EasyDescribe：一个方便的R语言基本统计集成包

聂秀泉

我们的日常统计分析中，对变量的基本统计描述和基础统计分析是经常需要做的，例如计算均值（标准差）、中位数（四分位间距），进行t检验、方差分析、多重检验矫正等等。然而，作为专门为统计而生的R语言，进行描述性统计的方法却“选择多得简直让人尴尬！”（《R语言实战•第二版》134页作者如是说），这对于许多初学者、统计学小白以及选择困难症患者来说，简直就是噩梦：每当要进行一项简单的统计分析时，就需要在多得让然尴尬的方法中进行对比与挑选，想想就让人头大。为了解决这一问题，我开发了EasyDescribe这个包，用一个函数解决几乎所有的常见基本统计描述，让R程序员不再选择困难。

接下来介绍一下EasyDescribe包的使用逻辑：

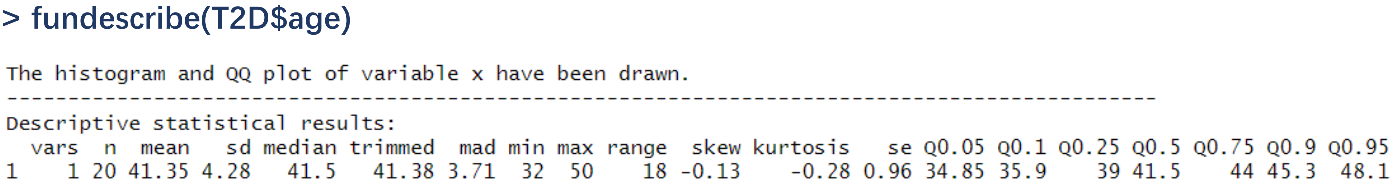
为了杜绝选择，EasyDescribe仅有fundescribe()这一个函数，不需要你再选择！那这一个函数是如何包办这些基本统计分析的呢？

fundescribe(x, y, data = NULL, na.rm = TRUE, norm.t = NULL)

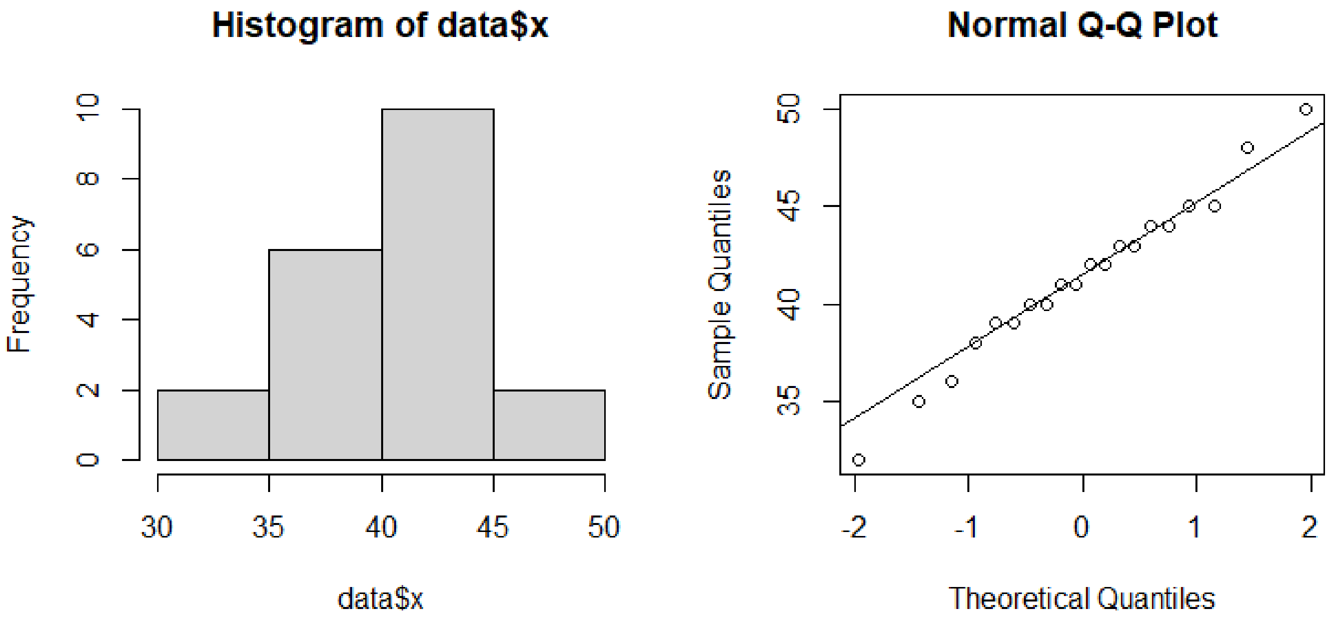
fundescribe()存在两个基本参数：x和y，它们就是你想分析的两个基本变量。

数据类型可以基本分成四大类：正态连续型变量、非正态连续变量、有序分类变量和无序分类变量，我们在做基本统计分析进行方法选择时，实际上大部分情况下就是在根据数据类型和实验设计进行方法选择。而fundescribe()函数就是自动根据你输入x和y的数据类型自动进行统计方法的选择。

比如，你单纯输入了一个连续型变量 fundescribe(T2D$age)，函数就会自动输出均值、标准差、中位数、四分位数等等，而且还会输出一个直方图和 QQ 图方便你了解数据的正态性与分布情况：



```
> fundescribe(T2D$age)
The histogram and QQ plot of variable x have been drawn.
-------------------------------------------------------------------------------------------
Descriptive statistical results:
  vars  n  mean   sd median trimmed  mad min max range  skew kurtosis   se Q0.05 Q0.1 Q0.25 Q0.5 Q0.75 Q0.9 Q0.95
1    1 20 41.35 4.28   41.5   41.38 3.71  32  50    18 -0.13    -0.28 0.96 34.85 35.9    39 41.5    44 45.3  48.1
```



