统计推断的基础

目录

1 常用统计术语的定义

由于样本来自总体，必然带有总体的信息，而统计推断是根据样本数据对总体性质进行推断的科学。统计推断包括，参数估计、假设检验和预测。

- 总体：将想要了解的统计对象，可以是有限的或者无限的。
- 样本：总体的一部分，可以是有代表性的或无代表性的，日常所使用的数据通常是样本。
- 参数：想了解的总体性质
- 估计量：统计估计，是数据的函数
- 估计值：将估计量应用于数据后得到的数值
- 样本容量：样本中所包含的个体数目
- 抽样：令 Y 为一个随机变量，代表着概率密度函数为 $f(y;\theta)$ 的一个总体，其中 $f(y;\theta)$ 依赖于单个参数 θ 。假定除了 θ 值未知外，Y 的概率密度函数 (pdf) 是已知的。不同的 θ 值将意味着不同的概率分布，因 此我们对 θ 值感兴趣。如果能得到该总体的某种样本，我们就能了解 θ 的某些情况。最容 易处理的抽样方案是随机抽样。

- 随机抽样: 若 $Y_1, Y_2, \cdots, Y_n$ 是具有同一概率密度函数 $f(y;\theta)$ 的独立随机变量, 我们称 $\{Y_1, Y_2, \cdots, Y_n\}$ 为一个来自 $f(y;\theta)$ 的随机样本 (random sample) [或者说 来自由 $f(y;\theta)$ 所代表的总体的一个随机样本]。
- 抽样方差：估计量的方差与一个抽样分布相联系, 所以通常称为抽样方差 (sampling variance)。要记住, 抽样方差不是一个随机变量, 而是一个常数。
- 标准差：样本中各个个体与其平均数的差的平方的算术平均数的平方根，反映的是一个数据集的离散程度，值越大，越离散，即个体间差异越大。
- 标准误差：反映样本平均数对总体平均数的变异程度，表示的是抽样的误差，代表的就是样本平均数与总体均数的相对误差，简单理解即是对平均数求标准差，比如一次实验会得到一个平均数，多次实验得到多个平均数，标准误差即是对这些平均数求标准差。

> 标准差是针对一次抽样的原始数据而言的；而标准误是针对多次抽样的样本均数而言的（也可以是其他统计量）。

2 估计量的有限样本性质

2.1 估计量的含义

- 给定一个随机样本 $\{Y_1, Y_2, \cdots, Y_n\}$, 它来自一个取决于某末知参数 θ 的总体分布, θ 的一个估计量 (estimator) 就是 赋予样本每个可能结果一个 θ 值的法则。这个法则在进行抽样之前就已经确立; 具体而言, 无论实际得到什么样的数据, 这个法则都不会改变。
- 作为估计量的一个例子, 令 $\{Y_1, \cdots, Y_n\}$ 为取自均值为 μ 的总体的一个随机样本。 μ 的一个很自然的估计量, 就是这个随机样本的均值:

$$\bar{Y}=n^{-1} \sum_{i=1}^{n} Y_i$$

- 我们把 $\bar{Y}$ 叫做样本均值 (sample average), 这里 $\bar{Y}$ 看作一个估计量。给定随机变量 $Y_1, \cdots, Y_n$ 的任何一种结果, 我们都用同样的法则去估计 μ, 即取其平均。对于实际结果 $\{y_1, \cdots, y_n\}$, 估计值 (estimate) 就是该样本的均值: $\bar{y}=\left(y_1+y_2+\cdots+y_n\right)/n$。

参数 θ 的一个估计量 W 可表示为一个抽象的数学公式: $W=h\left(Y_1, Y_2, \cdots, Y_n\right)$ h 代表随机变量 $Y_1, Y_2, \cdots, Y_n$ 的某个已知函数。如同样本均值的特殊情形那样, W 也 因取决于随机样本而成为一个随机变量: W 随着我们从总体中抽到不同的随机样本而可能改变。当我们把一个特定的数集 [比如 $\{y_1, y_2, \cdots, y_n\}$] 代人函数 h 中时, 便得到 θ 的 一个估计值, 记为 $w=h\left(y_1, y_2, \cdots, y_n\right)$ 。有时把 W 叫做点估计量, 而把 w 叫做点估计值。为了评价不同的估计方法, 我们研究随机变量 W 的概率分布的各种性质。一个估计量的分布常被称为抽样分布 (sampling distribution), 因为这个分布描述了 W 在不同随机样本上取各种结果的可能性。

