南開大學

第四讲 基本做图及相关性分析

Need to think

Programming isn't about what you know

It's about what you can figure out

提　纲

1.基本绘图

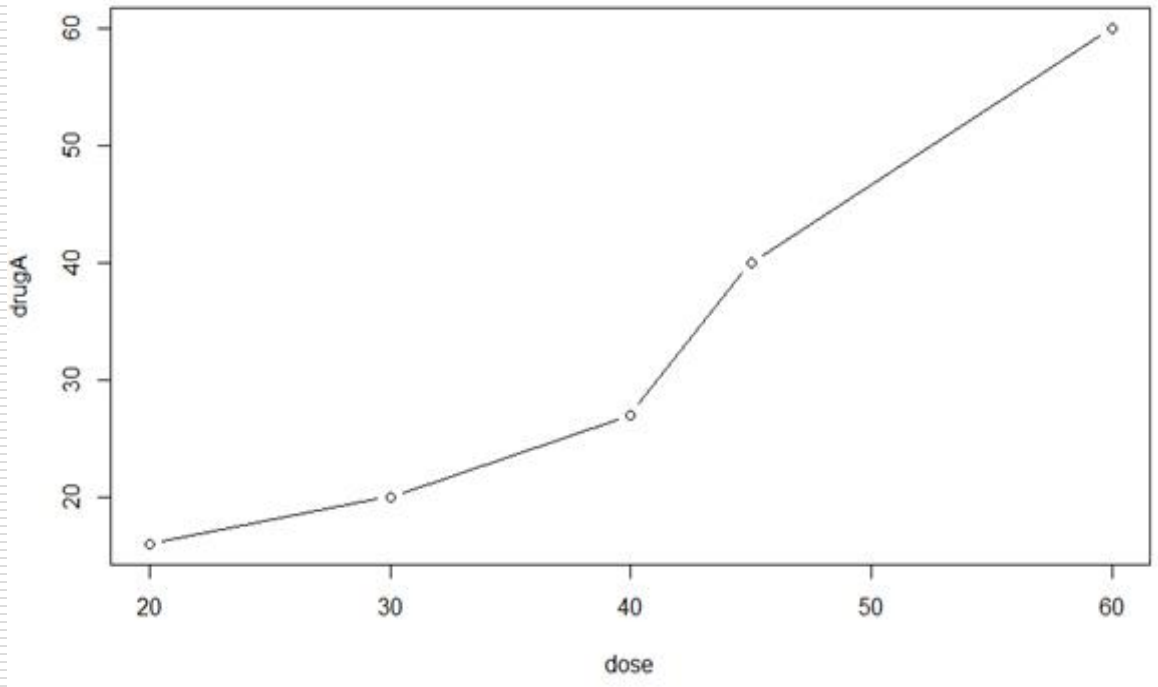
R是一个惊艳的图形构建平台。plot()是R中为对象作图的一个泛型函数（它的输出将根据所绘制对象类型的不同而变化）。

【例】假定一个数据集描述了病人对两种药物五个剂量水平上的响应。请可视化这些数据。

剂　量	对药物A的响应	对药物B的响应
20	16	15
30	20	18
40	27	25
45	40	31
60	60	40

1.基本绘图

```
dose <- c(20, 30, 40, 45, 60)
drugA <- c(16, 20, 27, 40, 60)
drugB <- c(15, 18, 25, 31, 40)
plot(dose, drugA, type = "b")
```



图形类型type：

“p”点图（默认）

“l”线图

“b”点线图，线不穿过点

“c”虚线图

“o”点线图，线穿过点

“h”直方图

“s”阶梯图

“S”步骤图

“n”无图

【练习】感受plot作图

（3分钟）

1.基本绘图

【操作练习】感受plot作图 （3分钟）

作的图如何保存成文件？

1.基本绘图

- 图形参数

可以通过修改称为图形参数的选项来自定义一幅图形的多个特征（字体、颜色、坐标轴、标题）。修改图形参数的一种方法是通过函数**par()**来指定这些选项。其调用格式为：

