统计学与 R 语言

第 15 讲　点估计与区间估计

张敬信

2022 年 4 月 18 日

哈尔滨商业大学

- 与总体有关的指标是**参数**；与样本有关的指标，是**统计量。**
- 统计推断的重要内容之一就是**参数估计**，即在抽样及抽样分布的基础上，根据样本统计量来推断所关心的总体参数。

一. 点估计与区间估计

参数估计主要有两种：

- **点估计**（准确/不一定可靠）：就是用样本统计量估计。比如估计哈尔滨成年男性的平均身高，样本均值 175cm 就是点估计；有一定把握落在 172~178cm 之间，就是区间估计。
- **区间估计**（更可靠/不很精确）：通常是指估计其 95% 置信区间，即有 95% 的把握认为该区间包含了总体参数，换句话说，如果抽样 100 次，将有 95 次该区间包含了总体参数[1]。

置信区间的越窄反映了参数估计的精确度越高，影响它因素一是置信水平，置信水平越高置信宽度越大；二是样本量，样本量越大置信宽度越小。

[1]不能理解成总体参数以 95% 的概率落在该区间.

将构造置信区间的步骤重复很多次，置信区间包含总体参数真值的次数所占的比例称为**置信水平**，表示为 $(1-\alpha)$. 即

$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = 1-\alpha$$

α 是总体参数未在区间内的比例，最常用的置信水平是 0.05。

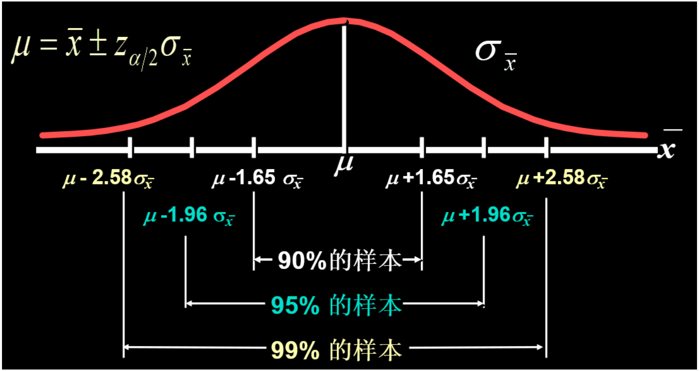

$\mu = \bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}}$
$\sigma_{\bar{x}}$
$\bar{x}$
μ
μ - 2.58$\sigma_{\bar{x}}$
μ -1.65 $\sigma_{\bar{x}}$
μ +1.65$\sigma_{\bar{x}}$
μ +2.58$\sigma_{\bar{x}}$
μ -1.96 $\sigma_{\bar{x}}$
μ +1.96$\sigma_{\bar{x}}$
90%的样本
95% 的样本
99% 的样本

图 1: 区间估计示意图

评价估计量的标准

- **无偏性**：估计量抽样分布的数学期望等于被估计的总体参数。
- **有效性**：对同一总体参数的两个无偏点估计量，有更小标准差的估计量更有效。
- **一致性**：随着样本量的增大，估计量的值越来越接近被估计的总体参数。

1. 用标准误计算置信区间

即使一个代表性非常好的样本，也是无法真正等同于总体的，总会存在一定的**抽样误差。**

比如用 100 人的平均身高作为总体参数 μ 的估计，如果再随机抽样 100 人，又得到另一个平均身高，再 100 人又一个平均身高，……做了 10 次抽样，就可以计算出样本统计量：10 个平均身高和 10 个标准差。这 10 个平均身高也可以计算标准差，这就是标准误（样本统计量的标准差），它反映了样本统计量之间差别（抽样误差）的大小。

然而，实际中不可能多次抽样计算每个样本的统计量，再计算各个统计量之间的差异，而是获取一个尽可能大的样本来计算标准误，理论方法是借助统计学家得到的计算公式[2]

$$se = s/\sqrt{n}$$

其中，s 为样本标准差，n 为样本量。可见样本量越大，标准误越小。

[2]计算具体的标准误时，真正需要的可能是某些真实值或来自总体的值，若无法得到，通常是用它们所对应的样本估计值来代替，某些估计值要保证能作为代替，可能离不开一些模型假定(理论保证).