2.2 估计量的性质

由于有无限种组合数据以估计参数的法则，所以我们需要一些有意义的准则来挑选估计量，或者至少能淘汰一些估计量。下面，我们介绍一些判断估计量好坏的性质。

2.2.1 无偏性

原则上, 给定 Y_i 的概率分布和函数 h, 我们就能求出 W 的整个抽样分布。通常, 在 评价 W 作为 θ 的一个估计量时, 集中考虑 W 分布的少数几个特征比较容易。一个估计量 的第一个重要性质就是关于它的期望值。

- 无偏估计量: 若 θ 的估计量 W 对一切可能的 θ 值, 都有

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

  则 W 是一个无偏估计量 (unbiased estimator)。

  > 一个估计量若是无偏的，则其概率分布的期望值就等于它所估计的参数。无偏性并不是说随便基于一个特定样本得到的估计值等于 θ，而是说如果我们能够从总体中抽取关于 Y 的无限多个样本，并且每次抽出一个样本都计算得到一个估计值，那么将所有随机样本的这些估计值平均起来，边可以得到 θ。由于大多数应用中，我们只能使用一个随机样本，所以这是个思想实验。

- 一个估计量的偏误: 如果 W 是 θ 的一个偏误估计量 (biased estimator), 则它的偏误 (bias) 可定义为

  $$\operatorname{Bias}(W) \equiv \mathrm{E}(W)-\theta$$

可以证明, 有些估计量在很一般的情形下是无偏的。现在我们来证明, 样本均值 $\bar{Y}$ 是 总体均值 μ 的一个无偏估计量, 不管其背后的总体如何分布。

$$\begin{aligned}
\mathrm{E}(\bar{Y}) &= \mathrm{E}\left((1/n)\sum_{i=1}^{n} Y_i\right)=(1/n)\mathrm{E}\left(\sum_{i=1}^{n} Y_i\right)=(1/n)\left(\sum_{i=1}^{n}\mathrm{E}\left(Y_i\right)\right) \\
&=(1/n)\left(\sum_{i=1}^{n}\mu\right)=(1/n)(n\mu)=\mu
\end{aligned}$$

为了做假设检验, 我们还有必要从均值为 μ 的总体中估计方差 σ^2 。令随机样本 $\{Y_1, Y_2,\cdots,Y_n\}$ 取自 $\mathrm{E}(Y)=\mu$ 和 $\operatorname{Var}(Y)=\sigma^2$ 的总体, 定义估计量为

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2$$

它常被称为样本方差 (sample variance)。可以证明, S^2 是 σ^2 的无偏估计量; $\mathrm{E}\left(S^2\right)=\sigma^2$ 。 用 $n-1$ 而不用 n 作除数, 是因为均值 μ 的使用是估计而非已知的。若 μ 已知, 则 σ^2 的一 个无偏估计量将是 $n^{-1}\sum_{i=1}^{n}\left(Y_i-\mu\right)^2$ ，然而实践中很少知道 μ。

