

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、多重检验校正等等。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,就需要在多得让人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)

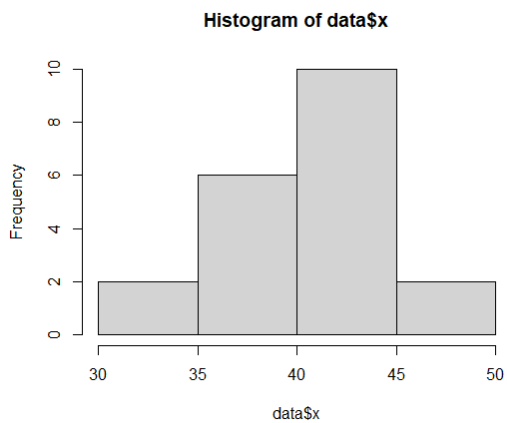
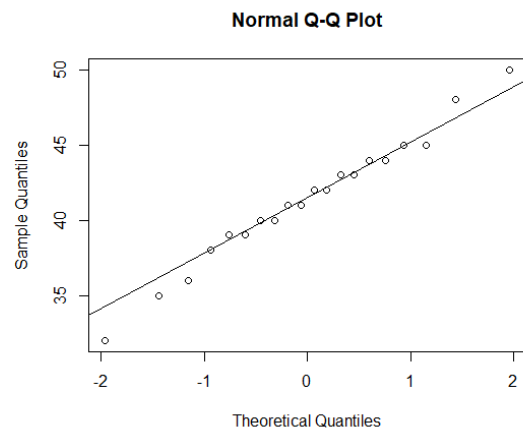
fundescribe() 存在两个基本参数: x 和 y,

x 就是你想分析的基本变量, y 就是对 x 进行分组的分组变量。

数据类型可以基本分成三大类:连续型变量、有序分类变量和无序分类变量,我们在做基本统计分析时进行方法选择,实际上大部分情况下就是在根据数据类型进行方法选择。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方法选择。

比如,你单纯输入了一个连续型变量 fundescribe(data\$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:

```
The histogram and QQ plot of variable x have been drawn.
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(data$gender)`，函数就会自动输出各个分类数量与占比：

```

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(data$age, data$gender)`

```

The histogram and QQ plot of variable x have been drawn.
-----
Two sample t-test:
    Welch Two Sample t-test
data:  x by y
t = 2.3267, df = 4961.6, p-value = 0.02002
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
95 percent confidence interval:
 0.1166846 1.3659512
sample estimates:
mean in group 1 mean in group 2
 54.65754      53.91623
-----
Wilcoxon rank sum test:
Mann-Whitney U test = Wilcoxon rank sum test
    wilcoxon rank sum test with continuity correction
data:  x by y
W = 5718685, p-value = 0.002617
alternative hypothesis: true location shift is not equal to 0
-----
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 7083 54.42 13.04  55.53  54.88 12.3  6.92 94.81 87.89 -0.35  0.09 0.15 30.92 36.84 46.25 55.53 63.08 70.34 74.82
-----
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 4802 54.66 13.46  55.83  55.21 12.47  6.92 94.81 87.89 -0.39  0.15 0.19 29.75 36.26 46.75 55.83 63.37  71 75.49
-----
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 2281 53.92 12.05  55.14  54.22 11.99 16.44 86.27 69.83 -0.25 -0.18 0.25 32.97 37.73 45.66 55.14 62.28 68.45 73.17
-----

