贝叶斯统计

续本达

清华大学 工程物理系

2023-12-13

复习

自助法 bootstrap

把样本看作总体的估计，用经验分布函数 F_n 估计总体分布函数 F
其它一切统计量都不变。

对应关系

$$\mathrm{E}_F(\hat{\theta}) \to \mathrm{E}_{F_n}(\hat{\theta}) \xrightarrow{\text{Monte Carlo}} \overline{\theta^*}$$

- $\theta \to \hat{\theta}$
  $\hat{\theta}$ 是 F_n 下的“真值”

应用

自助法通过给出统计量的分布估计，可以与点估计、区间估计、假设检验、回归分析、方差分析相结合。

二项总体

引子：橙子

工物系学生节的魔术节目需要橙子道具。你接到了购买橙子的任务，发现某电商有两家店：

- 溯源店出货 200 件，好评率 97%
- 跃迁店出货 20 年，好评率 100%

问题

- 应该从哪家店购买？哪一家店更可能给你良好的购物体验？

几点考虑

1. 好评数越大越好
2. 好评率越高越好
3. 打分的人越多，好评率越可靠
4. 最好能有一种通用规则适用不同人数

Benda 指数

- 满足二项分布的问题，可使用“本达指数”
- 假想有额外两位顾客，分别给店铺打“好评”与“差评”。
- 重新算出的好评率，称为本达指数 B。

$$B = \frac{k+1}{n+2}$$

店铺	顾客数 (n)	好评率	好评数 (k)	本达指数 (B)
溯源	200	97%	194	
跃迁	20	100%	20	

-

Benda 指数

- 满足二项分布的问题，可使用“本达指数”
- 假想有额外两位顾客，分别给店铺打“好评”与“差评”。
- 重新算出的好评率，称为本达指数 B。

$$B = \frac{k+1}{n+2}$$

店铺	顾客数 (n)	好评率	好评数 (k)	本达指数 (B)
溯源	200	97%	194	0.965
跃迁	20	100%	20	0.955

- 溯源店胜出！

Laplace 法则

- 满足二项分布的问题，可以使用 ~~本达指数~~ Laplace 法则
- 假想有额外的两位顾客，分别给店铺打“好评”与“差评”，重新算出好评率，称为 Laplace 指数 L

$$L = \frac{k+1}{n+2}$$

- Laplace 法则来源于哲学问题：我今年 36 岁，我知道每天太阳都升起了，明天太阳升起的概率是多少?
  - 每一天都是一次 0-1 试验，太阳升起“成功”还是“失败”。过年的 36 年里，试验都成功了。
    $k = n = 365.25 \times 36 = 13149$
  - 明天太阳升起的概率是

$$\frac{13149+1}{13149+2} = 99.992\,\%$$

注意：这里的“概率”是贝叶斯概率

Laplace 法则

- 满足二项分布的问题，可以使用 ~~本达指数~~ Laplace 法则
- 假想有额外的两位顾客，分别给店铺打“好评”与“差评”，重新算出好评率，称为 Laplace 指数 L

$$L = \frac{k+1}{n+2}$$

- Laplace 法则来源于哲学问题：我今年 36 岁，我知道每天太阳都升起了，明天太阳升起的概率是多少?
  - 每一天都是一次 0-1 试验，太阳升起“成功”还是“失败”。过年的 36 年里，试验都成功了。
    $k = n = 365.25 \times 36 = 13149$
  - 明天太阳升起的概率是

$$\frac{13149+1}{13149+2} = 99.992\,\%$$

注意：这里的“概率”是贝叶斯概率

频率派

不用 Laplace 法则，如何做？

假设

- 店铺的服务质量不随时间变化，顾客对溯源和跃适打“好评”的概率分别是 p_A 和 p_B，
- 顾客对服务打分时相互独立
- n 位顾客给出的“好评”数 k 服从二项分布