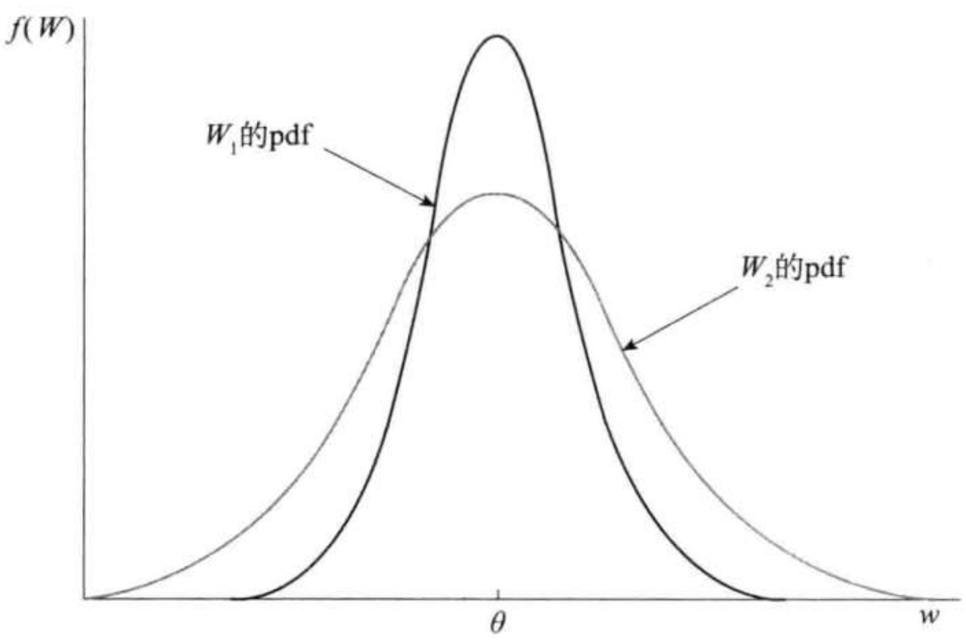
- 无偏性的一个缺点： 一些合理甚至相当好的估计量却不是无偏的。

- 无偏性的另一重要缺点：实际上存在着相当糟糕的无偏估计量。

> 考虑我们在估计一 个总体的均值 μ 。假定我们不用样本均值 $\bar{Y}$ 去估计 μ, 而是在收集一个容量为 n 的样本后, 只保留第一个观测, 并抛弃所有其余观测, 然后用 $W \equiv Y_1$ 作为 μ 的估计量, 因为 $\mathrm{E}\left(Y_1\right)=\mu$, 所以这个估计量也是无偏的。很可能你已经意识到, 忽略除第一个以外的所有观测值不是一个明智的估计方法: 它把样本中的大部分信息都丢弃了。例如, 取 $n=100$, 我们有 随机变量 Y 的 100 个观测结果, 但是我们只使用了其中的第一个去估计 $\mathrm{E}(Y)$ 。

2.2.2 估计量的抽样方差

无偏性仅保证估计量的概率分布有一个等于它所估计参数的均值，但我们仍然需要了解这个估计量的分布有多分散。一个估计量可以在平均的意义上等于 θ,但他仍然会以很大的概率偏离到很远处。下图表明，W_1 和 W_2 都是 θ 的无偏估计量，但 W_1 的分布更加紧密地集中在 θ 周围。用 W_1 作为估计量意味着，我们得到的一个具体估计值非常远离 θ 的随机样本，可能性相对较小。



将以上思想，用一个估计量的方差 (或标准差)来表达。估计量的方差与一个抽样分布相联系, 所以通常称为抽样方差 (sampling variance)。要记住, 抽样方差不是一个随机变量, 而是一个常数。 现在, 我们来求用以估计总体均值 μ 的样本均值的方差:

$$
\begin{aligned}
\operatorname{Var}(\bar{Y}) & =\operatorname{Var}\left((1 / n) \sum_{i=1}^n Y_i\right)=\left(1 / n^2\right) \operatorname{Var}\left(\sum_{i=1}^n Y_i\right)=\left(1 / n^2\right)\left(\sum_{i=1}^n \operatorname{Var}\left(Y_i\right)\right) \\
& =\left(1 / n^2\right)\left(\sum_{i=1}^n \sigma^2\right)=\left(1 / n^2\right)\left(n \sigma^2\right)=\sigma^2 / n
\end{aligned}
$$

- 若 $\{Y_i: i=1,2, \cdots, n\}$ 是取自均值为 μ 和方差为 σ^2 的总体的一个随机样本, 则 $\bar{Y}$ 有和总体一样的均值, 但它的方差等于总体方差 σ^2 除以样本容量 n 。