如果你单纯输入一个分类变量 fundescribe(T2D$gender)，函数就会自动输出各个分类数量与占比：

```
> fundescribe(T2D$gender)
```

```
   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  20 


             |         F |         M | 
             |-----------|-----------|
             |         9 |        11 | 
             |   0.45000 |   0.55000 | 
             |-----------|-----------|
```

所以，我们可以看到，fundescribe()函数的使用逻辑就是极简，不需要你操心输入的数据类型，它会根据你输入的变量类型进行自动方法选择。

上面是仅输入 x 的情况，如果同时输入 x 和 y，fundescribe()同样可以自动识别 x 和 y 的数据类型进行自动选择所对应的基本统计方法：

例 1、x 连续型变量，y 无序分类变量：

```
> fundescribe(T2D$age, T2D$gender)
```

基本统计描述

```
The histogram and QQ plot of variable x have been drawn.
-----------------------------------------------------------------------------------------
Descriptive statistical results:
  vars  n  mean   sd median trimmed  mad min max range  skew kurtosis   se Q0.05 Q0.1 Q0.25 Q0.5 Q0.75 Q0.9 Q0.95
1    1 20 41.35 4.28   41.5   41.38 3.71  32  50    18 -0.13    -0.28 0.96 34.85 35.9    39 41.5    44 45.3  48.1
-----------------------------------------------------------------------------------------
```

分层基本统计描述

```
Descriptive statistical results stratified by y:

 Descriptive statistics by group 
group: F
  vars n  mean   sd median trimmed  mad min max range  skew kurtosis   se Q0.05 Q0.1 Q0.25 Q0.5 Q0.75 Q0.9 Q0.95
1    1 9 39.89 2.15     40   39.89 1.48  35  42     7 -1.09     0.22 0.72  36.6 38.2    39   40    41   42    42
---------------------------------------------------------------------------------------------------------------
group: M
  vars  n  mean   sd median trimmed  mad min max range  skew kurtosis   se Q0.05 Q0.1 Q0.25 Q0.5 Q0.75 Q0.9 Q0.95
1    1 11 42.55 5.26     44   42.89 1.48  32  50    18 -0.58    -0.78 1.59    34   36  40.5   44    45   48    49
------------------------------------------------------------
```

两独立样本 t 检验

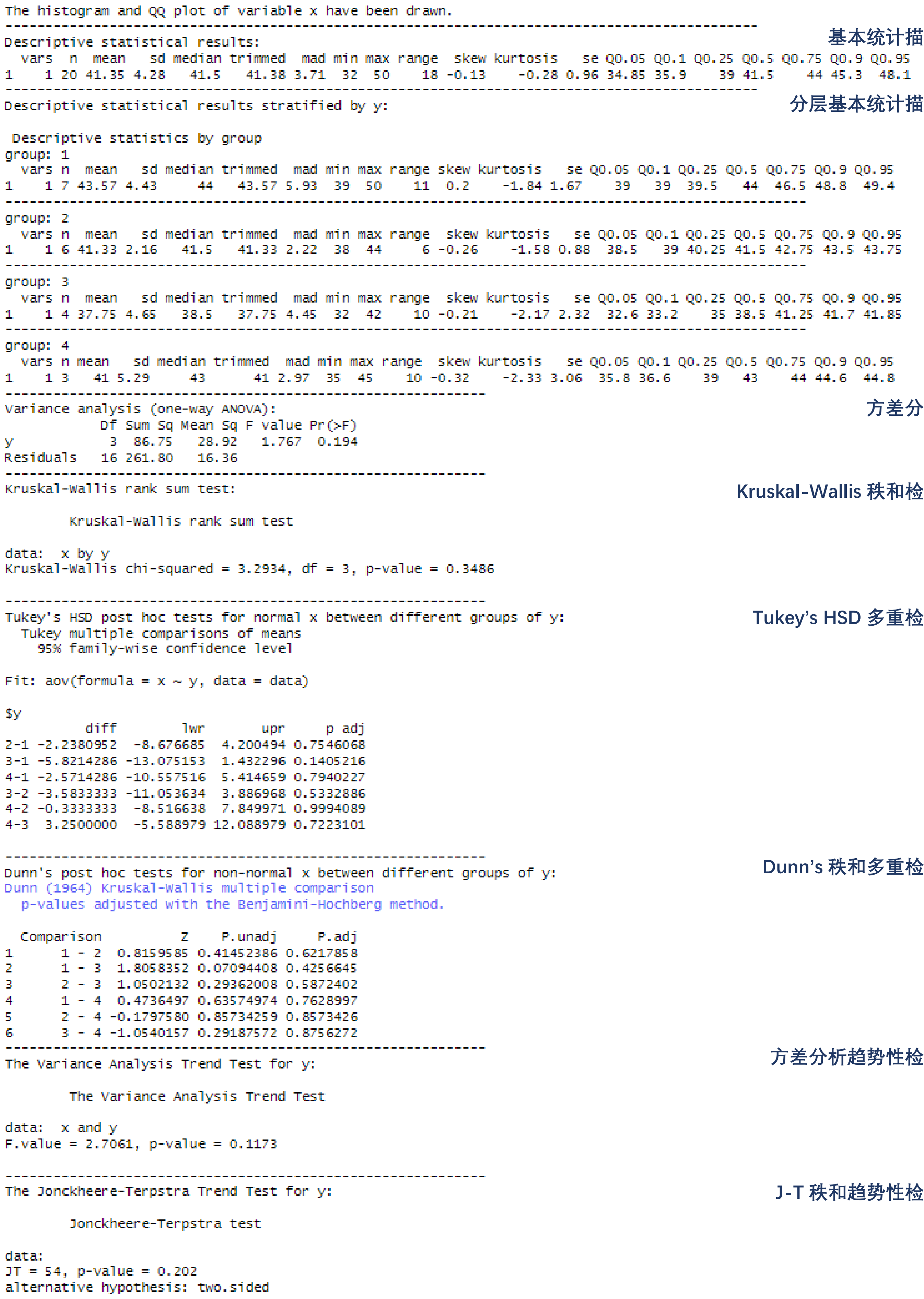
```
Two sample t-test:

        Welch Two Sample t-test

data:  x by y
t = -1.5266, df = 13.774, p-value = 0.1495
alternative hypothesis: true difference in means between group F and group M is not equal to 0
95 percent confidence interval:
 -6.394528  1.081397
sample estimates:
mean in group F mean in group M 
       39.88889        42.54545 

------------------------------------------------------------
```