标准误几乎在所有统计方法中都会出现，因为标准误的大小直接反映了抽样是否有足够的代表性，进而结果是否有足够的可靠性（可信度）。

由于抽样误差的存在，如果用样本统计量直接估计总体参数，则肯定会有一定的偏差。所以在估计总体参数时需要考虑到这种偏差大小，即用置信区间（参数估计值 ± 估计误差）来估计总体参数。

根据中心极限定理，从任何分布中抽样，只要样本量足够大，其统计量最终会服从正态分布。因此，估计误差通常用对应一定正态分布位数的 Z 值再乘以表示抽样误差的标准误来表示。例如，95% 置信区间一般表示为参数估计值 ±1.96× 标准误。

不同样本统计量的标准差的计算过程不同，其标准误也不同。

(1) 均值的置信区间

由于 $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, 故

$$\bar{x} \pm z_{\alpha/2} \times s/\sqrt{n}$$

若样本量较小，建议用相应 t 值代替 z 值。

反过来用上式，将第 2 项记为容许误差 E, 反解 n, 即确定样本量（向上取整）:

$$n = \frac{z_{\alpha/2}^2 \, s^2}{E^2}$$

注： $z_{\alpha/2}$ 表示 $N(0,1)$ 曲线该点右侧的面积，即 $P(Z \geq z_{\alpha/2}) = \alpha/2$;
由对称性 $P(Z \leq -z_{\alpha/2}) = \alpha/2$.

```r
height = c(159,158,164,169,161,161,160,157,158,163,
           161,154,166,168,159)      # 15 个身高数据
mu = mean(height)                    # 点估计：样本均值
mu
#> [1] 161
s = sd(height)
n = length(height)
se = s / sqrt(n)                     # 标准误
## 基于标准误的置信区间
alpha = 0.05
mu + c(-1,1) * qnorm(1-alpha/2) * se
#> [1] 159 163
```

```r
## 估计样本量
E = 0.1                    # 容许误差
N = ceiling(qnorm(1-alpha/2)^2 * s^2 / E)
N
#> [1] 666
```

(2) 比例的置信区间

由于 $\frac{\hat{p}-\pi}{\sqrt{\pi(1-\pi)/n}} \sim N(0,1)$, 故

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

反过来用上式，将第 2 项记为容许误差 E, 反解 n, 即确定样本量（向上取整）:

$$n = \frac{z_{\alpha/2}^2\,\hat{p}(1-\hat{p})}{E^2}$$

例如，随机抽取了 418 名学生，发现有 280 人有不同程度的视力下降，计算该比例的点估计及置信区间

```r
# 点估计
p = 280 / 418
# 置信区间
p + c(-1,1) * qnorm(1-alpha/2) * sqrt(p*(1-p) / 418)
#> [1] 0.625 0.715
# binom.test(280, 418)      # 二项检验
```

(3) 方差的置信区间

总体方差 σ^2 的点估计是 s^2, 且 $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$. 则

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}$$

```
## 对前文身高数据
(n-1) * s^2 / qchisq(c(1-alpha/2, alpha/2), n-1)
#> [1]  9.28 43.06
```

(4) 均值之差的置信区间

由于 $\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \sim N(0,1)$, 故

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

对于配对样本，均值之差的置信区间是基于对应差值计算的：

$$\bar{d} \pm z_{\alpha/2}\frac{\sigma_d}{\sqrt{n}}$$

注： 小样本建议将 z 换成 t, 自由度计算参阅 (贾俊平, 2018)。

(5) 比例之差的置信区间

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

例如，某项研究统计结果如下：

	有效	无效
控制组	32	30
实验组	19	42

```r
n1 = 32 + 30
p1 = 32 / n1
n2 = 19 + 42
p2 = 19 / n2
# 点估计
p1 - p2
#> [1] 0.205
# 置信区间
(p1 - p2) + c(-1, 1) * qnorm(1-alpha/2) *
  sqrt(p1*(1-p1)/n1 + p2*(1-p2)/n2)
#> [1] 0.0344 0.3749
```