- $\operatorname{Var}(\bar{Y})=\sigma^2/n$ 的一个重要含义是, 增大样本容量可以使它非常接近于零,这是一个合理估计量的关键特征。
- 在无偏估计量中, 我们偏好有最小方差的估计量。这就使我们能淘汰某些估计量。

> 对一个来自均值为 μ 和方差为 σ^2 的总体的随机样本, 我们知道 $\bar{Y}$ 是无偏的, 随机抽取, 就应有 $\operatorname{Var}\left(Y_1\right)=\sigma^2$ 。因此, 即便样本较小, $\operatorname{Var}\left(Y_1\right)$ 和 $\operatorname{Var}(\bar{Y})$ 之间的差异 也可能很大。如果 $n=10$, 那么 $\operatorname{Var}\left(Y_1\right)$ 便是 $\operatorname{Var}(\bar{Y})=\sigma^2/10$ 的 10 倍。这就给了我们一 个规范方法, 排除 Y_1 作为 μ 的估计量。

2.2.3 有效性

相对有效性: 假如 W_1 和 W_2 是 θ 的两个无偏估计量, 那么, 如果对所有的 θ 都有 $\operatorname{Var}\left(W_1\right) \leqslant \operatorname{Var}\left(W_2\right)$ 且至少对一个 θ 值不等式严格成立, 则称 W_1 比 W_2 更有效。前面我们证明了, 为估计总体均值 μ, 只要 $n>1$, 对任何 σ^2 值都有 $\operatorname{Var}(\bar{Y})<\operatorname{Var}\left(Y_1\right)$ 。但我们不能一味根据最小方差准则在无偏估计量之中进行选择。给定 θ 的两个无偏估计量, 一个可能对某些 θ 值有较小方差, 而另一个则对另一些 θ 值有较小方差。

比较不一定无偏估计量的一个方法是计算估计量的均方误 (mean squared error, MSE) 。如果 W 是 θ 的一个估计量, 则它的 MSE 被定义为 $\operatorname{MSE}(W)=\mathrm{E}\left[(W-\theta)^2\right]$ 。 MSE 度量了估计量 W 与 θ 的平均距离。可以证明, $\operatorname{MSE}(W)=\operatorname{Var}(W)+[\operatorname{Bias}(W)]^2$ 。 因此 $\operatorname{MSE}(W)$ 取决于方差和偏误（如果有偏误的话)。这样一来, 对于任何两个估计量, 即使其中之一或两者都有偏误, 我们也能加以比较。

3 估计量的渐进或大样本性质

前面我们可以看到，作为总体均值 μ 的估计量，Y_1 尽管无偏, 却是一个糟糕的估计量(方差可能比样本均值的方差大得多)。 Y_1 的一个明显特征是, 对任何样本容量, 它都具有同样的方差。为了估计总体均值 μ, Y 的方差随着 n 逐渐变大而递减, 在这个意义上, 它变得越来越好; 而Y_1 就没有这种变化。

- 研究估计量的渐近 (asymptotic) 或大样本 (large sample) 性质可以避免一些可笑的估计量。
- 渐近分析涉及如何近似一个估计量的抽样分布特征。这些近似取决于样本容量。不幸的是, 我们无法说出要使渐近分析适当需要多大的样本容量, 因为这取决于潜在的总体分布。

3.1 一致性

估计量的第一个渐进性质告诉我们，随着样本容量的无限增加，用以估计某参数的估计量距离该参数有多远。

- **一致性：令 W_n 为 θ 基于容量为 n 的一个样本 $Y_1, Y_2, \cdots, Y_n$ 的一个估计量。那么, 若随着 $n \rightarrow \infty$, 对任一 $\varepsilon>0$, 都有**

$$\mathrm{P}\left(\left|W_n-\theta\right| \geqslant \varepsilon\right) \rightarrow 0$$

W_n 便是 θ 的一个一致估计量 (consistent estimator)。若 W_n 不是 θ 的一致的估计量, 则说 它是非一致的（inconsistent）。