Wilcoxon 秩和检验

```
Wilcoxon rank sum test:
Mann-Whitney U test = Wilcoxon rank sum test

        Wilcoxon rank sum test with continuity correction

data:  x by y
W = 25, p-value = 0.06752
alternative hypothesis: true location shift is not equal to 0
```

例 2、x 连续型变量，y 有序分类变量：

> fundescribe(age, education, data = T2D)



```
The histogram and QQ plot of variable x have been drawn.
---------------------------------------------------------------------------------------------
```

基本统计描述

```
Descriptive statistical results:
  vars  n  mean   sd median trimmed  mad min max range  skew kurtosis   se Q0.05 Q0.1 Q0.25 Q0.5 Q0.75 Q0.9 Q0.95
1    1 20 41.35 4.28   41.5   41.38 3.71  32  50    18 -0.13    -0.28 0.96 34.85 35.9    39 41.5    44 45.3  48.1
---------------------------------------------------------------------------------------------
```

分层基本统计描述

```
Descriptive statistical results stratified by y:

 Descriptive statistics by group 
group: 1
  vars n  mean   sd median trimmed  mad min max range skew kurtosis   se Q0.05 Q0.1 Q0.25 Q0.5 Q0.75 Q0.9 Q0.95
1    1 7 43.57 4.43     44   43.57 5.93  39  50    11  0.2    -1.84 1.67    39   39  39.5   44  46.5 48.8  49.4
--------------------------------------------------------------------------------------------------- 
group: 2
  vars n  mean   sd median trimmed  mad min max range  skew kurtosis   se Q0.05 Q0.1 Q0.25 Q0.5 Q0.75 Q0.9 Q0.95
1    1 6 41.33 2.16   41.5   41.33 2.22  38  44     6 -0.26    -1.58 0.88  38.5   39 40.25 41.5 42.75 43.5 43.75
--------------------------------------------------------------------------------------------------- 
group: 3
  vars n  mean   sd median trimmed  mad min max range  skew kurtosis   se Q0.05 Q0.1 Q0.25 Q0.5 Q0.75 Q0.9 Q0.95
1    1 4 37.75 4.65   38.5   37.75 4.45  32  42    10 -0.21    -2.17 2.32  32.6 33.2    35 38.5 41.25 41.7 41.85
--------------------------------------------------------------------------------------------------- 
group: 4
  vars n mean   sd median trimmed  mad min max range  skew kurtosis   se Q0.05 Q0.1 Q0.25 Q0.5 Q0.75 Q0.9 Q0.95
1    1 3   41 5.29     43      41 2.97  35  45    10 -0.32    -2.33 3.06  35.8 36.6    39   43    44 44.6  44.8
-----------------------------------------------------------
```

方差分析

```
Variance analysis (one-way ANOVA):
            Df Sum Sq Mean Sq F value Pr(>F)
y            3  86.75   28.92   1.767  0.194
Residuals   16 261.80   16.36               
-----------------------------------------------------------
```

Kruskal-Wallis 秩和检验

```
Kruskal-Wallis rank sum test:

	Kruskal-Wallis rank sum test

data:  x by y
Kruskal-Wallis chi-squared = 3.2934, df = 3, p-value = 0.3486

-----------------------------------------------------------
```

Tukey’s HSD 多重检验

```
Tukey's HSD post hoc tests for normal x between different groups of y:
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = x ~ y, data = data)

$y
          diff        lwr       upr     p adj
2-1 -2.2380952  -8.676685  4.200494 0.7546068
3-1 -5.8214286 -13.075153  1.432296 0.1405216
4-1 -2.5714286 -10.557516  5.414659 0.7940227
3-2 -3.5833333 -11.053634  3.886968 0.5332886
4-2 -0.3333333  -8.516638  7.849971 0.9994089
4-3  3.2500000  -5.588979 12.088979 0.7223101

-----------------------------------------------------------
```

Dunn’s 秩和多重检验

```
Dunn's post hoc tests for non-normal x between different groups of y:
Dunn (1964) Kruskal-Wallis multiple comparison
  p-values adjusted with the Benjamini-Hochberg method.

  Comparison          Z    P.unadj     P.adj
1      1 - 2  0.8159585 0.41452386 0.6217858
2      1 - 3  1.8058352 0.07094408 0.4256645
3      2 - 3  1.0502132 0.29362008 0.5872402
4      1 - 4  0.4736497 0.63574974 0.7628997
5      2 - 4 -0.1797580 0.85734259 0.8573426
6      3 - 4 -1.0540157 0.29187572 0.8756272
-----------------------------------------------------------
```