$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- 溯源，200 位顾客，其中 194 位顾客给出好评，假设 $p_A = 0.97$，那么
  - $P(k=194) = \binom{200}{194} 0.97^{194} (1-0.97)^6 = 0.16$
  - $P(k \geq 194) = 0.61$
- 跃迁，20 位顾客，其中 20 位顾客给出好评，假设 $p_B = 0.97$，那么
  - $P(k=20) = \binom{20}{20} 0.97^{20} (1-0.97)^0 = 0.54$
  - $P(k \geq 20) = 0.54$

不用 Laplace 法则，如何做？

假设

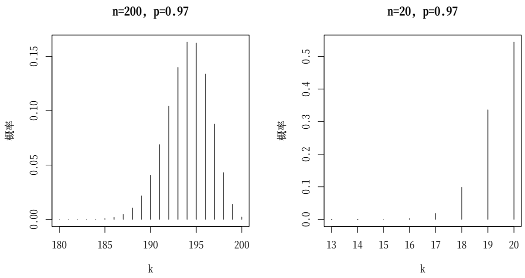
- 店铺的服务质量不随时间变化，顾客对溯源和跃适打“好评”的概率分别是 p_A 和 p_B，
- 顾客对服务打分时相互独立
- n 位顾客给出的“好评”数 k 服从二项分布

$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- 溯源，200 位顾客，其中 194 位顾客给出好评，假设 $p_A = 0.97$，那么
  - $P(k=194) = \binom{200}{194} 0.97^{194} (1-0.97)^6 = 0.16$
  - $P(k \geq 194) = 0.61$
- 跃迁，20 位顾客，其中 20 位顾客给出好评，假设 $p_B = 0.97$，那么
  - $P(k=20) = \binom{20}{20} 0.97^{20} (1-0.97)^0 = 0.54$
  - $P(k \geq 20) = 0.54$

二项分布的直方图

```r
par(mfrow = c(1, 2))
k <- seq(180, 200)
plot(k, dbinom(k, 200, 0.97), type='h', ylab="概率", main="n=200, p=0.97")
k <- seq(13, 20)
plot(k, dbinom(k, 20, 0.97), type='h', ylab="概率", main="n=20, p=0.97")
```



试验次数不同，分布的形状差异明显。综合考虑很难。

频率派的困难

- 比较 p_A 和 p_B 的最大似然估计：
  - $\hat{p}_A = 0.97, \hat{p}_B = 1$，与 Laplace 法则的结论相反
- 比较 p_A 和 p_B 的区间估计，对 $p_A \geq p_B$ 进行假设检验：
  - 二项分布的置信区间不对称，当 n 较小时，难以找到枢轴量，无简洁的构造方法

0-1 分布参数的区间估计

从一个 0-1 总体，获得一个容量为 n 的样本，那么成功次数 k 是二项分布。

- 由中心极限定理， n 较大时，可构造枢轴量：

$$\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} = \frac{n\overline{X} - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

- 根据 z 统计量构造 p 的置信区间。

我们学习的区间估计和假设检验方法围绕正态总体构造。离开正态总体遇到困难。

频率派的困难

- 比较 p_A 和 p_B 的最大似然估计：
  - $\hat{p}_A = 0.97, \hat{p}_B = 1$，与 Laplace 法则的结论相反
- 比较 p_A 和 p_B 的区间估计，对 $p_A \geq p_B$ 进行假设检验：
  - 二项分布的置信区间不对称，当 n 较小时，难以找到枢轴量，无简洁的构造方法

0-1 分布参数的区间估计

从一个 0-1 总体，获得一个容量为 n 的样本，那么成功次数 k 是二项分布。

- 由中心极限定理，n 较大时，可构造枢轴量：

$$\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} = \frac{n\overline{X} - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

- 根据 z 统计量构造 p 的置信区间。

我们学习的区间估计和假设检验方法围绕正态总体构造。离开正态总体遇到困难。

频率派的困难

- 比较 p_A 和 p_B 的最大似然估计：
  - $\hat{p}_A = 0.97, \hat{p}_B = 1$，与 Laplace 法则的结论相反
- 比较 p_A 和 p_B 的区间估计，对 $p_A \geq p_B$ 进行假设检验：
  - 二项分布的置信区间不对称，当 n 较小时，难以找到枢轴量，无简洁的构造方法

