

第6章 统计分析基础

1. 请简答常用描述性统计的方法有哪些, R 语言中常用描述性统计的函数有哪些, 以及这些函数的用法。

【参考答案】描述性统计, 是指运用制表、分类、图形以及计算概括性数据来描述数据特征的各项活动。描述性统计分析要对调查总体所有变量的有关数据进行统计性描述, 主要包括数据的频数分析、集中趋势分析、离散程度分析、分布以及一些基本的统计图形。

`summary()` 函数提供了最小值、最大值、四分位数和数值型变量的均值, 以及因子向量和逻辑型向量的频数统计。`stat.desc()` 函数计算种类繁多的描述性统计量。`Hmisc` 包中的 `describe()` 函数可以返回变量和观测变量的缺失值和唯一值的数目、平均值、分位数、以及五个最大值和五个最小值。`aggregate()` 函数分组获取描述性统计量。`doBy` 包中的 `summaryBy()` 函数分组计算描述性统计量。`psych` 包中的 `describeBy()` 函数分组计算描述性统计量。

2. 请简答常用统计推断的方法有哪些, R 语言中常用统计推断的函数有哪些, 以及这些函数的用法。

【参考答案】统计推断(statistical inference), 是指根据带有随机性的观测数据(样本)以及问题的条件和假定(模型), 而对未知事物作出的以概率形式表述的推断。它是数理统计学的主要任务, 其理论和方法构成数理统计学的主要内容。

统计推断是从总体中抽取部分样本, 通过对抽取部分所得到的带有随机性的数据进行合理的分析, 进而对总体作出科学的判断。它是伴随着一定概率的推测。统计推断的基本问题可以分为两大类: 一类是参数估计问题; 另一类是假设检验问题。这些都需要从总体中抽取样本, 通过对样本观察值分析来估计和推断, 即根据样本来推断总体分布的未知参数, 称为参数估计。参数估计有两种方法: 点估计和区间估计。

t 检验是极常用的检验方法, 除了单样本推断之外, `t.test` 命令还可以实现两样本推断和配对样本推断。如果要对总体比率或总体方差进行推断, 则可以使用 `prop.test` 和 `var.test`。 t 检验的前提条件是总体服从正态分布, 因此我们有必要先检验正态性。而且在评价回归模型时, 对残差也需要检验正态性。检验正态性的函数是 `shapiro.test`。

利用非参数方法推断中位数。`wilcoxon.test` 函数可实现符号秩检验。

卡方分布有一个重要应用就是根据样本数据来检验两个分类变量的独立性, 使用 `chisq.test` 函数。

3. 请简答 p 值、检验统计量与拒绝域的作用。

【参考答案】由于抽样误差的存在, 样本观察值与原假设中的期望值往往是不一致的, 但只有在两者之间的差异达到一定程度时, 才能有把握地认定这种误差不是由随机误差引起的; 当两者之间的误差达到某一临界值时, 就将这一观测结果称为“显著的”, 这一结果是非偶然的, 即不是由随机因素能解释的, 通过计算在原假设成立的前提下, 随机样本出现备择假设方向(小于、等于或大于)某一特定观察值的概率, 并与显著性水平比较, 从而确定这一观测结果是否显著, 这就是所谓的 p 值, 如果 $p \leq \alpha$, 则认为这一事件是显著

的,如果观察值极端的概率竟然比小概率的标准还小,那么这个事件肯定是小概率事件,所以认为这一结果是显著的,即否定这一结果是由随机误差引起的,从而应当修正原来的期望拒绝原假设,如果 $p > \alpha$,则认为这一事件是不显著的,因为它还是可以用随机误差来解释,就没有必要修正原来的期望不拒绝原假设。

4. 请简答 t 检验与正态性检验的用途以及两者之间的区别。

【参考答案】 t 检验,主要用于检验样本含量较小(例如 $n < 30$)、总体标准差 σ 未知的正态分布。 t 检验是用 t 分布理论来推论差异发生的概率,从而比较两个平均数的差异是否显著。总体单侧 t 检验是检验一个样本平均数与一个已知的总体平均数的差异是否显著;总体双侧 t 检验是检验两个样本平均数与其各自所代表的总体的差异是否显著。

总体分布的形态通常是未知的。在检验之前对总体数据是否符合正态分布作出检验,通常叫正态性检验。正态性检验通常属于非参数检验, `shapiro.test()` 命令可用于样本容量小于 5000 时的单变量总体正态性检验。正态性检验原假设都是所检验的数据来自正态总体,备择假设则是所检验的数据来自非正态总体。

5. 编程题:某老师开设了计算机网络原理和计算机程序设计两门课程,为了解学生对这两门课程喜欢上的差异,在学期末对 53 名学生进行了数据调查,得到下列数据:

喜欢计算机网络原理的人数: 24 人

喜欢计算机程序设计的人数: 22 人

不能区分的人数: 7 人

分析在显著性水平 $\alpha = 0.05$ 下,能否认为这两门课程被学生喜欢的程度存在差异。编写程序并解释主要结果。

【参考答案】

```
> x <- 20
> n <- 46
> p <- 0.5
> x2 <- (x-n*p)^2/(n*p*(1-p))
> P <- 1 - pchisq(x2, 1)
> sprintf('X-squared = %f, df = 1, p-value = %f', x2, P)
[1] "X-squared = 0.782609, df = 1, p-value = 0.376344"
```

$p > 0.05$, 所以满足原假设,即在显著性水平 $\alpha = 0.05$ 下,认为这两门课程被学生喜欢的程度不存在差异。