方差分析趋势性检验

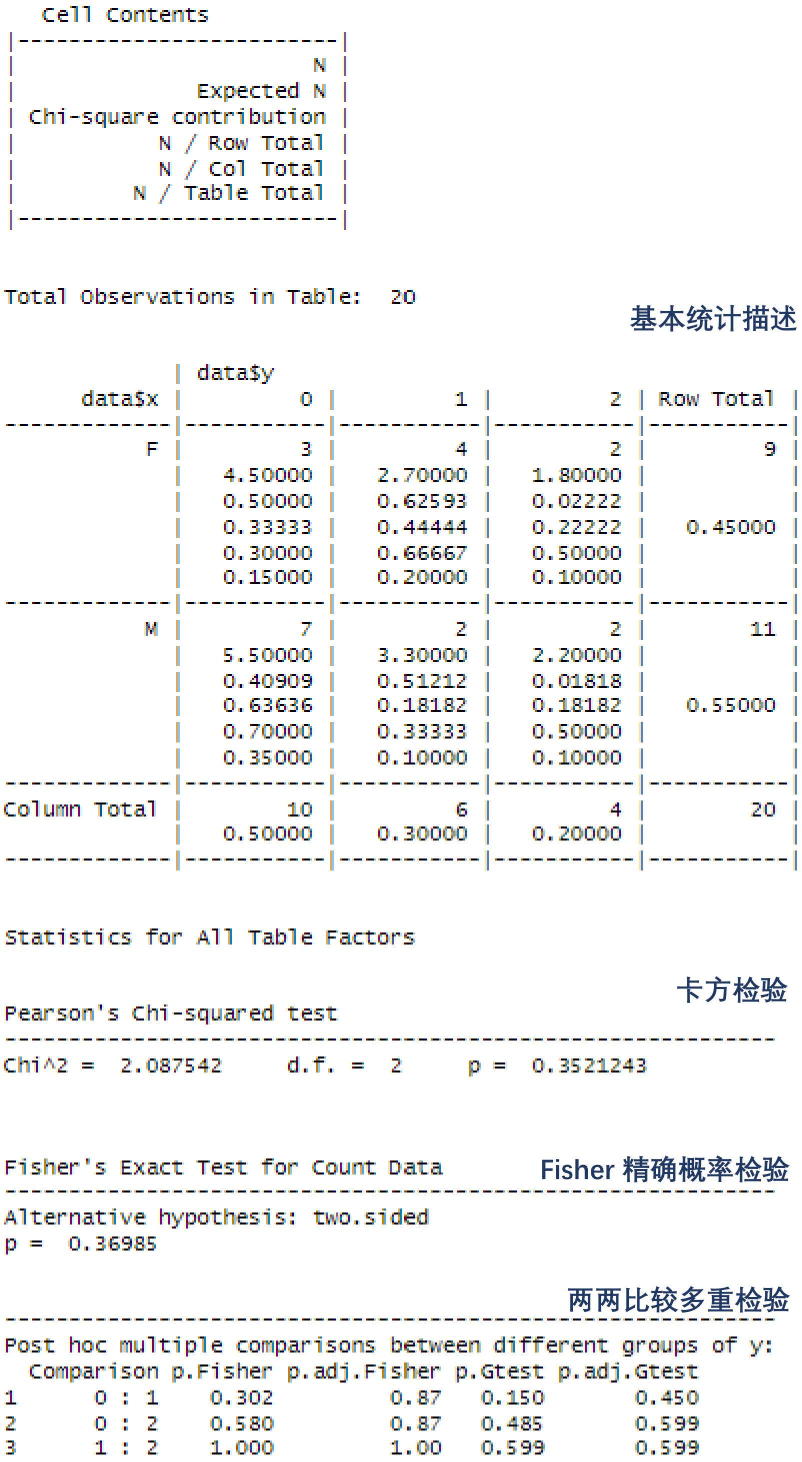
```
The Variance Analysis Trend Test for y:

	The Variance Analysis Trend Test

data:  x and y
F.value = 2.7061, p-value = 0.1173

-----------------------------------------------------------
```

J-T 秩和趋势性检验

```
The Jonckheere-Terpstra Trend Test for y:

	Jonckheere-Terpstra test

data:  
JT = 54, p-value = 0.202
alternative hypothesis: two.sided
```

例 3：x 无序分类变量，y 无序分类变量：

> fundescribe(gender, smoke, data = T2D)



```
   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  20
```

基本统计描述

```
             | data$y
      data$x |         0 |         1 |         2 | Row Total |
-------------|-----------|-----------|-----------|-----------|
           F |         3 |         4 |         2 |         9 |
             |   4.50000 |   2.70000 |   1.80000 |           |
             |   0.50000 |   0.62593 |   0.02222 |           |
             |   0.33333 |   0.44444 |   0.22222 |   0.45000 |
             |   0.30000 |   0.66667 |   0.50000 |           |
             |   0.15000 |   0.20000 |   0.10000 |           |
-------------|-----------|-----------|-----------|-----------|
           M |         7 |         2 |         2 |        11 |
             |   5.50000 |   3.30000 |   2.20000 |           |
             |   0.40909 |   0.51212 |   0.01818 |           |
             |   0.63636 |   0.18182 |   0.18182 |   0.55000 |
             |   0.70000 |   0.33333 |   0.50000 |           |
             |   0.35000 |   0.10000 |   0.10000 |           |
-------------|-----------|-----------|-----------|-----------|
Column Total |        10 |         6 |         4 |        20 |
             |   0.50000 |   0.30000 |   0.20000 |           |
-------------|-----------|-----------|-----------|-----------|


Statistics for All Table Factors
```

卡方检验

```
Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  2.087542     d.f. =  2     p =  0.3521243
```

```
Fisher's Exact Test for Count Data
------------------------------------------------------------
Alternative hypothesis: two.sided
p =  0.36985
```

