.)^2$	n-1			

k = Number of groups
n = Total sample size (all groups combined)
n_k = Sample size of group k

$\bar{x}_k$ = Sample mean of group kk
$\bar{x}.$ = Grand mean (i.e., mean for all groups combined)
SS = Sum of squares
MS = Mean square
df = Degrees of freedom
F = F-ratio (the test statistic)

https://onlinecourses.science.psu.edu/stat200/node/67

Analysis of Variance and the Kruskal-Wallis Test

- ❖ One-way analysis of variance
- ❖ Kruskal–Wallis test
- ❖ Two-way analysis of variance
- ❖ The Friedman test
- ❖ The ANOVA table in regression analysis

Kruskal–Wallis test

- ❖ 목적: 두 개 이상 다수의 집단을 비교
  단, 샘플이 아주 작은 경우
- ❖ 가설 검정 방법: 순위를 내어 비교하여 검정 (고유 값들이 순위에 남기 때문에, 그룹 간의 평균, 표준편차는 가설 검정에 의미를 갖지 않음)

Kruskal–Wallis test

- ❖ A nonparametric counterpart of a one-way analysis of variance
- ❖ Based on the between-group sum of squares calculated from the average ranks
- ❖ The function kruskal.test ()

Kruskal–Wallis test

http://dermabae.tistory.com/168

Two-way analysis of variance

❖ 목적: 두 개 이상 다수의 집단을 비교
단, one-way ANOVA와 다르게 측정값에 영향을 미치는 factor가 2개인 경우

❖ Multiple measurements on the same experimental unit (generally the paired t-test)

```
> heart.rate
    hr subj time
1   96    1    0
2  110    2    0
3   89    3    0
4   95    4    0
5  128    5    0
6  100    6    0
7   72    7    0
8   79    8    0
9  100    9    0
10  92    1   30
11 106    2   30
```

Two-way Analysis of Variance

$\boldsymbol{x_{ij}}$	
$\boldsymbol{i}$	Row of the m x n table
$\boldsymbol{j}$	Column of the m x n table
$\boldsymbol{\bar{x}_{i\cdot}}$	Row average
$\boldsymbol{\bar{x}_{j\cdot}}$	Column average

```
> heart.rate
    hr subj time
1   96    1    0
2  110    2    0
3   89    3    0
4   95    4    0
5  128    5    0
6  100    6    0
7   72    7    0
8   79    8    0
9  100    9    0
10  92    1   30
11 106    2   30
```

Variation between rows

$$\mathrm{SSD}_R = n\sum_i (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2$$

Variation between columns

$$\mathrm{SSD}_C = m\sum_j (\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})^2$$

Residual variation

$$\mathrm{SSD}_{\mathrm{res}} = \sum_i \sum_j (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}_{\cdot\cdot})^2$$

Two-way Analysis of Variance

Partition

- ❖ Normalizing the sums of squares
- ❖ According to the degrees of freedom

SSD_R <- $m - 1$
SSD_C <- $n - 1$
SSD_{res} <- $(m - 1)(n - 1)$

- ❖ A set of mean squares

The function gl ()

- ❖ Generating patterned factors for balanced experimental designs

The function anova ()

- ❖ The data values in one vector with the two classifying factors parallel to it

Analysis of Variance and the Kruskal-Wallis Test

- One-way analysis of variance
- Kruskal–Wallis test
- Two-way analysis of variance
- The Friedman test
- The ANOVA table in regression analysis

The Friedman test

- A nonparametric counterpart of two-way analysis of variance
- Based on ranking observations within each row
- Calculated and normalized to give a $\chi 2$ –distributed test statistic
- Less sensitive test than the Wilcoxon signed-rank test
- The function friedman.test ()

Analysis of Variance and the Kruskal-Wallis Test

- ❖ One-way analysis of variance
- ❖ Kruskal–Wallis test
- ❖ Two-way analysis of variance
- ❖ The Friedman test
- ❖ The ANOVA table in regression analysis

