丁香园临床科研方法学课程：跟我学R语言 视频课程

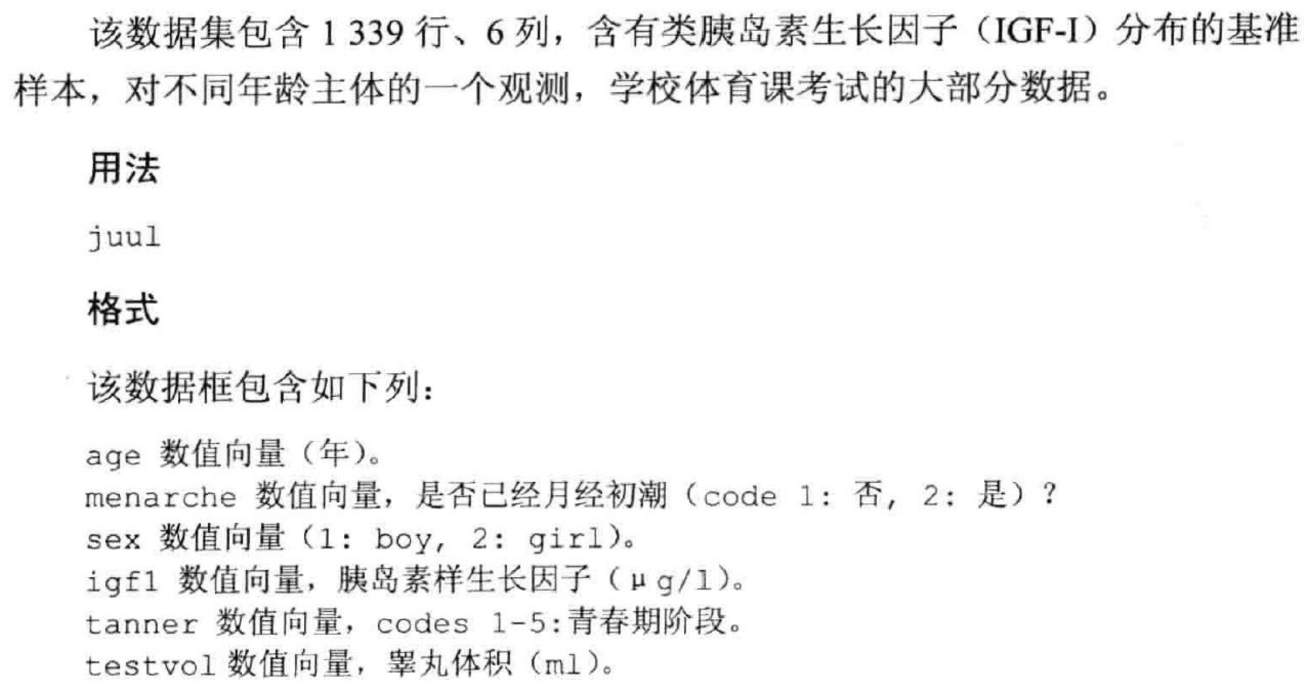
第三章 统计描述与基础统计分析方法

复旦大学附属肿瘤医院 周支瑞

3.1 描述性统计分析

单组数据汇总统计量

- 均数、标准差、中位数、分位数计算
- 统计描述过程中的缺失值处理
- ISwR包中的 juul 数据集



该数据集包含 1 339 行、6 列，含有类胰岛素生长因子（IGF-I）分布的基准样本，对不同年龄主体的一个观测，学校体育课考试的大部分数据。

用法

```
juul
```

格式

该数据框包含如下列：

age 数值向量（年）。
menarche 数值向量，是否已经月经初潮（code 1：否，2：是）？
sex 数值向量（1：boy，2：girl）。
igf1 数值向量，胰岛素样生长因子（μg/l）。
tanner 数值向量，codes 1-5：青春期阶段。
testvol 数值向量，睾丸体积（ml）。