0-1 分布参数的区间估计

从一个 0-1 总体，获得一个容量为 n 的样本，那么成功次数 k 是二项分布。

- 由中心极限定理， n 较大时，可构造枢轴量：

$$\frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} = \frac{n\overline{X} - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

- 根据 z 统计量构造 p 的置信区间。

我们学习的区间估计和假设检验方法围绕正态总体构造。离开正态总体遇到困难。

频率派的困难

- 比较 p_A 和 p_B 的最大似然估计：
  - $\hat{p}_A = 0.97, \hat{p}_B = 1$，与 Laplace 法则的结论相反
- 比较 p_A 和 p_B 的区间估计，对 $p_A \geq p_B$ 进行假设检验：
  - 二项分布的置信区间不对称，当 n 较小时，难以找到枢轴量，无简洁的构造方法

0-1 分布参数的区间估计

从一个 0-1 总体，获得一个容量为 n 的样本，那么成功次数 k 是二项分布。

- 由中心极限定理， n 较大时，可构造枢轴量：

$$\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} = \frac{n\overline{X} - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

- 根据 z 统计量构造 p 的置信区间。

我们学习的区间估计和假设检验方法围绕正态总体构造。离开正态总体遇到困难。

使用 bootstrap 解决能否解决?

- 0-1 总体，容量为 200 的样本，194 个 1，6 个 0
  - bootstrap 是从 $B(200, 0.97)$ 采样。
- 容量为 20 的样本，20 个 1。
  - bootstrap 是从 $B(20, 1)$ 采样。
- 可见 bootstrap 方法等价于最大似然估计，仍无法解决非正态分布的问题。

使用 bootstrap 解决能否解决?

- 0-1 总体，容量为 200 的样本，194 个 1，6 个 0
  - bootstrap 是从 $B(200, 0.97)$ 采样。
- 容量为 20 的样本，20 个 1。
  - bootstrap 是从 $B(20, 1)$ 采样。
- 可见 bootstrap 方法等价于最大似然估计，仍无法解决非正态分布的问题。

使用 bootstrap 解决能否解决?

- 0-1 总体，容量为 200 的样本，194 个 1，6 个 0
  - bootstrap 是从 $B(200, 0.97)$ 采样。
- 容量为 20 的样本，20 个 1。
  - bootstrap 是从 $B(20, 1)$ 采样。
- 可见 bootstrap 方法等价于最大似然估计，仍无法解决非正态分布的问题。

贝叶斯方法

明日见日问题

- Laplace 在 18 世纪回答一个哲学问题
  - 明天我见到太阳从东方升起的概率是多少?
- 把太阳升起与否看成 Bernoulli 实验
  - 明日见日的概率是 Bernoulli 实验的成功率 p
  - 如何合理估算 p?
- 关键问题：p 是不是随机变量?
  - 贝叶斯派：是
  - 频率派：不是

明日见日问题

- Laplace 在 18 世纪回答一个哲学问题
  - 明天我见到太阳从东方升起的概率是多少?
- 把太阳升起与否看成 Bernoulli 实验
  - 明日见日的概率是 Bernoulli 实验的成功率 p
  - 如何合理估算 p?
- 关键问题：p 是不是随机变量?
  - 贝叶斯派：是
  - 频率派：不是

明日见日问题

- Laplace 在 18 世纪回答一个哲学问题
  - 明天我见到太阳从东方升起的概率是多少?
- 把太阳升起与否看成 Bernoulli 实验
  - 明日见日的概率是 Bernoulli 实验的成功率 p
  - 如何合理估算 p?
- 关键问题：p 是不是随机变量?
  - 贝叶斯派：是
  - 频率派：不是