The ANOVA table in regression analysis

- ❖ 목적: 회귀 직선이 모집단에서 유의한 지 아닌지 평가. 모집단에서 모든 회귀계수가 0인지 아닌지를 검정
- ❖ 가설 검정: 모든 회귀계수가 0 -> 선형적인 상관성 X -> 회귀 직선 유의 X

The ANOVA table in regression analysis

- ❖ The use of analysis of variance tables in grouped and cross-classified experimental designs
- ❖ Corresponding to a regression analysis
- ❖ The function anova ()

Model variation $$\mathrm{SSD}_{\mathrm{model}} = \sum_i (\hat{y}_i - \bar{y}_.)^2$$

Residual variation $$\mathrm{SSD}_{\mathrm{res}} = \sum_i (y_i - \hat{y}_i)^2$$

8. Tabular data

Tabular data

- Single proportions
- Two independent proportions
- k proportions test for trend
- r × c tables

	Distribution	
Single proportions	Binomial distribution	*N*: size parameter *p*: probability parameter
Large sample size	Normal distribution	*Np*: mean *Np(1-P)*: variance

Yates correction

2 x 2 분할 표에서 카이-제곱 값에 적용 가능
비연속적인 이항분포에서 확률이나 비율을 알기 위한 연속적 분포인 정상 분포나 x^2 분포를 이용할 때는 연속성을 가지도록 비연속성을 교정해야 한다. 이때 사용하는 방법이다.

$$u = \frac{x - Np_0}{\sqrt{Np_0(1-p_0)}}$$

- x = the observed number of "success"
- $p = p_0$

- Mean = 0
- SD =1
- An approximate x^2 distribution with 1 degree of freedom

The function prop.test ()

- Number of positive outcomes
- Total number
- Probability parameter

```
> prop.test(39,215,.15)

         1-sample proportions test with continuity correction

data:  39 out of 215, null probability 0.15
X-squared = 1.425, df = 1, p-value = 0.2326
alternative hypothesis: true p is not equal to 0.15
95 percent confidence interval:
 0.1335937 0.2408799
sample estimates:
        p
0.1813953
```

The function binom.test ()

- Doing more than testing single proportions

- Obtaining the p-value
  Calculating the point probabilities for all the possible values of x
  Summing those that are less than or equal to the point probability of the observed x

```
> binom.test(39,215,.15)

         Exact binomial test

data:  39 and 215
number of successes = 39, number of trials = 215, p-value = 0.2135
alternative hypothesis: true probability  ...  not equal to 0.15
95 percent confidence interval:
 0.1322842 0.2395223
sample estimates:
probability of success
             0.1813953
```

Tabular data

- ❖ Single proportions
- ❖ Two independent proportions
- ❖ k proportions test for trend
- ❖ r × c tables

Two independent proportions

- ❖ Comparing two or more proportions
- ❖ Two vectors

  1st number of positive outcomes

  2nd total number for each group

```
> lewitt.machin.success <- c(9,4)
> lewitt.machin.total <- c(12,13)
> prop.test(lewitt.machin.success,lewitt.machin.total)

        2-sample test for equality of proportions with continuity
        correction

data:  lewitt.machin.success out of lewitt.machin.total
X-squared = 3.2793, df = 1, p-value = 0.07016
alternative hypothesis: two.sided
95 percent confidence interval:
 0.01151032 0.87310506
sample estimates:
   prop 1    prop 2
0.7500000 0.3076923
```

Two independent proportions

- Without Yates correction (correct=F)
- Fisher's exact test for *p*-value correction
- The function fisher.test ()

```
> matrix(c(9,4,3,9),2)
     [,1] [,2]
[1,]    9    3
[2,]    4    9

> lewitt.machin <- matrix(c(9,4,3,9),2)
> fisher.test(lewitt.machin)

        Fisher's Exact Test for Count Data

data:  lewitt.machin
p-value = 0.04718
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.9006803 57.2549701
sample estimates:
odds ratio
  6.180528
```