> 它意味着 W_n 的分布越来越集中于 θ, 粗略地讲, 对 于越来越大的样本容量, W_n 离开 θ 很远的可能性越来越小。

- 若 W_n 是一致的, 我们也说 θ 是 W_n 的概率极限（probability limit), 记作 $\operatorname{plim}\left(W_n\right)=\theta_0$

3.1.1 无偏性和一致性的区别

- 样本容量：无偏性估计量强调的是给定的、有限的样本容量下的一个特征；一致性是强调估计量的抽样分别在样本容量变大时候的一个特征。
- 无偏估计量不一定是一致的,只有那些随着样本容量增大而方差缩减至零的无偏估计量才是一致的。换言之，若 W_n 是 θ 的无偏估计量, 且 随着 $n \rightarrow \infty$ 而有 $\operatorname{Var}\left(W_n\right) \rightarrow 0$, 则 $\operatorname{plim}\left(W_n\right)=\theta$ 。

> **利用全部数据样本的无偏估计量, 通常其方差都随着样本容量的扩大而缩减至零, 因而是一致的。**
> 一致估计量的一个很好的例子是, 取自均值为 μ 和方差为 σ^2 的总体的一个随机样本的均值。我们已经证明该样本均值对 μ 是无偏的。在方程 (C. 6) 中, 对于任何一个样本容量 n, 我们都推导出 $\operatorname{Var}\left(Y_n\right)=\sigma^2 / n$ 。因此, 随着 $n \rightarrow \infty$, 有 $\operatorname{Var}\left(\bar{Y}_n\right) \rightarrow 0$, 所以, Y_n 除了 无偏外, 还是 μ 的一致估计量。

3.1.2 大数定律

$\bar{Y}_n$ 是 μ 的一致估计量, 即使不存在 $\operatorname{Var}\left(\bar{Y}_n\right)$, 这一结论也是成立的。这个经典结论被 称为大数定律 (law of large numbers, LLS)。

大数定律: 令 $Y_1, Y_2, \cdots, Y_n$ 是均值为 μ 的独立同分布随机变量, 于是 $\operatorname{plim}\left(\bar{Y}_n\right)=\mu$

> 大数定律意味着, 我们若对估计总体均值 μ 感兴趣, 通过选取一个足够大的样本, 便能得到一个任意接近 μ 的数。这个基本结论和概率极限的基本性质相结合, 便可以证明一 些相当复杂的估计量是一致的。

性质 PLIM. 1

- 令 θ 为某参数并定义一个新参数 $\gamma=g(\theta), g(\bullet)$ 代表某连续函数。假设 $\operatorname{plim}\left(W_n\right)=\theta$, 定义 γ 的一个估计量为 $G_n=g\left(W_n\right)$ 。于是,$\operatorname{plim}\left(G_n\right)=\gamma$
- 这个结论常常可表述为: 对于一个连续函数 $g(\theta)$, 我们有$\operatorname{plim} g\left(W_n\right)=g\left(\operatorname{plim} W_n\right)$

作为一个一致但偏误估计量的重要例子, 考虑从均值为 μ 和方差为 σ^2 的总体中估计标准差 σ 。

- 样本方差 $S_n^2=\frac{1}{n-1} \sum_{i=1}^n\left(Y_i-\bar{Y}\right)^2$ 是方差 σ^2 的无偏估计。利用大数定律和一些数学知识, 可以证明 S_n^2 也是 σ^2 的一致估计。

- $\sigma=\sqrt{\sigma^2}$ 的一个自然估计量为 $S_n=\sqrt{S_n^2}$ (其中, 平方根总取其正值)。因为平方根的期望值不是期望值的平方根, 所以称为样本标准差 (sample standard deviation) 的 S_n 不是一个无偏的估计量。然 而, 由性质 PLIM. 1, plimS $S_n=\sqrt{\operatorname{plim} S_n^2}=\sqrt{\sigma^2}=\sigma$, 因此, S_n 是 σ 的一个一致估计量。