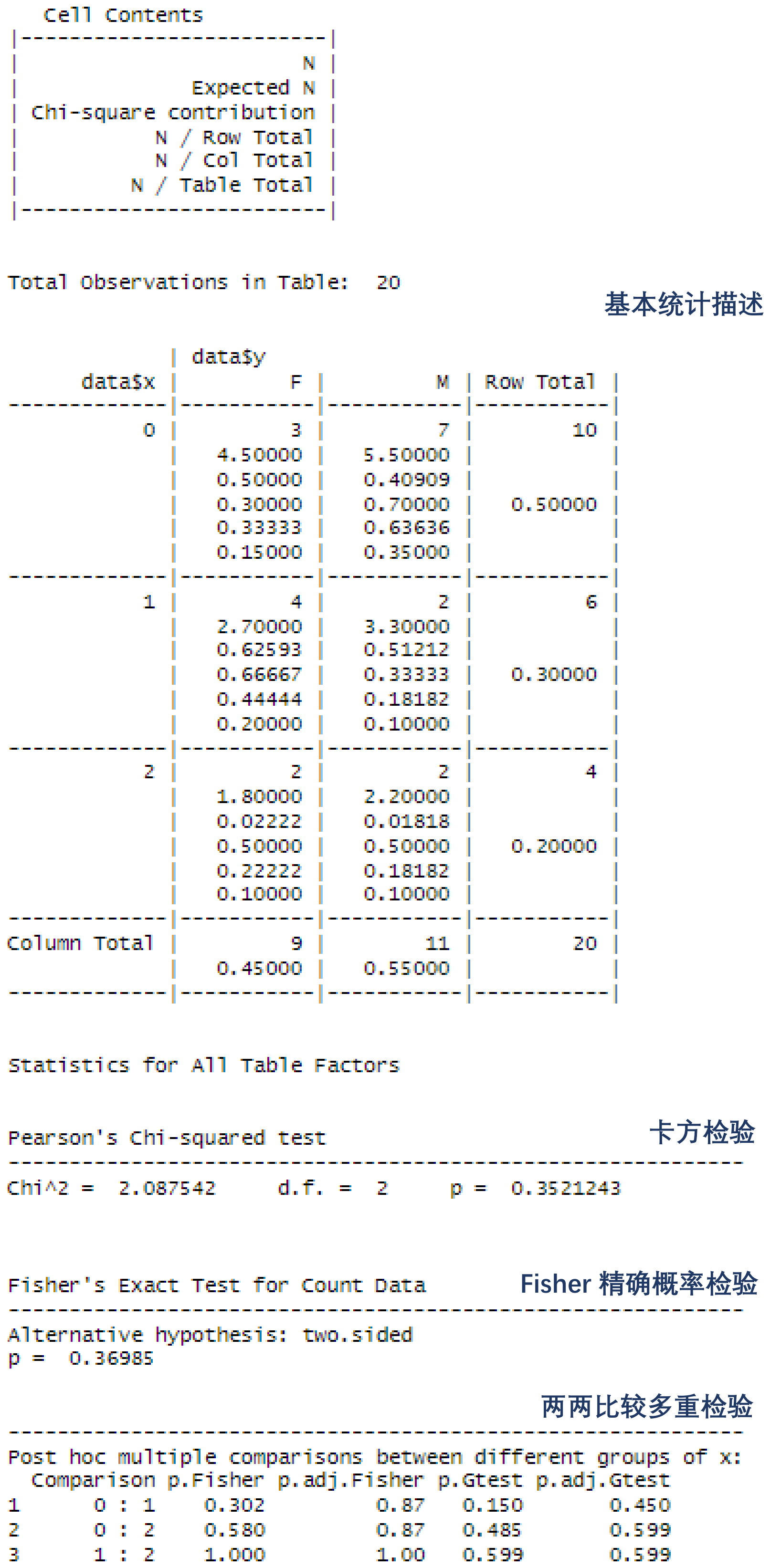
Fisher 精确概率检验

两两比较多重检验

```
------------------------------------------------------------
Post hoc multiple comparisons between different groups of y:
  Comparison p.Fisher p.adj.Fisher p.Gtest p.adj.Gtest
1      0 : 1    0.302         0.87   0.150       0.450
2      0 : 2    0.580         0.87   0.485       0.599
3      1 : 2    1.000         1.00   0.599       0.599
```

例 4：x 无序分类变量，y 无序分类变量：

> fundescribe(T2D$smoke, T2D$gender)



```
   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  20
```

基本统计描述

```
             | data$y
      data$x |         F |         M | Row Total |
-------------|-----------|-----------|-----------|
           0 |         3 |         7 |        10 |
             |   4.50000 |   5.50000 |           |
             |   0.50000 |   0.40909 |           |
             |   0.30000 |   0.70000 |   0.50000 |
             |   0.33333 |   0.63636 |           |
             |   0.15000 |   0.35000 |           |
-------------|-----------|-----------|-----------|
           1 |         4 |         2 |         6 |
             |   2.70000 |   3.30000 |           |
             |   0.62593 |   0.51212 |           |
             |   0.66667 |   0.33333 |   0.30000 |
             |   0.44444 |   0.18182 |           |
             |   0.20000 |   0.10000 |           |
-------------|-----------|-----------|-----------|
           2 |         2 |         2 |         4 |
             |   1.80000 |   2.20000 |           |
             |   0.02222 |   0.01818 |           |
             |   0.50000 |   0.50000 |   0.20000 |
             |   0.22222 |   0.18182 |           |
             |   0.10000 |   0.10000 |           |
-------------|-----------|-----------|-----------|
Column Total |         9 |        11 |        20 |
             |   0.45000 |   0.55000 |           |
-------------|-----------|-----------|-----------|


Statistics for All Table Factors
```

卡方检验

```
Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  2.087542     d.f. =  2     p =  0.3521243
```