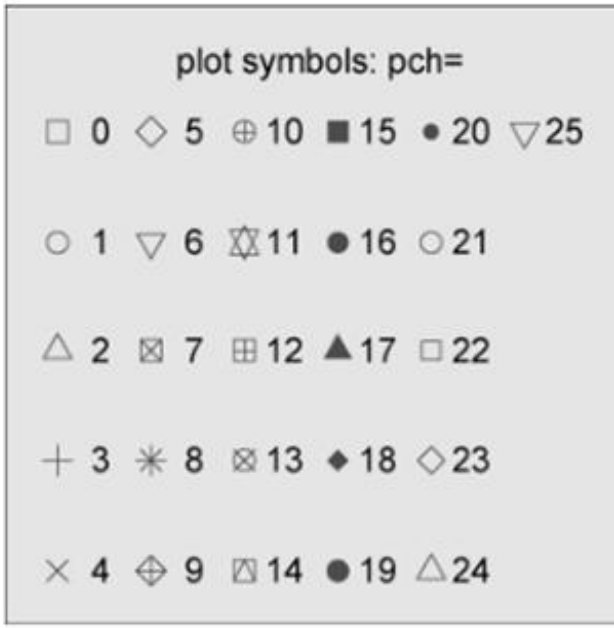
par(optionname=value,…, no.readonly=FALSE)

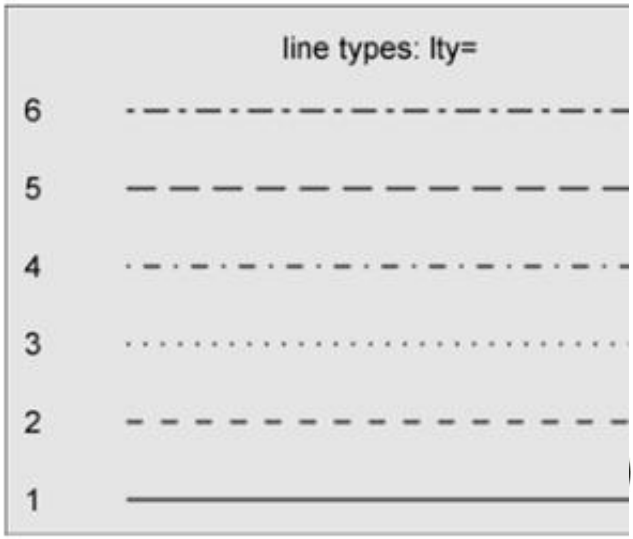
如果不加参数执行**par()**将生成一个当前图形参数设置的列表。参数**no.readonly=TRUE**可以生成一个可以修改的当前图形参数列表。

1.基本绘图

符号和线条类型

参　数	描　述
pch	指定绘制点时使用的符号
cex	指定符号的大小。cex是一个数值，表示绘图符号相对于默认大小的缩放倍数。默认大小为1，1.5表示放大为默认值的1.5倍，0.5表示缩小为默认值的50%，等等
lty	指定线条类型
lwd	指定线条宽度。lwd是以默认值的相对大小来表示的（默认值为1）。例如，lwd=2将生成一条两倍于默认宽度的线条





颜色、文本格式、图形尺寸和边界大小等也有相应的参数，使用时根据需要查表即可。

1.基本绘图

【操作练习】使用图形参数控制图形外观（5分钟）。

```
opar <- par(no.readonly = TRUE) #记录原始参数
par(pin = c(2, 3))    #设置当前图的长宽
par(lwd = 2, cex = 1.5)
par(cex.axis = 0.75, font.axis = 3)
plot(dose, drugA, type = "b", pch = 9, lty = 2, col =
"red")
plot(dose, drugB, type = "b", pch = 25, lty = 6, col =
"blue")
par(opar)  #恢复原始参数
```

1.基本绘图

【操作练习】带标题等的图形示例（5分钟）。

```
plot(dose, drugA, type = "b", col = "red", lty = 2,
   pch = 2, lwd = 2, main = "Clinical Trials for Drug
A",  sub = "This is hypothetical data",
   xlab = "Dosage", ylab = "Drug Response", xlim
= c(0, 60), ylim = c(0, 70))
```

其他的请同学们需要时自己看资料：

https://www.jianshu.com/p/36d135173a54

1.基本绘图

- 添加文本、自定义坐标轴、参考线和图例等

除了图形参数，许多高级绘图函数（例如plot、hist、boxplot）也允许自行设定坐标轴和文本标注选项。例如在图形上添加了标题（main）、副标题（sub）、坐标轴标签（xlab、ylab）并指定了坐标轴范围（xlim、ylim）。某些高级绘图函数已经包含了默认的标题和标签。你可以通过在plot()语句或单独的par()语句中添加ann=FALSE来移除它们。