```

例 2、x 连续型变量，y 有序分类变量：

fundescribe(age, income, data=data)

```
The histogram and QQ plot of variable x have been drawn.
-----
Variance analysis (one-way ANOVA):
      Df Sum Sq Mean Sq F value Pr(>F)
y      3  14993    4998   29.75 <2e-16 ***
Residuals 7079 1189213    168
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
-----
Kruskal-Wallis rank sum test:

      Kruskal-Wallis rank sum test

data: x by y
Kruskal-Wallis chi-squared = 88.649, df = 3, p-value < 2.2e-16
-----
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 -0.2567654 -1.642604  1.1290734  0.9643674
3-1 -0.9763462 -2.398963  0.4462706  0.2911377
4-1 -4.5323027 -6.158426 -2.9061792  0.0000000
3-2 -0.7195808 -1.636949  0.1977877  0.1822280
4-2 -4.2755373 -5.484671 -3.0664037  0.0000000
4-3 -3.5559565 -4.807073 -2.3048396  0.0000000
-----
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  0.5630095  5.734284e-01  5.734284e-01
2  1 - 3  1.2157638  2.240749e-01  3.361123e-01
3  2 - 3  1.0348356  3.007457e-01  3.608948e-01
4  1 - 4  7.1188952  1.087956e-12  2.175912e-12
5  2 - 4  8.9286762  4.311367e-19  2.586820e-18
6  3 - 4  7.8702772  3.538564e-15  1.061569e-14
-----
The Variance Analysis Trend Test for y:

      The Variance Analysis Trend Test

data: x and y
F.value = 64.336, p-value = 1.219e-15
-----
Descriptive statistical results:
      vars  n mean  sd median trimmed mad min max range skew kurtosis se
1  1 7083 54.42 13.04 55.53 54.88 12.3 6.92 94.81 87.89 -0.35 0.09 0.15
  q0.05 q0.1 q0.25 q0.5 q0.75 q0.9 q0.95
1 30.92 36.84 46.25 55.53 63.08 70.34 74.82
-----
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
1  1  715 55.5 11.59 56.47 55.93 11.02 10.25 94.81 84.56 -0.47 0.86 0.43
  q0.05 q0.1 q0.25 q0.5 q0.75 q0.9 q0.95
1 36.07 41.28 48.23 56.47 63.02 68.69 73.62
-----
group: 2
      vars  n mean  sd median trimmed mad min max range skew kurtosis se
1  1 3005 55.24 12.67 56.13 55.61 12.06 7.17 92.08 84.91 -0.32 0.26 0.23
  q0.05 q0.1 q0.25 q0.5 q0.75 q0.9 q0.95
1 32.72 39.33 47.39 56.13 63.83 70.68 75.3
-----
group: 3
      vars  n mean  sd median trimmed mad min max range skew kurtosis se
1  1 2348 54.52 13.73 56 55.13 12.73 7 89.86 82.86 -0.41 -0.06 0.28
  q0.05 q0.1 q0.25 q0.5 q0.75 q0.9 q0.95
1 29.23 35.25 46.05 56 63.54 71.17 75.25
-----
group: 4
      vars  n mean  sd median trimmed mad min max range skew kurtosis se
1  1 1015 50.97 12.89 51.78 51.14 13.48 6.92 86.27 79.35 -0.16 -0.32 0.4
  q0.05 q0.1 q0.25 q0.5 q0.75 q0.9 q0.95
1 29.68 33.4 41.76 51.78 60.37 67.14 71.89
```

例 3: x 无序分类变量, y 无序分类变量:

fundescribe(gender, exercise, data=data)

```

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

Total Observations in Table:  7083

                                画出R×C列联表

  data$x | data$y
-----|-----|-----|-----|
          | 0 | 1 | Row Total |
-----|-----|-----|-----|
    1 | 1330 | 998 | 2328 |
      | 1336.38967 | 991.61033 | |
      | 0.03055 | 0.04117 |
      | 0.57131 | 0.42869 | 0.32867 |
      | 0.32710 | 0.33079 |
      | 0.18777 | 0.14090 |
-----|-----|-----|-----|
    2 | 2736 | 2019 | 4755 |
      | 2729.61033 | 2025.38967 | |
      | 0.01496 | 0.02016 |
      | 0.57539 | 0.42461 | 0.67133 |
      | 0.67290 | 0.66921 |
      | 0.38628 | 0.28505 |
-----|-----|-----|-----|
Column Total | 4066 | 3017 | 7083 |
              | 0.57405 | 0.42595 |
-----|-----|-----|-----|

Statistics for All Table Factors

                                卡方检验

Pearson's Chi-squared test
-----|
Chi^2 = 0.1068395    d.f. = 1    p = 0.7437714

Pearson's Chi-squared test with Yates' continuity correction
-----|
Chi^2 = 0.09077302    d.f. = 1    p = 0.7631967

Fisher's Exact Test for Count Data                                Fisher精确概率
-----|
Sample estimate odds ratio: 0.9834229

Alternative hypothesis: true odds ratio is not equal to 1
p = 0.758903
95% confidence interval: 0.8885556 1.088572

Alternative hypothesis: true odds ratio is less than 1
p = 0.3814504
95% confidence interval: 0 1.071151

Alternative hypothesis: true odds ratio is greater than 1
p = 0.6378859
95% confidence interval: 0.9029556 Inf