Fisher 精确概率检验

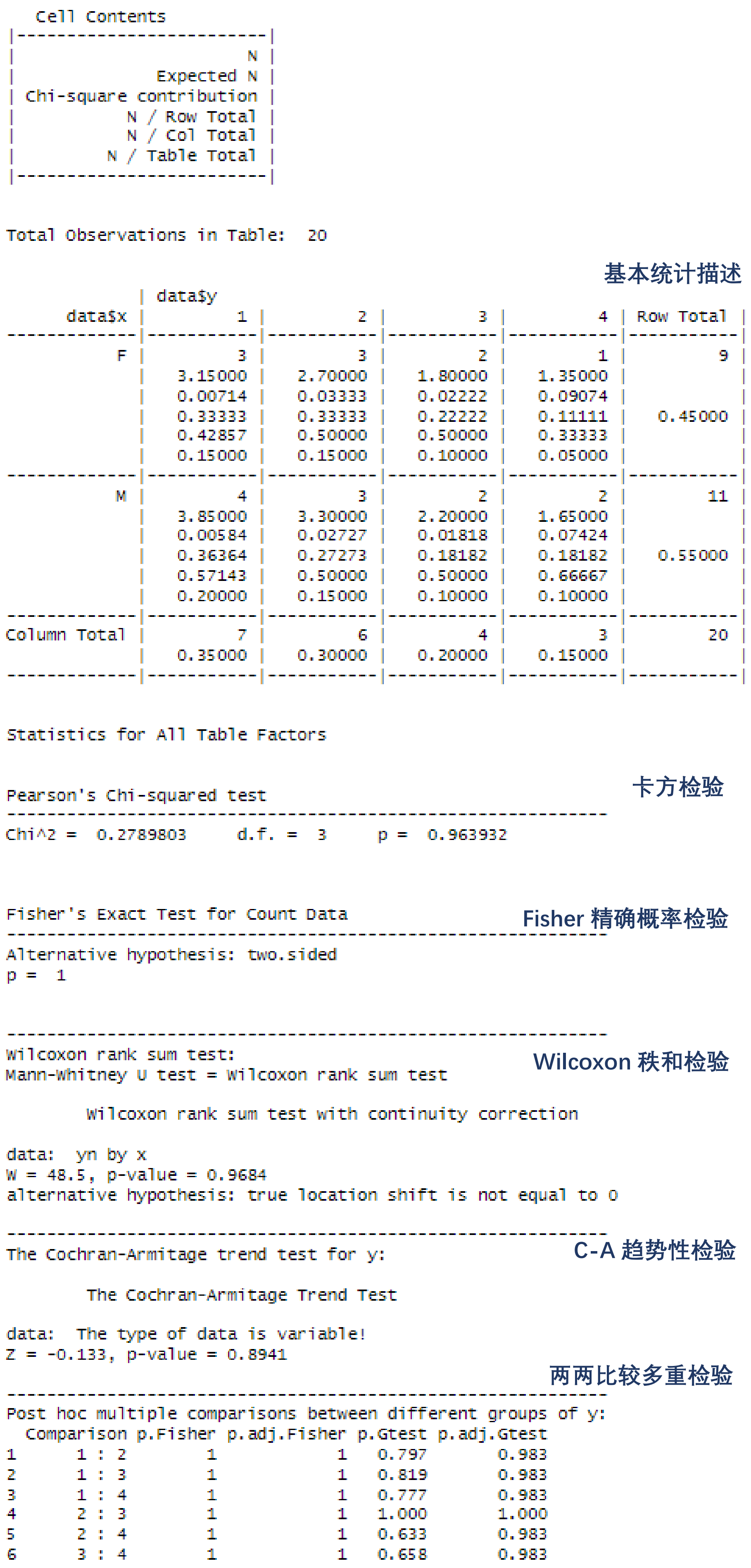
```
Fisher's Exact Test for Count Data
------------------------------------------------------------
Alternative hypothesis: two.sided
p =  0.36985
```

两两比较多重检验

```
------------------------------------------------------------
Post hoc multiple comparisons between different groups of x:
  Comparison p.Fisher p.adj.Fisher p.Gtest p.adj.Gtest
1      0 : 1    0.302         0.87   0.150       0.450
2      0 : 2    0.580         0.87   0.485       0.599
3      1 : 2    1.000         1.00   0.599       0.599
```

例 5：x 无序分类变量，y 有序分类变量：

> fundescribe(T2D$gender, T2D$education)



```
   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  20
```

基本统计描述

```
             | data$y
      data$x |         1 |         2 |         3 |         4 | Row Total |
-------------|-----------|-----------|-----------|-----------|-----------|
           F |         3 |         3 |         2 |         1 |         9 |
             |   3.15000 |   2.70000 |   1.80000 |   1.35000 |           |
             |   0.00714 |   0.03333 |   0.02222 |   0.09074 |           |
             |   0.33333 |   0.33333 |   0.22222 |   0.11111 |   0.45000 |
             |   0.42857 |   0.50000 |   0.50000 |   0.33333 |           |
             |   0.15000 |   0.15000 |   0.10000 |   0.05000 |           |
-------------|-----------|-----------|-----------|-----------|-----------|
           M |         4 |         3 |         2 |         2 |        11 |
             |   3.85000 |   3.30000 |   2.20000 |   1.65000 |           |
             |   0.00584 |   0.02727 |   0.01818 |   0.07424 |           |
             |   0.36364 |   0.27273 |   0.18182 |   0.18182 |   0.55000 |
             |   0.57143 |   0.50000 |   0.50000 |   0.66667 |           |
             |   0.20000 |   0.15000 |   0.10000 |   0.10000 |           |
-------------|-----------|-----------|-----------|-----------|-----------|
Column Total |         7 |         6 |         4 |         3 |        20 |
             |   0.35000 |   0.30000 |   0.20000 |   0.15000 |           |
-------------|-----------|-----------|-----------|-----------|-----------|


Statistics for All Table Factors
```

卡方检验

```
Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  0.2789803     d.f. =  3     p =  0.963932
```

Fisher 精确概率检验

```
Fisher's Exact Test for Count Data
------------------------------------------------------------
Alternative hypothesis: two.sided
p =  1


------------------------------------------------------------
Wilcoxon rank sum test:
Mann-Whitney U test = Wilcoxon rank sum test
```

Wilcoxon 秩和检验

```
        Wilcoxon rank sum test with continuity correction

data:  yn by x
W = 48.5, p-value = 0.9684
alternative hypothesis: true location shift is not equal to 0

------------------------------------------------------------
The Cochran-Armitage trend test for y:
```