(6) 方差比的置信区间

$$\frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1-1,n_2-1)} \le \frac{\sigma_1^2}{\sigma_2^2} \le F_{\alpha/2}(n_2-1,n_1-1)\; s_1^2/s_2^2$$

例如，两种蜘蛛进食的数据：

$$n_1 = 10 \qquad n_2 = 12$$

$$\bar{X} = 10.26\ \text{mm} \qquad \bar{Y} = 9.02\ \text{mm}$$

$$s_X^2 = (2.51)^2 \qquad s_Y^2 = (1.90)^2$$

```r
# 点估计
s2x = 2.51 ^ 2
s2y = 1.90 ^ 2
s2x / s2y
#> [1] 1.75
# 置信区间
n1 = 10
n2 = 12
F1 = qf(1-alpha/2, n1-1, n2-1)
c(1/F1, F1) * (s2x / s2y)
#> [1] 0.486 6.262
```

传统方法依赖于中心极限定理，要求大样本近似正态分布，统计量有计算公式。对于某些抽样分布未知或难以计算的统计量，想要根据一个样本研究抽样样本变化带来的变异，就需要 Bootstrap（自助）重抽样法[3]。

Bootstrap 法的基本思想是：样本是从总体中随机抽取的，则包含总体的全部信息，那么不妨就把该样本视为” 总体”，进行多次有放回抽样生成一系列经验样本，再对每个经验样本计算统计量，就可以得到统计量的分布，进而用于统计推断。

可以证明：**在初始样本量足够大且是从总体中随机抽取的情况下，bootstrap 抽样能够无偏接近总体的分布。**

[3]Bootstrap 法手工实现极其麻烦，但特别适合用计算机实现，已广泛用于统计推断（点估计/置信区间/假设检验）、回归模型诊断以及机器学习等.

以 `Bootstrap` 法估计统计量的置信区间为例，基本步骤如下：

- 从原始样本中有放回地随机抽取 n 个构成子样本
- 对子样本计算想要的统计量
- 重复前两步 K 次，得到 K 个统计量的估计值
- 根据 K 个估计值获得统计量的分布，并计算置信区间

`tidymodels` 系列的 `infer` 包提供了统一的、`tidy` 的统计推断工作流，主要函数有：

- `specify()`: 设定感兴趣的变量或变量关系
- `hypothesize()`: 设定零假设
- `generate()`: 基于零假设生成数据
- `calculate()`: 根据上述数据，计算统计量的分布
- `visualize()`: 可视化

