

16

Frequentist Inference

频率派统计推断

参数固定，但不可知，将概率解释为反复抽样的极限频率



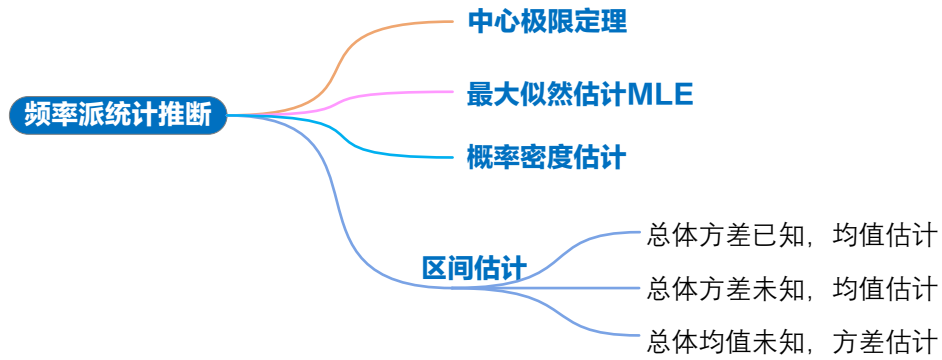
审视数学，你会发现，它不仅是颠扑不破的真理，而且是至高无上的美丽——那种冷峻而朴素的美，不需要唤起人们任何的怜惜，没有绘画和音乐的浮华装饰，纯粹，只有伟大艺术才能展现出来的严格完美。

Mathematics, rightly viewed, possesses not only truth, but supreme beauty — a beauty cold and austere, like that of sculpture, without appeal to any part of our weaker nature, without the gorgeous trappings of painting or music, yet sublimely pure, and capable of a stern perfection such as only the greatest art can show.

—— 伯特兰·罗素 (Bertrand Russell) | 英国哲学家、数学家 | 1872 ~ 1970



- ▶ `scipy.stats.binom_test()` 计算二项分布的 p 值
- ▶ `scipy.stats.norm.interval()` 产生区间估计结果
- ▶ `seaborn.heatmap()` 产生热图
- ▶ `seaborn.lineplot()` 绘制线型图
- ▶ `scipy.stats.ttest_ind()` 两个独立样本平均值的 t -检验



16.1 统计推断：两大学派

统计有两大分支：统计描述、统计推断。

本书第 2 章专门介绍了如何用图形和汇总统计量描述样本数据。而**统计推断** (statistical inference) 的数学工具来自于概率，本书“概率”、“高斯”、“随机”这三个板块给我们提供了足够的数学工具。因此，这个板块和下一板块正式进入统计推断这个话题。

本书前文提过，统计推断通过样本推断总体，在数据科学、机器学习应用颇为广泛。统计推断有两大大学派——**频率学派推断** (Frequentist inference) 和**贝叶斯学派推断** (Bayesian inference)。

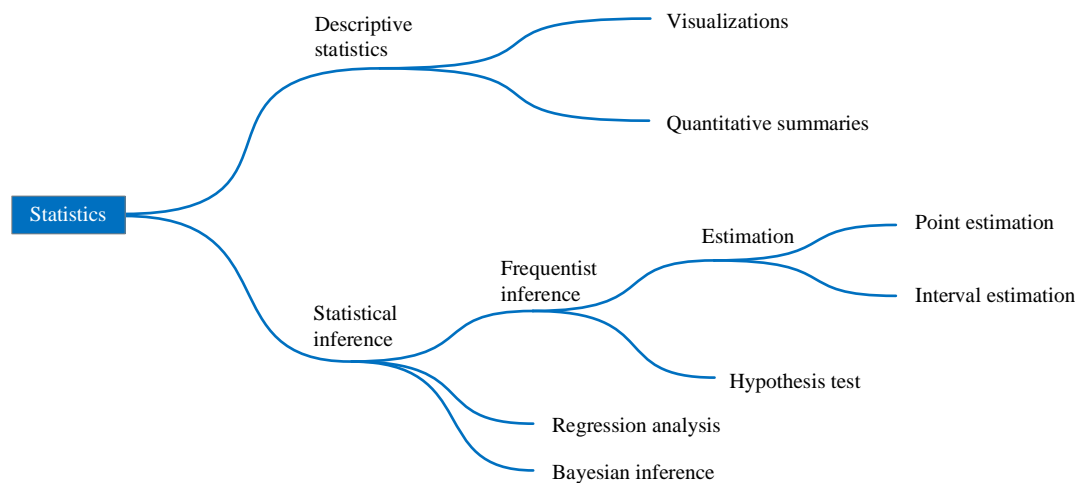


图 1. 本书统计学版图

频率学派

频率学派认为真实参数确定，但一般不可知。真实参数就好比上帝视角能够看到一切随机现象表象下的本质。

而我们观察到的样本数据都是在这个参数下产生的。真实参数对于我们不可知，频率派强调通过样本数据计算得到的频数、概率、概率密度等而得出有关总体的推断结论。

频率学派认为事件的概率是大量重复独立试验中频率的极限值。事件的可重复性、减小抽样误差对于频率派试验很重要。

频率学派方法的结论主要有两类：1) 显著性检验的“真或假”结论；2) 置信区间是否覆盖真实参数的结论。为了得出这些结论，我们需要掌握**区间估计** (interval estimation)、**最大似然估计** (maximum likelihood estimation, MLE)、**假设检验** (hypothesis test) 等数学工具。