C-A 趋势性检验

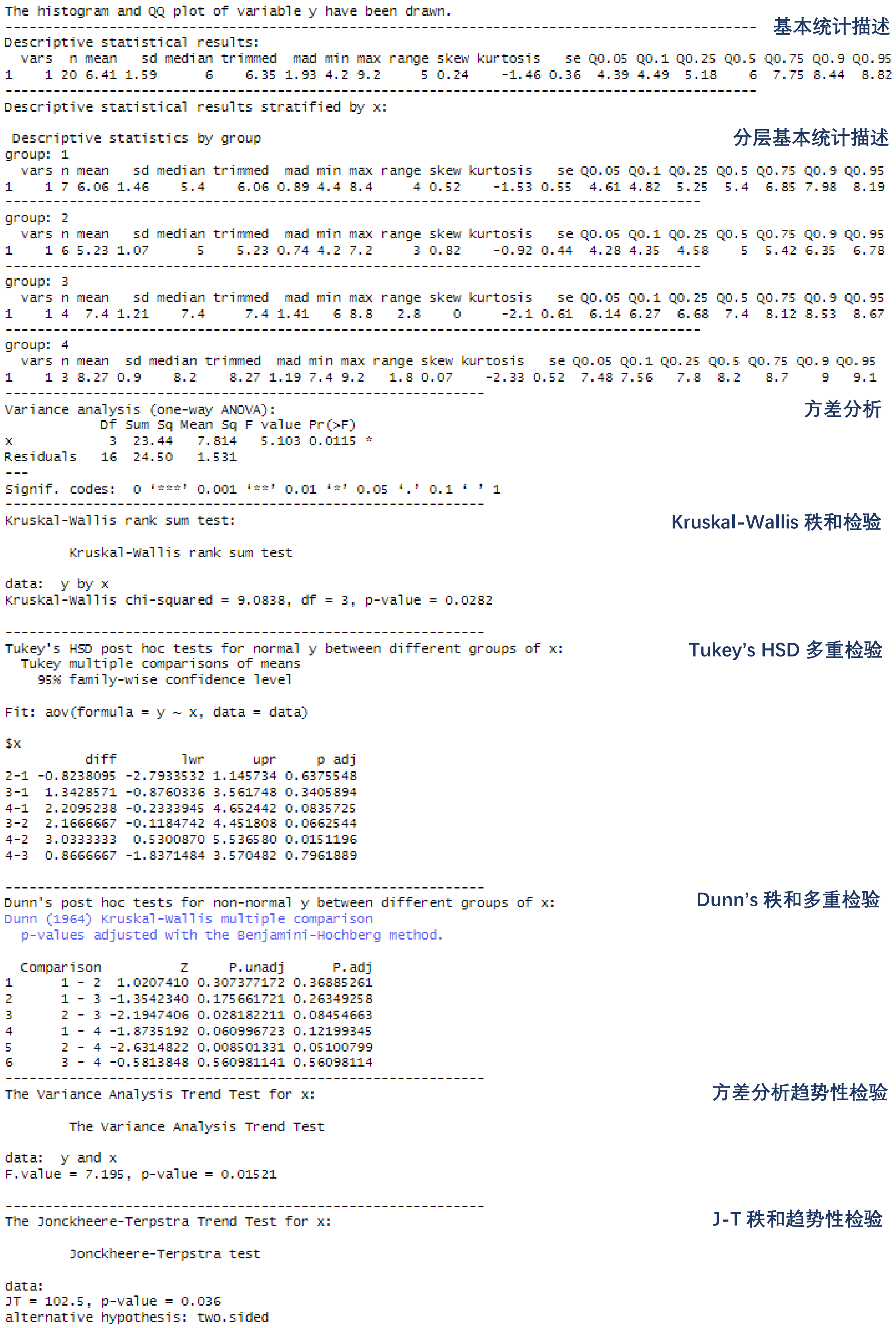
```
        The Cochran-Armitage Trend Test

data:  The type of data is variable!
Z = -0.133, p-value = 0.8941
```

两两比较多重检验

```
------------------------------------------------------------
Post hoc multiple comparisons between different groups of y:
  Comparison p.Fisher p.adj.Fisher p.Gtest p.adj.Gtest
1      1 : 2        1            1   0.797       0.983
2      1 : 3        1            1   0.819       0.983
3      1 : 4        1            1   0.777       0.983
4      2 : 3        1            1   1.000       1.000
5      2 : 4        1            1   0.633       0.983
6      3 : 4        1            1   0.658       0.983
```

例 6、x 有序分类变量，y 连续型变量：

> fundescribe(T2D$education, T2D$glucose)



基本统计描述

```
The histogram and QQ plot of variable y have been drawn.
------------------------------------------------------------------------------------------------
Descriptive statistical results:
  vars  n mean   sd median trimmed  mad min max range skew kurtosis   se Q0.05 Q0.1 Q0.25 Q0.5 Q0.75 Q0.9 Q0.95
1    1 20 6.41 1.59      6    6.35 1.93 4.2 9.2     5 0.24    -1.46 0.36  4.39 4.49  5.18    6  7.75 8.44  8.82
------------------------------------------------------------------------------------------------
Descriptive statistical results stratified by x:
```

分层基本统计描述

```
 Descriptive statistics by group 
group: 1
  vars n mean   sd median trimmed  mad min max range skew kurtosis   se Q0.05 Q0.1 Q0.25 Q0.5 Q0.75 Q0.9 Q0.95
1    1 7 6.06 1.46    5.4    6.06 0.89 4.4 8.4     4 0.52    -1.53 0.55  4.61 4.82  5.25  5.4  6.85 7.98  8.19
-----------------------------------------------------------------------------------------
group: 2
  vars n mean   sd median trimmed  mad min max range skew kurtosis   se Q0.05 Q0.1 Q0.25 Q0.5 Q0.75 Q0.9 Q0.95
1    1 6 5.23 1.07      5    5.23 0.74 4.2 7.2     3 0.82    -0.92 0.44  4.28 4.35  4.58    5  5.42 6.35  6.78
-----------------------------------------------------------------------------------------
group: 3
  vars n mean   sd median trimmed  mad min max range skew kurtosis   se Q0.05 Q0.1 Q0.25 Q0.5 Q0.75 Q0.9 Q0.95
1    1 4  7.4 1.21    7.4     7.4 1.41   6 8.8   2.8    0     -2.1 0.61  6.14 6.27  6.68  7.4  8.12 8.53  8.67
-----------------------------------------------------------------------------------------
group: 4
  vars n mean  sd median trimmed  mad min max range skew kurtosis   se Q0.05 Q0.1 Q0.25 Q0.5 Q0.75 Q0.9 Q0.95
1    1 3 8.27 0.9    8.2    8.27 1.19 7.4 9.2   1.8 0.07    -2.33 0.52  7.48 7.56   7.8  8.2   8.7    9   9.1
-------------------------------------------------------------
```

方差分析

```
Variance analysis (one-way ANOVA):
            Df Sum Sq Mean Sq F value Pr(>F)  
x            3  23.44   7.814   5.103 0.0115 *
Residuals   16  24.50   1.531                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
-------------------------------------------------------------
```