还有获取/绘制 `p` 值/置信区间的函数。

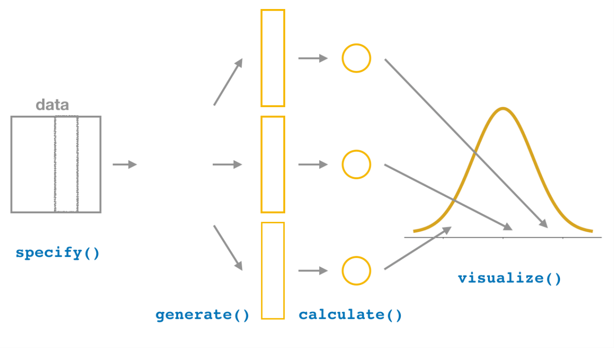


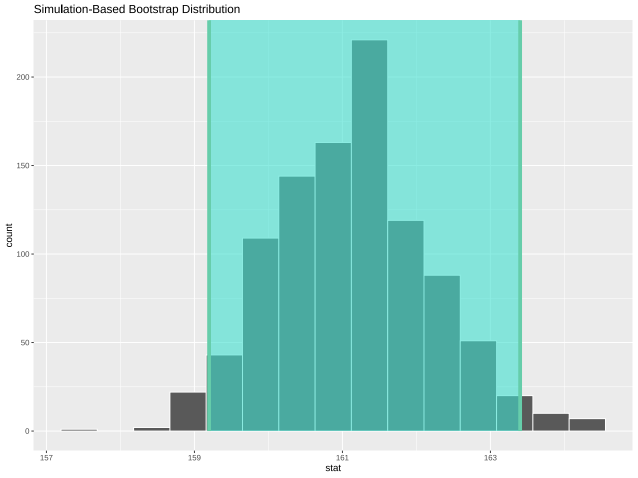
图 2: 用 infer 包实现 Bootstrap 置信区间的一般流程

- 基于 Bootstrap 法计算学生身高的置信区间

```
library(infer)
boot_means = tibble(height) %>%
  specify(response = height) %>%   # 1000 次 bootstrap
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean")         # 计算统计量：样本均值
boot_means
#> Response: height (numeric)
#> # A tibble: 1,000 x 2
#>   replicate  stat
#>       <int> <dbl>
#> 1         1  161.
#> 2         2  160.
#> 3         3  161.
#> # ... with 997 more rows
```

```r
mean(boot_means$stat)                 # 点估计
#> [1] 161
boot_ci = boot_means %>%              # bootstrap 置信区间
  get_ci(level = 0.95, type = "percentile")
boot_ci
#> # A tibble: 1 x 2
#>   lower_ci upper_ci
#>      <dbl>    <dbl>
#> 1     159.     163.
```

```r
visualize(boot_means) +
  shade_ci(endpoints = boot_ci)    # 可视化
```


Simulation-Based Bootstrap Distribution
count
200
150
100
50
0
157
159
161
163
stat

二. 矩估计

设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本，根据大数定律，对 $\forall \varepsilon > 0$, 有

$$\lim_{n\to+\infty} P\{|\bar{X} - E(X)| \geq \varepsilon\} = 0$$

并且对任意 k, 只要 $E(X^k)$ 存在，同样有

$$\lim_{n\to+\infty} P\left\{|\frac{1}{n}\sum_{i=1}^{n} X_i^k - E(X^k)| \geq \varepsilon\right\} = 0, \qquad k = 1, 2, \cdots$$

因此用样本矩估计总体矩，就得到总体分布中参数的一种（无偏）估计。比如，用样本均值估计总体均值。

矩估计的优点是简单易行，不需要事先知道总体是什么分布，缺点是当总体分布已知时，没有充分利用分布提供的信息，且矩估计量不具有唯一性。

k 个未知参数的矩估计一般步骤：

- 令 1 阶样本矩等于 1 阶理论矩：$M_1 = \frac{1}{n}X_i = E(X)$;
- 令 2 阶样本矩等于 2 阶理论矩：$M_2 = \frac{1}{n}X_i^2 = E(X^2)$;
- ……
- 令 k 阶样本矩等于 k 阶理论矩：$M_k = \frac{1}{n}X_i^k = E(X^k)$;
- 求解方程组，得到 k 个未知参数的估计值。

注：以上用的是原点矩，也可以用中心矩列方程：$M_k^* = E\big[(X-\mu)^k\big]$.

例如，X_1, $X_2 m \cdots$, X_n 来自均值为 μ, 方差为 σ^2 的正态总体 X, 用矩估计法估计 μ 和 σ^2.

$$\begin{cases} \frac{1}{n}\sum_{i=1}^{n} X_i = E(X) = \mu \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 = E(X^2) = \sigma^2 + \mu^2 \end{cases}$$

解得

$$\begin{cases} \hat{\mu} = \bar{X} \\ \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \end{cases}$$

```r
## 对前文身高数据为例
mu = mean(height)
mu
#> [1] 161
s2 = mean((height - mu)^2)
s2
#> [1] 16.2
```

本篇主要参阅 (张敬信, 2022), (贾俊平, 2018), (冯国双, 2018), STAT 415 Introduction to Mathematical Statistics, 以及包文档，模板感谢 (黄湘云, 2021), (谢益辉, 2021).

参考文献

冯国双 (2018). 白话统计. 电子工业出版社, 北京, 1 edition.

张敬信 (2022). *R* 语言编程：基于 *tidyverse*. 人民邮电出版社, 北京.

谢益辉 (2021). *rmarkdown: Dynamic Documents for R*.

贾俊平 (2018). 统计学. 中国人民大学出版社, 北京, 7 edition.

黄湘云 (2021). *Github: R-Markdown-Template*.