贝叶斯概率

- n 次 p 成功率的 Bernoulli 实验的成功次数 k，满足二项分布

$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- 把 p 看成随机变量，按贝叶斯公式，

$$P(p|k,n) = \frac{P(k|p,n)P(p|n)}{P(k|n)}$$

  - $P(p|n)$ 是 p 的 先验分布
  - $P(p|k,n)$ 是 p 的 后验分布
  - $P(k|p,n)$ 是 p 的 似然函数
  - $P(k|n)$ 叫做 evidence

- 当先验分布 $P(p|n) = 1$ 时， p 作为随机变量的后验分布为

$$P(p|k,n) \propto P(k|p,n) \propto p^k(1-p)^{n-k} \sim \mathrm{Beta}(k+1, n-k+1)$$

服从贝塔分布。

贝叶斯概率

- n 次 p 成功率的 Bernoulli 实验的成功次数 k，满足二项分布

$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- 把 p 看成随机变量，按贝叶斯公式，

$$P(p|k,n) = \frac{P(k|p,n)P(p|n)}{P(k|n)}$$

  - $P(p|n)$ 是 p 的 先验分布
  - $P(p|k,n)$ 是 p 的 后验分布
  - $P(k|p,n)$ 是 p 的 似然函数
  - $P(k|n)$ 叫做 evidence

- 当先验分布 $P(p|n) \equiv 1$ 时， p 作为随机变量的后验分布为

$$P(p|k,n) \propto P(k|p,n) \propto p^k (1-p)^{n-k} \sim \mathrm{Beta}(k+1, n-k+1)$$

服从贝塔分布。

贝叶斯概率

- n 次 p 成功率的 Bernoulli 实验的成功次数 k，满足二项分布

$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- 把 p 看成随机变量，按贝叶斯公式，

$$P(p|k,n) = \frac{P(k|p,n)P(p|n)}{P(k|n)}$$

  - $P(p|n)$ 是 p 的 先验分布
  - $P(p|k,n)$ 是 p 的 后验分布
  - $P(k|p,n)$ 是 p 的 似然函数
  - $P(k|n)$ 叫做 evidence

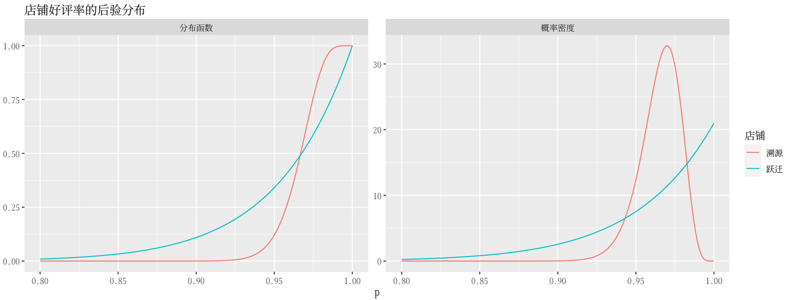
- 当先验分布 $P(p|n) \equiv 1$ 时， p 作为随机变量的后验分布为

$$P(p|k,n) \propto P(k|p,n) \propto p^k (1-p)^{n-k} \sim \mathrm{Beta}(k+1, n-k+1)$$

服从贝塔分布。

后验分布

```
library(ggplot2)
d <- data.frame(店铺=c("跃迁", "溯源"), 顾客数=c(20, 200), 好评率=c(1, 0.97))
d$好评数 <- d$顾客数 * d$好评率
post <- data.frame()
x <- seq(0.8, 1, 0.001)
for (i in 1:2) {
  post <- rbind(post, rbind(data.frame(p=x, 店铺=d$店铺[i], 值="概率密度",
                               value=dbeta(x, d$好评数[i] + 1, d$顾客数[i] - d$好评数[i] + 1)),
                data.frame(p=x, 店铺=d$店铺[i], 值="分布函数",
                           value=pbeta(x, d$好评数[i] + 1, d$顾客数[i] - d$好评数[i] + 1))))}
p <- ggplot(post, aes(x=p, y=value, color=店铺)) + geom_line() + ggtitle("店铺好评率的后验分布")
print(p + ylab("") + facet_wrap(~值, scales="free_y"))
```