```

例 4: x 无序分类变量, y 无序分类变量:

fundescribe(data\$drink, data\$gender)

```

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

```

Total Observations in Table: 7083

R×C列联表

data\$x	data\$y		Row Total
	1	2	
1	863	204	1067
	350.69547	716.30453	
	748.38701	366.40273	
	0.80881	0.19119	0.15064
	0.37070	0.04290	
	0.12184	0.02880	
2	278	94	372
	122.26684	249.73316	
	198.35974	97.11493	
	0.74731	0.25269	0.05252
	0.11942	0.01977	
	0.03925	0.01327	
3	1187	4457	5644
	1855.03770	3788.96230	
	240.57428	117.78274	
	0.21031	0.78969	0.79684
	0.50988	0.93733	
	0.16758	0.62925	
Column Total	2328	4755	7083
	0.32867	0.67133	

Statistics for All Table Factors

卡方检验

Pearson's Chi-squared test

Chi^2 = 1768.621 d.f. = 2 p = 0

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	p.Chisq	p.adj.Chisq
1 1 : 2	1.41e-02	1.41e-02	0.0131	0.0131	1.44e-02	1.44e-02
2 1 : 3	6.74e-309	2.02e-308	0.0000	0.0000	0.00e+00	0.00e+00
3 2 : 3	3.62e-100	5.43e-100	0.0000	0.0000	3.43e-120	5.14e-120

例 5: x 无序分类变量, y 有序分类变量:

fundescribe(data\$gender, data\$income)

```

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

Total Observations in Table: 7083

RxC列联表

  data$x | data$y
-----|-----|-----|-----|-----|
| 1 | 2 | 3 | 4 | Row Total | |
|---|---|---|---|---|---|
| 1 | 216 | 981 | 752 | 379 | 2328 |
|   | 235.00212 | 987.66624 | 771.72723 | 333.60440 |
|   | 1.53650 | 0.04499 | 0.50428 | 6.17726 |
|   | 0.09278 | 0.42139 | 0.32302 | 0.16280 |
|   | 0.30210 | 0.32646 | 0.32027 | 0.37340 |
|   | 0.03050 | 0.13850 | 0.10617 | 0.05351 | |
|---|---|---|---|---|---|
| 2 | 499 | 2024 | 1596 | 636 | 4755 |
|   | 479.99788 | 2017.33376 | 1576.27277 | 681.39560 |
|   | 0.75225 | 0.02203 | 0.24689 | 3.02432 |
|   | 0.10494 | 0.42566 | 0.33565 | 0.13375 |
|   | 0.69790 | 0.67354 | 0.67973 | 0.62660 |
|   | 0.07045 | 0.28575 | 0.22533 | 0.08979 | |
|---|---|---|---|---|---|
| Column Total | 715 | 3005 | 2348 | 1015 | 7083 |
|   | 0.10095 | 0.42426 | 0.33150 | 0.14330 |
-----|-----|-----|-----|-----|

Statistics for All Table Factors

卡方检验

Pearson's Chi-squared test

-----
Chi^2 = 12.30852    d.f. = 3    p = 0.006397676

-----

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

秩和检验

    Wilcoxon rank sum test with continuity correction

data:  yn by x
W = 5715198, p-value = 0.01758
alternative hypothesis: true location shift is not equal to 0

-----

Post hoc multiple comparisons between different groups of y:
两两比较多重检验
  Comparison p.Fisher p.adj.Fisher p.Gtest p.adj.Gtest p.Chisq p.adj.Chisq
1 1 : 2 0.21300 0.32000 0.20800 0.31200 0.22700 0.34000
2 1 : 3 0.38300 0.46000 0.35900 0.43100 0.38500 0.46200
3 1 : 4 0.00241 0.00861 0.00203 0.00867 0.00250 0.00942
4 2 : 3 0.63800 0.63800 0.63100 0.63100 0.65200 0.65200
5 2 : 4 0.00722 0.01440 0.00655 0.01310 0.00705 0.01410
6 3 : 4 0.00287 0.00861 0.00289 0.00867 0.00314 0.00942

-----

The Cochran-Armitage trend test for y:

趋势性检验

    The Cochran-Armitage Trend Test

data:  The type of data is variable!
Z = 2.169, p-value = 0.0301

```

从上面五个例子，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，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。