数据分布类型的图形描述

- 直方图
- 经验累积分布图形
- Q-Q图
- 箱式图

分组数据汇总统计量

- tapply: 分组计算统计量
- aggreate和by函数：分组计算统计量

分组数据的图形描述

- 直方图
- 并联箱式图
- 带状图

表格

- 生成表格
- 边际表格和频数

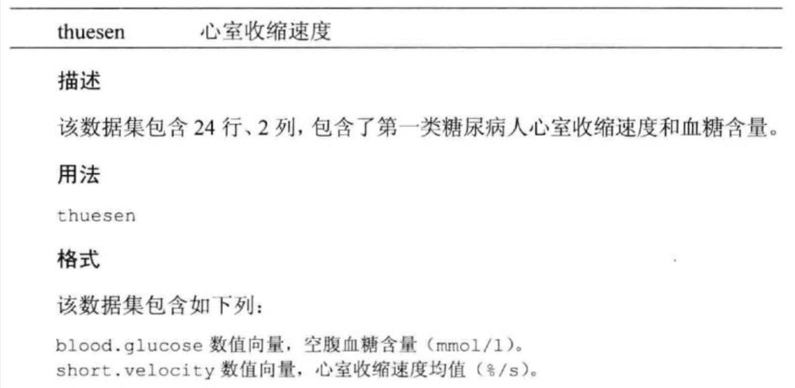
表格的图形展示

- 条形图
- 点图
- 饼图

3.2 相关 分析

三种相关系数

- Pearson相关系数
- Spearman秩相关
- Kendall相关系数



thuesen　　　心室收缩速度

描述

该数据集包含 24 行、2 列，包含了第一类糖尿病人心室收缩速度和血糖含量。

用法

```
thuesen
```

格式

该数据集包含如下列：

`blood.glucose` 数值向量，空腹血糖含量（mmol/1）。
`short.velocity` 数值向量，心室收缩速度均值（%/s）。

- cor()函数可以计算这三种相关系数，而cov()函数可用来计算协方差。两个函数的参数有很多，其中与相关系数的计算有关的参数可以简化为：cor(x, use= , method=)

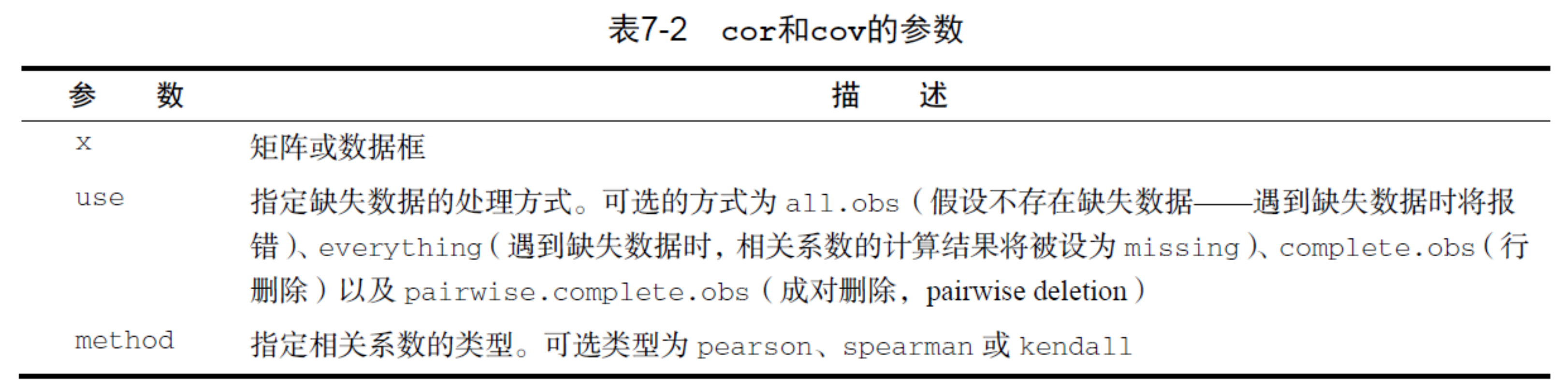


表7-2　cor和cov的参数

参　　数	描　　述
x	矩阵或数据框
use	指定缺失数据的处理方式。可选的方式为 all.obs（假设不存在缺失数据——遇到缺失数据时将报错）、everything（遇到缺失数据时，相关系数的计算结果将被设为 missing）、complete.obs（行删除）以及 pairwise.complete.obs（成对删除，pairwise deletion）
method	指定相关系数的类型。可选类型为 pearson、spearman 或 kendall