点估计：后验期望估计量

取后验分布的期望，是贝叶斯参数的一样点估计方法

$$p \sim \mathrm{Beta}(a, b), p \in [0, 1]$$

有 $\mathrm{E}(p) = \dfrac{a}{a+b} = \dfrac{k+1}{n+2}$ ，即 Laplace 指数。

- 溯源店 p_A

$$\hat{p}_A = \mathrm{E}(p_A) = 0.965$$

- 跃迁店 p_B

$$\hat{p}_B = \mathrm{E}(p_B) = 0.955$$

后验期望估计是 Laplace 法则的数学依据。

点估计：后验期望估计量

取后验分布的期望，是贝叶斯参数的一样点估计方法

$$p \sim \mathrm{Beta}(a,b), p \in [0,1]$$

有 $\mathrm{E}(p) = \dfrac{a}{a+b} = \dfrac{k+1}{n+2}$ ，即 Laplace 指数。

- 溯源店 p_A
$$\hat{p}_A = \mathrm{E}(p_A) = 0.965$$
- 跃迁店 p_B
$$\hat{p}_B = \mathrm{E}(p_B) = 0.955$$

后验期望估计是 Laplace 法则的数学依据。

区间估计：后验分布的区间

- 贝叶斯的区间估计里，直接使用参数的后验分布，不必构造枢轴量
  1. ~~选取枢轴量~~
  2. 由分位点定义建立不等式
  3. ~~解出不等式~~

```
qbeta(1-0.9, 195, 7)
qbeta(1-0.9, 21, 1)
```

```
[1] 0.9481917
[1] 0.8961505
```

- 溯源店 p_A 的 90% 置信区间为 [0.95, 1]
- 跃迁店 p_B 的 90% 置信区间为 [0.90, 1]

模型选择

蒙特卡罗法从后验分布采样

- 大多数情况下，参数 θ 的后验分布无法解析表达，

$$P(\theta|D,H) = \frac{P(D|\theta,H)P(\theta|H)}{P(D|H)}$$

其中 H 代表理论假设，D 代表数据。

- $P(D|H)$ 是最难计算的

$$P(D|H) = \int P(D|\theta,H)P(\theta|H)\mathrm{d}\theta$$

  - 它与 θ 无关，通过从 $P(\theta|D,H) \propto P(D|\theta,H)P(\theta|H)$ 采样取得一组 $\{\theta_i\}$，使用马尔可夫链法进行
  - 可构造 后验期望估计量

$$\mathrm{E}(\theta) = \overline{\theta}$$

  - 可用 $\{\theta_i\}$ 的样本分位数构造 θ 后验分布的区间

多层级模型与假设检验

- 考虑贝叶斯公式的先验分布 $P(\theta|H)$，如果 H 本身也可以两种可能 H_0, H_1。
- 同样可以通过数据进行假设检验，也叫作 模型选择

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{P(D|H_0)P(H_0) + P(D|H_1)P(H_1)}$$

- $P(D|H_i)$ 的积分较难计算，可使用马尔可夫链蒙特卡罗法完成。

总结

两大统计学派

贝叶斯学派	频率学派
“好评率”是随机变量	“好评率”是常数
概率依赖人的认知	概率是假想重复实验的平均
需要先验知识	不需要先验知识
计算复杂	计算相对简单
无需正态假设	大多依赖正态假设
18 世纪出现	20 世纪由概率论公理化成为主流
21 世纪被广泛使用	

21 世纪的实验物理

- 学术界从频率派死忠向“哪个好用就用哪个”的骑墙派转变

Forecasting neutrino masses from combining KATRIN and the CMB observations: Frequentist and Bayesian analyses

Ole Host, Ofer Lahav, Filipe B. Abdalla, and Klaus Eitel
Phys. Rev. D **76**, 113005 – Published 27 December 2007

- 在偏离正态总体时，两种方法往往给出不同结果
  - 例如，中微子质量不能为负，参数空间有界，正态假设不成立。
- 有学者开始同时使用两种方法，由读者自己选择。