使用R进行常见的统计分析

Alex / 2020-05-15 / free_learner@163.com / learning-archive.org

更新于2023-08-26，主要是文字排版上的更新，内容基本保持不变。

汇总一下使用R进行常见的统计分析的方法，包括相关分析、T检验、方差分析等。20230826更新 过去几年也一直在坚持学习统计学，但是始终感觉没有入门。以下内容请特别谨慎参考。

一、衡量两个变量的线性关系

1. 皮尔逊相关（Pearson's correlation coefficient）

```
cor.test(x, y)
## 计算变量x和y的相关系数以及显著性等。
## 通过设置method参数可以计算斯皮尔曼相关和肯德尔相关。
```

2. 偏相关（partial correlation）

```
pcor.test(x, y, z)
## 使用ppcor包计算在排除变量z的影响后，计算x和y的相关系数。
## z是包含一个或多个变量的数据框。
```

3. 比较相关系数

如果有两个独立样本的皮尔逊相关系数（R1和R2分别是两个样本得到的相关系数，样本量分别为N1和N2），想要检验这两个相关系数是否有显著差异，可以使用下面的方法：

```
Z1 <- atanh(R1); Z2 <- atanh(R2);
## 对相关系数进行Fisher-Z转换使得采样分布为正态，atanh=0.5ln((1+R)/(1-R))。
Zdiff <- (Z1 - Z2)/sqrt(1/(N1-3)+1/(N2-3))
## sqrt表示开方运算。
p <- 2*(1-pnorm(abs(Zdiff)))
## pnorm表示标准正态分布的累积密度函数（cumulative density function）。
```

4. 多重线性回归（multiple linear regression）

```
lm(y ~ x1+x2)
## 当只有一个自变量时，线性回归等价于皮尔逊相关系数。
## 多重线性回归和相关的区别在于（假设我们关心的是y和x1的关系），在考虑y和x1的关系的时候，排除了x2的影响。
## 20230826更新 多重线性回归和偏相关p值是一致的，只是前者得到的是回归系数，而后者得到的是相关系数。
```

二、比较两组的均值

1. T检验

```
t.test(y ~ x)
## y表示观测值，x表示一个两个水平的因子变量。
## 通过设置paired参数可以进行配对T检验。
## 默认使用的是Welch's t-test，可用于校正方差不齐。
```

```
lm(y ~ x)
## 除了使用t.test，也可以使用lm函数比较两组的均值差异。
## 这种方法的好处是可以添加协变量。
```

2. 非参数方法

```
wilcox.test(y ~ x)
## Wilcoxon rank-sum test
## 通过设置paired参数可以用于配对数据，这个时候称为Wilcoxon signed-rank test。
## 这种非参数方法的思路是将数据转换为排序(rank)，然后对排序进行分析。
```

三、比较三组（或以上）的均值

1. 单因素方差分析（one-way ANOVA）

```
lm(y ~ x)
## y表示观测值，x表示一个三个（或以上）水平的（被试间）因子变量。
```

```
aov(y ~ x)
## aov对lm进行了封装，输出上更符合方差分析的形式。
```

```
oneway.test(y ~ x)
## Welch's F-ratio，可用于校正方差不齐。
```

```
## 如果x是一个有顺序的分类变量，可以检测是否存在线性/二次/三次等趋势，需要设置contrast。
contrasts(x) <- contr.poly(n)，n表示因子水平数目。
```

2. 事后检验（post-hoc tests）

方差分析只能告诉我们三组均值之间是否有显著差异，但是不能说明差异出现在哪两组之间，因此需要做事后检验。事后检验就是在任意两组之间进行T检验，并且需要校正多次比较引起的第一类错误概率的增高。

```
pairwise.t.test(y, x, p.adjust.method='bonferroni')
## p.adjust.method对p值进行多重比较校正，有一系列的方法
```

3. 双因素方差分析（two-way ANOVA）

```
contrast(x1) <- contr.helmert(n1)
## x1表示一个被试间分类变量，设置正交的contrast，n1表示水平数目。
contrast(x2) <- contr.helmert(n2)
## x2表示一个被试间分类变量，设置正交的contrast，n1表示水平数目。
model <- lm(y ~ x1+x2+x1:x2)
Anova(model, type="III")
## Anova函数需要car包，ANOVA分为I/II/III三种类型。如果进行第三种类型的ANOVA，需要正交的contrast。在ANOVA之后，同样需要做事后检验。
```

4. 协方差分析（ANCOVA）

```
contrast(x1) <- contr.helmert(n1)
## x1表示一个被试间分类变量，设置正交的contrast。
model <- lm(y ~ x1+x2)
## x2表示一个连续变量，假设x1和x2没有交互作用（可以加上交互项检验一下）。
model <- Anova(model, type="III")
summary(glht(model, linfct=mcp(x1="Tukey")))
## glht函数来自于multcomp包，在协方差分析中需要对排除x2影响后的均值（adjusted mean）进行事后检验。
```

5. 重复测量方差分析（repeated-measures ANOVA）

```
contrast(x1) <- contr.helmert(n1)
## x1表示一个被试内分类变量，设置正交的contrast。
contrast(x2) <- contr.helmert(n2)
## x2表示一个被试内分类变量，设置正交的contrast。
ezANOVA(data=df, dv=.(y), wid=.(subject), within=.(x1,x2), detailed=TRUE,
type=3)
## ezANOVA函数来自ez包，df表示存放变量的数据框，wid表示被试变量（即被试名是分类变量），
within表示被试内变量。
## 处理被试内变量也可以使用multi-level linear model，我现在还没有太理解该模型。
```

6. 混合设计方差分析（mixed design ANOVA）

```
contrast(x1) <- contr.helmert(n1)
## x1表示一个被试内分类变量，设置正交的contrast。
contrast(x2) <- contr.helmert(n2)
## x2表示一个被试间分类变量，设置正交的contrast。
ezANOVA(data=df, dv=.(y), wid=.(subject), within=.(x1), between=.(x2),
detailed=TRUE, type=3)
```

参考

Andy Field, Jeremy Miles, Zoe Field. (2012). Discovering Statistics Using R.