1.基本绘图

- 图形的组合

使用函数**par()**或**layout()**可以很容易地将多幅图形组合为一幅总括图形。

可以在**par()**函数中使用图形参数**mfrow=c(*nrows*, *ncols*)**来创建按行填充的、行数为***nrows***、列数为***ncols***的图形矩阵。另外，可以使用**nfcol=c(*nrows*, *ncols*)**按列填充矩阵。

1.基本绘图

【操作练习】 图形组合par()（5分钟）。

```
attach(mtcars)
opar <- par(no.readonly = TRUE)
par(mfrow = c(2, 2))
plot(wt, mpg, main = "Scatterplot of wt vs. mpg")
plot(wt, disp, main = "Scatterplot of wt vs disp")
hist(wt, main = "Histogram of wt")
boxplot(wt, main = "Boxplot of wt")
par(opar)
opar <- par(no.readonly = TRUE)
par(mfrow = c(3, 1))
hist(wt)
hist(mpg)
hist(disp)
par(opar)
detach(mtcars)
```

1.基本绘图

- layout（略）
- 图形布局的精细控制（略）

2. 相关性分析

【提出问题】

R基础安装中的state.x77数据集，它提供了美国50个州在1977年的人口、收入、文盲率、预期寿命、谋杀率和高中毕业率数据。数据集中还收录了气温和土地面积数据（同学们可以使用help(state.x77)查看数据集）。

2. 相关性分析

利用R探究：

- 收入和预期寿命的相关性如何？它是否明显不为零（显著相关）？
- 不同地区的差别是否在统计上显著？
- 不同地区的文盲率是否一样？

……

2. 相关性分析

【主要解决的问题】

分析数据集中不同组数据的相关性，主要包括：

- 相关系数的计算
- 基于相关系数的显著性检验

2. 相关性分析

【基本概念】

相关分析是研究两个或两个以上处于同等地位的随机变量间的相关关系的统计分析方法。例如，人的身高和体重之间；空气中的相对湿度与降雨量之间的相关关系都是相关分析研究的问题。

相关系数用来描述变量之间的相关关系。

2. 相关性分析

相关系数的符号（±）表明关系的方向（正相关或负相关），其值的大小表示关系的强弱程度（完全不相关时为0，完全相关时为1）。

R可以计算多种相关系数，包括Pearson相关系数、Spearman相关系数、Kendall相关系数、偏相关系数、多分格（polychoric）相关系数和多系列（polyserial）相关系数。

2. 相关性分析

◆ Pearson积差相关系数

Pearson积差相关系数衡量了两个变量之间的线性相关程度。它具有+1和-1之间的值，其中1是总正线性相关性，0是非线性相关性，并且-1是总负线性相关性。计算公式：

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

https://baike.baidu.com/item/%E7%A7%AF%E5%B7%AE%E7%9B%B8%E5%85%B3/1278536

1. 相关性分析

◆ **Spearman等级相关系数**

两个连续变量间不满足Pearson积差相关分析的适用条件时，使用Spearman等级相关系数来描述。Spearman等级相关系数衡量分级定序变量之间的相关程度。计算公式：

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

https://baike.baidu.com/item/spearman%E7%9B%B8%E5%85%B3%E7%B3%BB%E6%95%B0?fromModule=lemma_search-box

1. 相关性分析

◆ 偏相关系数

偏相关是指在控制一个或多个变量时，另外两个变量之间的相互关系，又称净相关或部分相关。例如我们可以通过控制收入、文盲率和高中毕业率的影响，研究人口和谋杀率之间的相关系数。

2. 相关性分析

◆ 显著性检验

在计算好相关系数以后，需要对它们进行统计显著性检验。为判断样本相关系数对总体相关程度的代表性，需要对相关系数进行显著性检验。

若在统计上是显著的，说明它可以作为总体相关程度的代表值，否则不能作为总体相关程度的代表值。

2. 相关性分析

显著性检验（significance test）：就是事先对总体（随机变量）的参数或总体分布形式做出一个假设，然后利用样本信息来判断这个假设是否合理，即判断总体的真实情况与原假设是否有显著性差异。或者说，显著性检验要判断样本与我们对总体所做的假设之间的差异是纯属机会变异，还是由我们所做的假设与总体真实情况之间不一致所引起的。

2. 相关性分析

显著性检验是针对我们对总体所做的假设做检验，其原理就是“小概率事件实际不可能性原理”来接受或否定假设。常用的原假设为变量间不相关（即总体的相关系数为0）。

协方差表示的是两个变量总体误差的期望。协方差为0的两个随机变量称为是不相关的。

2. 相关性分析

【使用R的具体处理过程及结果解读】

- cor()函数可以计算三种相关系数，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=)

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

默认参数为use="everything"和method="pearson".