这一章仅仅蜻蜓点水地介绍几个常用的频率学派工具，需要大家必须掌握的是最大似然估计 MLE。

▲ 注意，本书不会介绍假设检验。《数据有道》中讲解线性回归时会涉及到常见假设检验。

贝叶斯学派

贝叶斯学派则认为参数本身也是不确定的，参数本身也是随机变量，因此也服从某种概率分布。也就是说，所有参数都可能是产生样本数据的参数，只不过不同的参数对应的概率有大有小。

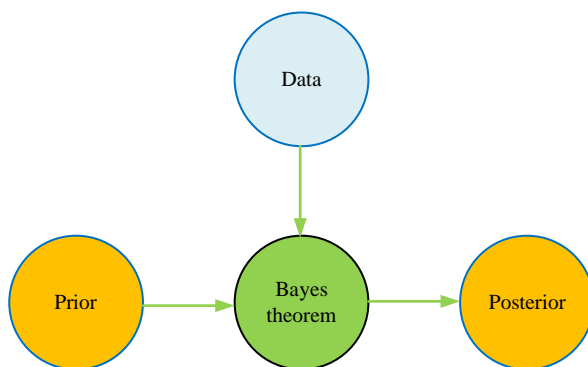


图 2. 贝叶斯推断

不同于频率派仅仅使用样本数据，贝叶斯学派结合过去的经验知识和样本数据。贝叶斯学派引入**先验分布** (prior distribution)、**后验分布** (posterior distribution)、**最大后验概率估计** (Maximum A Posteriori estimation, MAP) 这样的概念来计算不同参数值的概率。

比较来看，频率学派推断只考虑证据，不考虑先验概率。频率派强调概率是可重复性事件发生的频率，而不是基于主观判断的个人信念或偏好。

此外，很多情况下，贝叶斯推断没有后验分布的解析解，因此经常利用蒙特卡罗模拟获取满足特定后验分布的随机数。本书中大家会看到 Metropolis–Hastings 抽样算法的应用。

有意思的是，当样本数据量趋近无穷时，频率学派和贝叶斯学派结果趋于一致，可谓殊途同归。

贝叶斯统计能够整合主观、客观不同来源的信息，并作出合理判断，这是频率派推断做不到的。机器学习算法中，贝叶斯统计的应用越来越广泛。

➡ 本书前文提到，机器学习算法中频率学派的方法有其局限性。因此和常见的概率统计教材不同，本书“厚”贝叶斯学派，“薄”频率学派。本章和下一章将简要介绍频率学派统计推断的常用工具。而本书下一个板块将用五章内容专门介绍贝叶斯学派统计推断。

回归分析

回归分析 (regression analysis) 经常被划分到频率学派的工具箱中。作者则认为解释回归分析的视角很多，比如最小二乘优化视角、投影视角、矩阵分解、条件概率、最大似然估计 MLE、最大后验估计 MAP。因此，本书不把回归分析划在频率学派下面。

➔ 本书将在第 24 章从多视角来看回归分析。另外，《数据有道》一册则有专门讲解回归分析的板块，其中大家会看到拟合优度、方差分析 ANOVA、 F 检验、 t 检验、置信区间等工具在回归分析中的应用。除了线性回归，《数据有道》还会介绍非线性回归、贝叶斯回归、基于主成分分析的回归算法。

16.2 频率学派的工具

以鸢尾花数据为例

鸢尾花数据集最初由 Edgar Anderson 于 1936 年在加拿大加斯帕半岛上采集获得。在开始本章之前，先给大家出个问题，如何设计试验估算：

- ◀ 加斯帕半岛上所有鸢尾花花萼长度均值；
- ◀ 半岛上三类鸢尾花 (setosa、versicolour、virginica) 的具体比例。

为了解决这些实际问题，统计学家想出来了两个方法来解决。

大数定理

第一个办法是尽可能多地采集样本，比如在估算加斯帕半岛上所有鸢尾花花萼长度均值时，尽量同一时间采集尽可能多的鸢尾花数据。

这里应用到的统计学原理是**大数定律** (law of large numbers)。大数定律指的是当样本数量越多时，样本的算术平均值有越大的概率接近其真实的概率分布的期望。