- Standard x^2 test in chisq.test ()

```
> chisq.test(lewitt.machin)

      Pearson's Chi-squared test with Yates' continuity
      correction

data:  lewitt.machin
X-squared = 3.2793, df = 1, p-value = 0.07016
```

Tabular data

- Single proportions
- Two independent proportions
- k proportions test for trend
- r × c tables

- Case of decreasing or increasing trend in the proportions with group number

```
> caesar.shoe
    <4  4 4.5  5 5.5  6+
Yes  5  7   6  7   8  10
No  17 28  36 41  46 140
```

```
> caesar.shoe.yes <- caesar.shoe["Yes",]
> caesar.shoe.total <- margin.table(caesar.shoe,2)
> caesar.shoe.yes
 <4   4 4.5   5 5.5  6+
  5   7   6   7   8  10
> caesar.shoe.total
 <4   4 4.5   5 5.5  6+
 22  35  42  48  54 150
```

k proportions test for trend

```
> prop.test(caesar.shoe.yes,caesar.shoe.total)
        6-sample test for equality of proportions without
        continuity correction

data:  caesar.shoe.yes out of caesar.shoe.total
X-squared = 9.2874, df = 5, p-value = 0.09814
alternative hypothesis: two.sided
sample estimates:
    prop 1     prop 2     prop 3     prop 4     prop 5     prop 6
0.22727273 0.20000000 0.14285714 0.14583333 0.14814815 0.06666667

Warning message:
In prop.test(caesar.shoe.yes, caesar.shoe.total) :
  Chi-squared approximation may be incorrect
```

Nonsignificant ??
Fine in view of the small number of caesarean sections

```
> prop.trend.test(caesar.shoe.yes,caesar.shoe.total)

      Chi-squared Test for Trend in Proportions

data:  caesar.shoe.yes out of caesar.shoe.total ,
 using scores: 1 2 3 4 5 6
X-squared = 8.0237, df = 1, p-value = 0.004617
```

The warning about the x^2 approximation
The function prop.trend.test ()
Parameter: x, n, score
(score: given to the groups)

A weighted linear regression of the proportion on the group scores

The rough type of alternative to which the test should be sensitive

Tabular data

- ❖ Single proportions
- ❖ Two independent proportions
- ❖ k proportions test for trend
- ❖ r × c tables

Tables with more that two classes on both sides

$$
\begin{array}{cccc|c}
n_{11} & n_{12} & \cdots & n_{1c} & n_{1\cdot} \\
n_{21} & n_{22} & \cdots & n_{2c} & n_{2\cdot} \\
\vdots & \vdots & & \vdots & \vdots \\
n_{r1} & n_{r2} & \cdots & n_{rc} & n_{r\cdot} \\
\hline
n_{\cdot 1} & n_{\cdot 2} & \cdots & n_{\cdot c} & n_{\cdot\cdot}
\end{array}
$$

Problems

- ❖ Several different sampling plans
- ❖ No relation between rows and columns

▼

- ❖ Fixing each rows or columns
- ❖ The same probabilities

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

r × c tables

Using the function chisq.test () or fisher.test ()
(r-1)(c-1) degrees of freedom

```
> caff.marital <- matrix(c(652,1537,598,242,36,46,38,21,218
+ ,327,106,67),
+ nrow=3,byrow=T)
> colnames(caff.marital) <- c("0","1-150","151-300",">300")
> rownames(caff.marital) <- c("Married","Prev.married","Single")
> caff.marital
               0 1-150 151-300 >300
Married      652  1537     598  242
Prev.married  36    46      38   21
Single       218   327     106   67
> chisq.test(caff.marital)

      Pearson's Chi-squared test

data:  caff.marital
X-squared = 51.6556, df = 6, p-value = 2.187e-09
```

Thank you for your attention