相关系数的显著性检验

- 可以使用cor.test()函数对单个的Pearson、Spearman和Kendall相关系数进行检验。简化后的使用格式为：cor.test(x, y, alternative = , method =)
- 其中的x和y为要检验相关性的变量，alternative则用来指定进行双侧检验或单侧检验（取值为"two.side"、"less"或"greater"），而method用以指定要计算的相关类型（"pearson"、"kendall" 或"spearman" ）。当研究的假设为总体的相关系数小于0 时， 请使用alternative="less" 。在研究的假设为总体的相关系数大于0 时， 应使用alternative="greater"。在默认情况下，假设为alternative="two.side"（总体相关系数不等于0）。

偏相关

- 偏相关是指在控制一个或多个定量变量时，另外两个定量变量之间的相互关系。你可以使用ggm包中的pcor()函数计算偏相关系数。ggm包没有被默认安装，在第一次使用之前需要先进行安装。函数调用格式为：pcor(u, S) 其中的u是一个数值向量，前两个数值表示要计算相关系数的变量下标，其余的数值为条件变量（即要排除影响的变量）的下标。S为变量的协方差阵。

```
# partial correlations
library(ggm)
# partial correlation of population and murder rate, controlling
# for income, illiteracy rate, and HS graduation rate
pcor(c(1,5,2,3,6), cov(states))
```

3.3 组间差异t检验

单样本t检验

- Here is an example concerning daily energy intake in kJ for 11 women (Altman, 1991, p. 183). First, the values are placed in a data vector:

```
< daily.intake <- c(5260,5470,5640,6180,6390,6515,
+ 6805,7515,7515,8230,8770)
```

两样本t检验

- We return to the daily energy expenditure data and consider the problem of comparing energy expenditures between lean and obese women.
- The factor stature contains the group and the numeric variable expend the energy expenditure in mega-Joules

方差齐性检验

- Even though it is possible in R to perform the two-sample t test without the assumption that the variances are the same, you may still be interested in testing that assumption, and R provides the var.test function for that purpose, implementing an F test on the ratio of the group variances. It is called the same way as t.test

配对样本t检验

- The data on pre- and postmenstrual energy intake in a group of women. There data are entered from the command line, but they are also available as a data set in the ISwR package

3.4 组间差异秩和检验

单样本Wilcoxon符号秩检验

- The t tests are fairly robust against departures from the normal distribution especially in larger samples, but sometimes you wish to avoid making that assumption. To this end, the distribution-free methods are convenient. These are generally obtained by replacing data with corresponding order statistics.

两样本Wilcoxon符号秩检验

- You might prefer a nonparametric test if you doubt the normal distribution assumptions of the t test. The two-sample Wilcoxon test is based on replacing the data by their rank (without regard to grouping) and calculating the sum of the ranks in one group, thus reducing the problem to one of sampling n1 values without replacement from the numbers 1 to n1 +n2. This is done using wilcox.test, which behaves similarly to t.test

配对Wilcoxon符号秩检验

- The paired Wilcoxon test is the same as a one-sample Wilcoxon signedrank test on the differences. The call is completely analogous to t.test

练习题

- 5.2 In the dataset ***vitcap***, use a t test to compare the vital capacity for the two groups. Calculate a 99% confidence interval for the difference. The result of this comparison may be misleading. Why?

  t.test(vital.capacity~group,conf=0.99,data=vitcap)

  The fact that age also differs by group may cause bias.

- 5.3 Perform the analyses of the ***react*** and ***vitcap*** data using nonparametric techniques.

  This is quite parallel to t.test usage

  wilcox.test(react)

  wilcox.test(vital.capacity~group, data=vitcap)

主要参考文献

【1】Robert I. Kabacoff 著, 《R语言实战 》(第2版), 人民邮电出版社, 2016

【2】Peter Dalgaard 著, 《R语言统计入门》》(第2版), 人民邮电出版社, 2014

【3】薛毅 陈立萍 著,《R语言实用教程》, 清华大学出版社, 2014

【4】张铁军 陈兴栋 刘振球 著,《R语言与医学统计图形》, 人民卫生出版社, 2018

Thanks！

感谢您的观看！