3.2 渐进正态性

一致性是点估计量的一个性质。虽然它告诉我们, 随着样本容量变大, 这个估计量的分布将围绕所估计的参数而集中到一点上, 但它根本没有告诉我们这个分布在给定样本容量时的形状。为了构造区间估计量和检验假设, 我们需要有一种方法来近似我们的估计量分布。大多数计量经济估计量的分布, 在大样本条件下, 都可用正态分布很好地近似。

- **渐近正态性: 令 $\left\{Z_n: n=1,2, \cdots\right\}$ 是一个随机变量序列, 使得对所有的数 z, 都有随着 $n \rightarrow \infty, P\left(Z_n \leqslant z\right) \rightarrow \Phi(z)$**

其中, $\Phi(z)$ 是标准正态累积分布函数。于是便称 Z_n 服从一个渐近标准正态分布。此时, 我们常记作 ${Z_n}^a \sim \operatorname{Normal}(0,1)$ 。波浪号上方的 “a” 表示 “渐近”或 “近似”的意思。

> 随着样本容量 n 变大, Z_n 的累积分布函数变得越来越接近于标准 正态分布的 cdf。当渐近正态性（asymptotic normality）成立时, 对于较大的 n, 我们近 似有 $\mathrm{P}\left(Z_n \leqslant z\right) \approx \Phi(z)$ 。因此, 可以用标准正态概率去近似关于 Z_n 的概率。

3.2.1 中心极限定理

中心极限定理 (central limit theorem, CLT) 是概率与统计学中最强有力的结论之 一。它表明任何 (具有有限方差的) 总体的一个随机样本的均值, 经过标准化后, 都服从一个渐近标准正态分布。

中心极限定理: 令 $\left\{Y_1, Y_2, \cdots, Y_n\right\}$ 为一个有均值 μ 和方差 σ^2 的随机样本。于是,

$$
Z_n=\frac{\bar{Y}_n-\mu}{\sigma / \sqrt{n}}
$$

服从一个渐近标准正态分布。

- 变量 Z_n 就是 $\overline{Y_n}$ 的标准化形式: 我们从 $\overline{Y_n}$ 减去了 $\mathrm{E}\left(\overline{Y_n}\right)=\mu$, 然后除以 $\operatorname{sd}\left(\overline{Y_n}\right)=\sigma / \sqrt{n}$ 。

  > 这样, 无论 Y 的总体分布是什么, Z_n 都有和标准正态分布一样的零均值和单位方差。令人惊奇的是, 随着 n 变大, Z_n 的整个分布便任意接近于标准正态分布。

- 我们可以把方程中的标准化变量写成 $\sqrt{n}\left(\bar{Y}_n-\mu\right) / \sigma$, 它向我们表明, 为了得到有用的极限分布, 我们必须将样本均值与总体均值之差乘以样本容量的平方根。

> 若不乘以 $\sqrt{n}$, 我们只能得到依概率收敛于零的 $\left(\bar{Y}_n-\mu\right) / \sigma$ 。换言之, 随着 $n \rightarrow \infty$, $\left(\bar{Y}_n-\mu\right) / \sigma$ 的 分布仅向一个点集中, 我们知道, 对于合理的样本容量, 这不可能是一个很好的近似。乘以 $\sqrt{N}$ 便确保了 Z_n 的方差保持不变。

4. 区间估计和置信区间

4.1 区间估计的性质

从一个特定样本得到的点估计值，本身还不足以对检验经济理论或进行政策探讨提供足够信息 个点估计值也许是研究者对总体值的最好猜测，但根据其性质，并不能告诉我们估计值到底离总体参数有多接近。举个例子，假定某研究者根据工人的一个随机样本报告说，培训使小时工资增加 6.4% 我们怎能知道一旦整个工人总体都接受了培训，其效果是否会接近这个数，由于我们不知道总体值，所以无法知道某一特定估计值究竟离它有多近。然而，我们能作出概率意义下的陈述，并由此诞生区间估计。