关于use:https://www.jianshu.com/p/6b9265bce085

2. 相关性分析

- **Pearson相关系数的变量要求：**

  ①两变量相互独立

  ②两变量为连续变量

  ③两变量的分布遵循正态分布

  ④两变量呈线性关系

【读懂Pearson相关分析结果：

https://blog.csdn.net/cheyennelam/article/details/62227477】

2. 相关性分析

当研究的假设为总体的相关系数小于0时，使用alternative=“less”。当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）。

cor.test每次只能检验一种相关关系。

2. 相关性分析

- 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"）。

2. 相关性分析

【操作练习】计算变量之间的相关系数（5分钟）

```
states<- state.x77[,1:6]
states
cov(states) #方差和协方差，在默认情况下得到的结果是一个方阵（所有变量之间两两计算相关）
cor(states) # 默认计算Pearson积差相关系数
cor(states, method="spearman")# Spearman等级相关系数
#计算非方形的相关矩阵
x <- states[,c("Population", "Income", "Illiteracy", "HS Grad")]
y <- states[,c("Life Exp", "Murder")]
cor(x,y)
```

2. 相关性分析

【操作练习】计算变量之间的偏相关系数（5分钟）

```
# 在控制了收入、文盲率和高中毕业率时人口和谋杀率的偏相关系数， ggm包中的pcor()函数计算偏相关系数。
install.packages("ggm")
library(ggm)
pcor(c(1,5,2,3,6), cov(states))#计算偏相关系数
```

```
0.346
```

在控制了收入、文盲率和高中毕业率的影响时，人口和谋杀率之间的相关系数为 0.346。偏相关系数常用于社会科学的研究中。

2. 相关性分析

【操作练习】显著性检验（5分钟）

```
cor.test(states[,3], states[,5])# 检验某种相关系数的显著性。
        Pearson's product-moment correlation data:
states[, 3] and states[, 5]
t = 6.8479, df = 48, p-value = 1.258e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval: 0.5279280 0.8207295
sample estimates:
cor 0.7029752
```

检验预期寿命和谋杀率的Pearson相关系数为0的原假设。假设总体的相关度为0，则预计在一千万次中只会有少于一次的机会见到0.703这样大的样本相关度（即p = 1.258e-08）。由于这种情况几乎不可能发生，所以我们可以拒绝原假设，从而支持了要研究的猜想，即预期寿命和谋杀率之间的总体相关度不为0。