Kruskal-Wallis 秩和检验

```
Kruskal-Wallis rank sum test:

	Kruskal-Wallis rank sum test

data:  y by x
Kruskal-Wallis chi-squared = 9.0838, df = 3, p-value = 0.0282

-------------------------------------------------------------
```

Tukey’s HSD 多重检验

```
Tukey's HSD post hoc tests for normal y between different groups of x:
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = y ~ x, data = data)

$x
          diff        lwr      upr     p adj
2-1 -0.8238095 -2.7933532 1.145734 0.6375548
3-1  1.3428571 -0.8760336 3.561748 0.3405894
4-1  2.2095238 -0.2333945 4.652442 0.0835725
3-2  2.1666667 -0.1184742 4.451808 0.0662544
4-2  3.0333333  0.5300870 5.536580 0.0151196
4-3  0.8666667 -1.8371484 3.570482 0.7961889

-------------------------------------------------------------
```

Dunn’s 秩和多重检验

```
Dunn's post hoc tests for non-normal y between different groups of x:
Dunn (1964) Kruskal-Wallis multiple comparison
  p-values adjusted with the Benjamini-Hochberg method.

  Comparison          Z     P.unadj      P.adj
1      1 - 2  1.0207410 0.307377172 0.36885261
2      1 - 3 -1.3542340 0.175661721 0.26349258
3      2 - 3 -2.1947406 0.028182211 0.08454663
4      1 - 4 -1.8735192 0.060996723 0.12199345
5      2 - 4 -2.6314822 0.008501331 0.05100799
6      3 - 4 -0.5813848 0.560981141 0.56098114
-------------------------------------------------------------
```

方差分析趋势性检验

```
The Variance Analysis Trend Test for x:

	The Variance Analysis Trend Test

data:  y and x
F.value = 7.195, p-value = 0.01521

-------------------------------------------------------------
```

J-T 秩和趋势性检验

```
The Jonckheere-Terpstra Trend Test for x:

	Jonckheere-Terpstra test

data:  
JT = 102.5, p-value = 0.036
alternative hypothesis: two.sided
```

例 7、x 连续型变量，y 连续型变量：

> fundescribe(T2D$age, T2D$glucose)

基本统计描述

```
The histogram and QQ plot of variable x and y have been drawn.
------------------------------------------------------------
Descriptive statistical results for x:
  vars  n  mean   sd median trimmed  mad min max range  skew kurtosis   se Q0.05 Q0.1 Q0.25 Q0.5 Q0.75 Q0.9
1    1 20 41.35 4.28   41.5   41.38 3.71  32  50    18 -0.13    -0.28 0.96 34.85 35.9    39 41.5    44 45.3
  Q0.95
1  48.1
------------------------------------------------------------
Descriptive statistical results for y:
  vars  n mean   sd median trimmed  mad min max range skew kurtosis   se Q0.05 Q0.1 Q0.25 Q0.5 Q0.75 Q0.9
1    1 20 6.41 1.59      6    6.35 1.93 4.2 9.2     5 0.24    -1.46 0.36  4.39 4.49  5.18    6  7.75 8.44
  Q0.95
1  8.82
------------------------------------------------------------
```

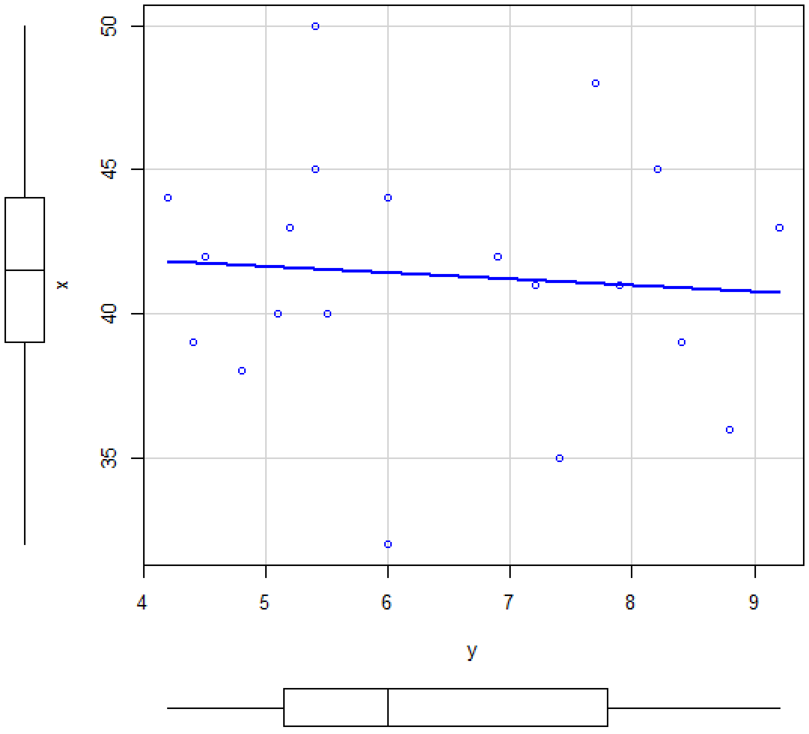
Pearson 相关

```
The Pearson's product-moment correlation test:

        Pearson's product-moment correlation

data:  data$x and data$y
t = -0.33484, df = 18, p-value = 0.7416
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.5036623  0.3769683
sample estimates:
        cor 
-0.07867712 

------------------------------------------------------------
```

Spearman 秩相关

```
The Spearman's rank correlation test:

        Spearman's rank correlation rho

data:  data$x and data$y
S = 1405.3, p-value = 0.8127
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho 
-0.05658252 

------------------------------------------------------------
The scatter plot have been drawn.
```

散点图



从上面七个例子，我想用户已经可以基本管中窥豹，对 EasyDescribe 这个包和 fundescribe()函数有所了解了，EasyDescribe-0.1.2 版本是 EasyDescribe 包的一次重大更新，希望大家喜欢。后面作者还会对这个包继续维护和更新，欢迎大家使用，更欢迎大家提出建议与意见，联系邮箱：niexiuquan1995@foxmail.com。