评价一个估计量的不确定性的一个方法: 求出它的抽样标准差。将点估计值连同估计量的标准差一起报告, 提供了估计值的某些精确度信息。然而, 这个标准差 依赖于未知的总体参数, 就算这个问题可以忽略, 连同点估计值一起报告标准差, 也并没 有直接陈述总体值很可能落在相对于估计值的什么地方。而通过构造置信区间 (confidence interval), 能克服这一局限。

- 我们用一个例子来解释置信区间的概念。假定总体服从 $\mathrm{Normal}(\mu, 1)$ 分布, 并令 $\{Y_1, \cdots, Y_n\}$ 是来自这个总体的一个随机样本。(为便于说明, 我们假定总体方差已知并等于 1 ; 然后我们再说明怎样对付方差未知的现实情形。)
- 样本均值 $\bar{Y}$ 服从一个均值为 μ 和方差为 $1/n$ 的正态分布: $\bar{Y} \sim \mathrm{Normal}(\mu, 1/n)$ 。现在将 $\bar{Y}$ 标准化, 因为标准化的 $\bar{Y}$ 服 从一个标准正态分布, 所以我们有

$$\mathrm{P}\left(-1.96<\frac{\bar{Y}-\mu}{1/\sqrt{n}}<1.96\right)=0.95$$

  括号中的事件等同于事件 $\bar{Y}-1.96/\sqrt{n}<\mu<\bar{Y}+1.96/\sqrt{n}$, 所以

$$\mathrm{P}(\bar{Y}-1.96/\sqrt{n}<\mu<\bar{Y}+1.96/\sqrt{n})=0.95$$

- 随机区间 $[\bar{Y}-1.96/\sqrt{n}, \bar{Y}+1.96/\sqrt{n}]$ 包含总体均值 μ 的 概率是 0.95 或 95%, 所以它有些意义。这一信息使我们能构造 μ 的区间估计值, 方法是 通过把均值的样本结果 $\bar{y}$ 代人而得到。因此,

$$[\bar{y}-1.96/\sqrt{n}, \bar{y}+1.96/\sqrt{n}]$$

给出了 μ 的一个区间估计例子。它又被称为一个 95% 置信区间, 这个区间可简记为 $\bar{y} \pm 1.96/\sqrt{n}$ 。

> 一旦观测到样本数据 $\{y_1, y_2, \cdots, y_n\}$, 方程中的置信区间便容易计算; $\bar{y}$ 是依赖于数据的唯一因素。例如, 假定 $n=16$, 并且 16 个数据点的均值是 7.3, 则 μ 的这 个 95% 置信区间是 $7.3 \pm 1.96/\sqrt{16}=7.3 \pm 0.49$, 即 $[6.81, 7.79]$ 。根据构造可知, $\bar{y}=7.3$ 是此区间的中心。

- 置信区间虽容易计算, 但较难理解。当我们说方程 $[\bar{y}-1.96/\sqrt{n}, \bar{y}+1.96/\sqrt{n}]$ 是 μ 的一个 95%置信区间时, 我们的意思是说, 随机区间

$$[\bar{Y}-1.96/\sqrt{n}, \bar{Y}+1.96/\sqrt{n}]$$

包含 μ 的概率是 0.95 。换言之, 在抽取随机样本之前,$[\bar{Y}-1.96/\sqrt{n}, \bar{Y}+1.96/\sqrt{n}]$ 便有 95% 的机会包含 μ 。 $[\bar{Y}-1.96/\sqrt{n}, \bar{Y}+1.96/\sqrt{n}]$是一个随机区间, 因为端点随不同的样本而变。

> 人们常常这样解释置信区间: “ μ 落在区间$[\bar{y}-1.96/\sqrt{n}, \bar{y}+1.96/\sqrt{n}]$上的概率是 0.95 ”。这种解释是错误的。一旦样本被观测到, 且 $\bar{y}$ 也已计算出来, 则置信区间的上下限只是两个数字 (在上述例子中就是 6.81 和 7.79) 而已。总体参数 μ 虽然未知, 也仅是个常数。因此 μ 或者落入或者不落入区间 (我们永远也不会确切知道是哪一种情形)。一旦利用现有的数 据把置信区间计算出来, 就不再有概率的问题。概率解释来自如下事实: 利用所有随机样本构造这样的置信区间, 其中有 95% 将包含 μ 。