2. 相关性分析

【操作练习】计算states变量的相关系数和显著性（8分钟）

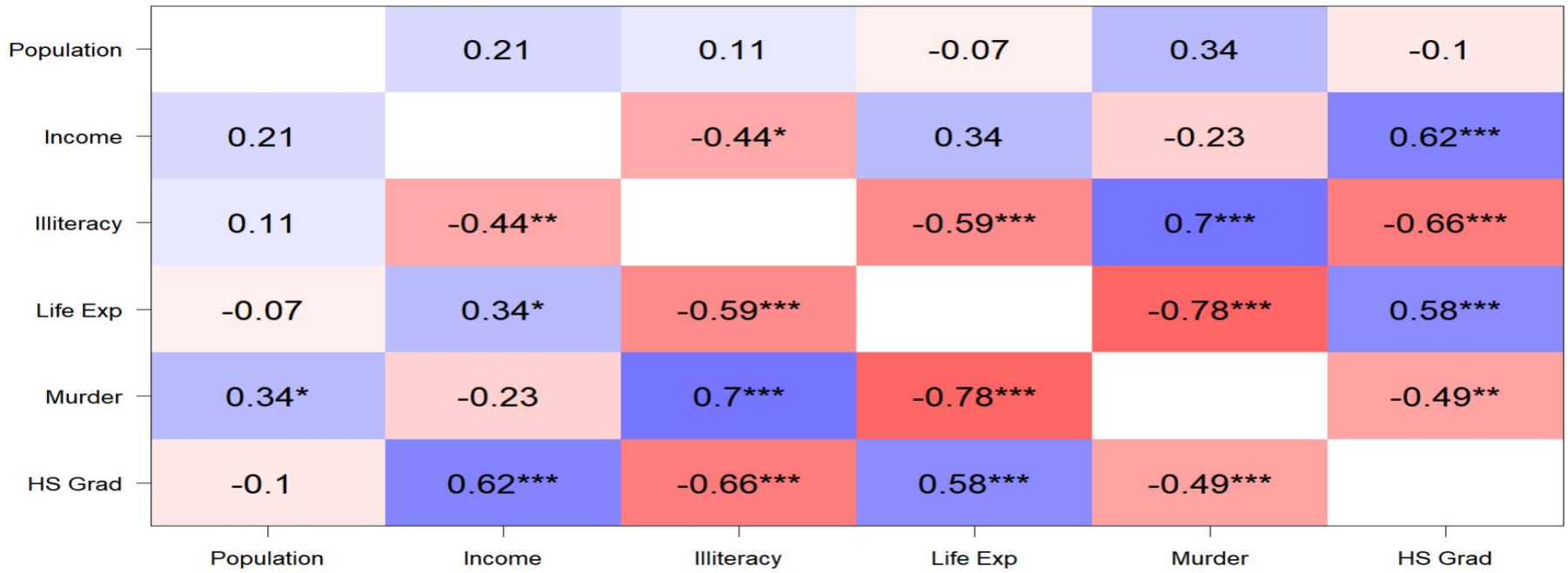
```
install.packages("psych")
library(psych) #加载R包
corr.test(states, use="complete")#计算相关系数及其显著性
```

通过corr.test计算相关矩阵并进行显著性检验。参数use=的取值可为"pairwise"或"complete"（分别表示对缺失值执行成对删除或行删除）。参数method=的取值可为"pearson"（默认值）、"spearman"或"kendall"。

结果：人口数量和高中毕业率的相关系数（-0.10）并不显著地不为0（p = 0.5）。

2. 相关性分析

```
data <- corr.test(states      )#可以绘制相关系数图
r <- data$r
p <- data$p
corPlot(r,pval=p,numbers=TRUE,diag=FALSE,stars=TRUE)
```



3. 相关性分析练习

【操作练习】

某科学基金会的管理人员欲分析从事数学研究工作的中等或较高水平的数学家的年工资额 y 与他们的研究成果(论文、著作)的质量指标 x_1 ,从事研究工作的时间 x_2 以及能成功获得资助的指标 x_3 之间的关系，为此按一定的试验设计方法调查了24位此类型的数学家，得到的数据见“数学家.xls”

	x1	x2	x3	y
2	3.50	9.00	4.00	33.20
3	5.30	20.00	6.00	40.30
4	5.10	18.00	5.90	38.70
5	5.80	33.00	6.40	46.80
6	4.20	31.00	5.00	41.40
7	6.00	13.00	6.70	37.50
8	6.80	25.00	7.50	39.00
9	5.50	30.00	6.00	40.70
10	3.10	5.00	3.50	30.10
11	7.20	47.00	8.00	52.90
12	4.50	25.00	5.00	38.20
13	4.90	11.00	5.80	31.80
14	8.00	23.00	8.30	43.30
15	6.50	35.00	7.00	44.10
16	6.60	39.00	7.40	42.80
17	3.70	21.00	4.30	33.60
18	6.20	7.00	7.00	34.20
19	7.00	40.00	7.60	48.00
20	4.00	35.00	4.90	38.00
21	4.50	23.00	5.00	35.90
22	5.90	33.00	6.40	40.40
23	5.60	27.00	6.10	36.80
24	4.80	34.00	5.50	45.20
25	3.90	15.00	4.40	35.10

3. 相关性分析练习

- 计算y，x_1，x_2和x_3的Pearson系数，并进行显著性分析。
- 根据分析结果，进行适当的偏相关分析。

提示：读Excel文件方法

- 保存为CSV格式文件，然后读CSV 文件
- data1<-read.csv("数学家.csv",header = TRUE)

课下作业

1、重复练习（但不限于）课上练习的内容，将操作界面截图到Word文档，发word文档到学堂云交作业。

2、你和你的小伙伴对获取的数据进行：

- 均值和标准差分析
- 数据可视化（作图）
- 相关性分析