- 为了强调置信区间的意义, 从 Normal $(2,1)$ 分布中取出的 20 个容量 为 $n=10$ 的随机样本 (或复制) 的计算结果, 对 20 个样本中的每一个, 都求出 $\bar{y}$, 并把 (C. 18) 计算为 $\bar{y} \pm 1.96/\sqrt{10}=\bar{y} \pm 0.62$ (保留两位小数)。你可以看到, 每个随机样本的 区间都有所变化。20 个区间中的 19 个包含了总体值 μ 。只有第 19 次复制中的 μ 没有落在 置信区间内。换句话说, 有 95%的样本给出了包含 μ 的置信区间。对于只有 20 次复制的 情形, 这个结果不是必然的, 但在这个特定的模拟实验中却得到了这一结果。

复制	$\bar{y}$	95%区间	包含 μ 吗?
1	1.98	(1.36, 2.60)	是
2	1.43	(0.81, 2.05)	是
3	1.65	(1.03, 2.27)	是
4	1.88	(1.26, 2.50)	是
5	2.34	(1.72, 2.96)	是
6	2.58	(1.96, 3.20)	是
7	1.58	(0.96, 2.20)	是
8	2.23	(1.61, 2.85)	是
9	1.96	(1.34, 2.58)	是
10	2.11	(1.49, 2.73)	是
11	2.15	(1.53, 2.77)	是
12	1.93	(1.31, 2.55)	是
13	2.02	(1.40, 2.64)	是
14	2.10	(1.48, 2.72)	是
15	2.18	(1.56, 2.80)	是
16	2.10	(1.48, 2.72)	是
17	1.94	(1.32, 2.56)	是
18	2.21	(1.59, 2.83)	是
19	1.16	(0.54, 1.78)	否
20	1.75	(1.13, 2.37)	是

4.2 正态分布总体均值的置信区间

由于实践中，我们并不知道 σ,因而区间$[\bar{y}-1.96\sigma/\sqrt{n}, \bar{y}+1.96\sigma/\sqrt{n}]$ 依然未知。为了考虑 σ 未知的情形, 我们必须用一个估计值:

$$S=\left(\frac{1}{n-1} \sum_{i=1}^{n}\left(y_{i}-\bar{y}\right)^{2}\right)^{1/2}$$

- 表示样本标准差。于是, 用 σ 的估计值 s 代替 σ, 我们就能求出一个完全 赖于观测数据的置信区间。不幸的是, 由于 s 取决于特定样本, 这就不能保持 95% 的置信水平。换言之, 因为常数 σ 已被随机变量 s 代替, 所以随机区间 $[\bar{Y} \pm 1.96(S/\sqrt{n})]$ 包含 μ 的概率不再是 0.95 。

我们该怎样做下去呢? 我们必须依靠 t 分布, 而不是使用标准正态分布。 t 分布得自 $\frac{Y-\mu}{S/\sqrt{n}} \sim t_{n-1}$

- 其中, $\bar{Y}$ 是样本均值, S 是随机样本 $\{Y_1, \cdots, Y_n\}$ 的样本标准差。

为了构造一个 95% 的置信区间, 令 c 表示 t_{n-1} 分布中的第 97.5 分位数。换言之, c 是 这样一个值, 使得 t_{n-1} 中的面积有 95% 落在 $-c$ 与 c 之间: $\mathrm{P}\left(-c<t_{n-1}<c\right)=0.95$ 。 (c 值 依赖于自由度 $n-1$, 但我们没有标明。) 一旦适当选定了 c, 随 机区间 $[Y-c \cdot S/\sqrt{n}, Y+c \cdot S/\sqrt{n}]$ 包含 μ 的概率就是 0.95 。对于一个特定的样本, 这 个 95% 置信区间将计算为

$$[\bar{y}-c \cdot s/\sqrt{n}, \bar{y}+c \cdot s/\sqrt{n}]$$