简单来说，大数定理告诉我们，当我们进行大量的随机实验时，随着实验次数的增加，实际观测值越来越接近真实值。这就是大数定理的“大数”之处，有点“大力出奇迹”的味道。

大数定律体现出一些随机事件的均值具有长期稳定性。本书前文提到，抛一枚硬币，硬币落地正面朝上还是反面朝上，是偶然的。但是，如果硬币质地均匀，让我们抛硬币的次数达到上千上万次，就会发现硬币朝上的次数约为 50%。因此，频率学派推断特别强调同一试验的可重复性。

然而，这种办法需要尽可能多地提高样本数量，这使得试验本身变得尤为困难。

中心极限定理

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

第二种方法是，多次地独立地从总体中抽取样本，并计算每次样本的平均值，并用这些样本平均值去估算总体的期望。这种方法在统计学中被称为**中心极限定理** (central limit theorem)。

中心极限定理成立的条件包括：1) 独立性：独立随机变量之和的概率分布，需要基于独立随机变量的样本。2) 相同分布：随机变量应当具有相同的概率分布，即从同一总体中独立抽取样本。3) 样本量要足够大。

中学物理课，我们用游标卡尺反复测量同一物体的厚度，然后计算平均值来估计物体的实际厚度，这一试验的思路实际上就是中心极限定理的应用。

具体来说，中心极限定理指一个总体中随机进行 n 次抽样，每次抽取 m 个样本，计算其平均数，一共能得到 n 个平均数。当 n 足够大时，这 n 个平均数的分布接近于正态分布，不管总体的分布如何。这个定理，常常也被戏谑地称为“上帝视角”，在他眼中正态分布仿佛如同宇宙终极分布一般。

游标卡尺反复测量同一物体的厚度，可能会出现一些误差。这些误差可能来自于游标卡尺的不稳定性、读数不准确、人为误差等等因素。如果我们对这些误差进行统计分析，通常可以得到一个误差分布，该分布的中心点表示这些测量的平均值，标准差表示这些测量的离散程度。

当我们进行大量的游标卡尺测量时，由于中心极限定理的作用，这些误差的分布将趋向于正态分布。因此，我们可以使用正态分布模型来描述这些误差，从而对它们进行统计分析。这些分析包括计算平均值、标准差、置信区间等，可以帮助我们评估测量结果的准确性和稳定性，以及确定测量误差的来源。

点估计

点估计 (point estimation)，顾名思义，是指用样本统计量的某单一具体数值直接作为某未知总体参数的最佳估值。

举个例子，农场有几万只鸡兔。为了估计兔子的平均体重，我们从农场动物中随机抽取 100 只兔子作为样本，计算它们的平均体重为 5 kg。如果我们选择用 5 kg 代表整个农场所有兔子的体重，这种方法就是点估计。

本章主要介绍**最大似然估计** (Maximum Likelihood Estimation, MLE)。最大似然估计 MLE 在机器学习中应用广泛，MLE 和贝叶斯学派的最大后验概率估计 MAP 地位并列。

此外，点估计也用在贝叶斯推断中。贝叶斯推断中最常用的点估计是后验分布的期望值，称为后验期望。

区间估计

在用多次抽样估计总体分布的期望时，抽样的次数总是有限的，也有可能存在极端的样本值，这都会对估算产生影响。统计学家就想到一个更有效的办法，在进行估算时将注意力集中到样本平均值可能的一个范围或区间内，并给出真实的期望值位于这个区间的概率。这个区间就被称为**置信区间** (confidence interval, CI)。

举个例子，每次抽样的次数不变，做 100 次抽样，分别计算得到 100 个对应的样本平均值，并且认定在“上帝视角”中这 100 个样本平均值服从正态分布。那么，在这个正态分布的中心区域的 95 个样本均值，就构成了一个区间。这个区间就是对应的 95% 置信区间。它告诉我们，有 95% 的可能性总体真正的期望值在这个置信区间范围内。

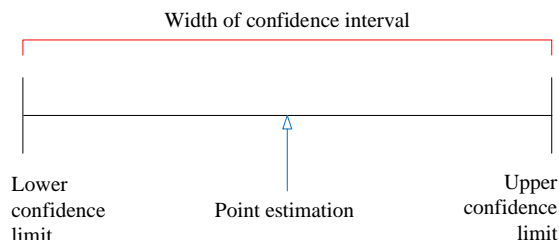


图 3. 对比点估计和区间估计

16.3 中心极限定理：渐近于正态分布

随机变量 X_1, X_2, \dots, X_n 独立同分布 IID，即相互独立且服从同一分布。 $X_k (k = 1, 2, \dots, n)$ 的期望和方差为：

$$E(X_k) = \mu, \quad \text{var}(X_k) = \sigma^2 \quad (1)$$

这 n 个随机变量的平均值 \bar{X} 近似服从如下正态分布：

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad (2)$$

注意，以上结论和 X_k 服从任何分布无关。

标准误 (standard error, SE) 的定义为：

$$SE = \frac{\sigma}{\sqrt{n}} \quad (3)$$

本节举两个例子来讲解中心极限定理。

离散

第一个例子是离散随机变量。

如图4所示为抛一枚色子结果 X 和对应的理论概率值。 X 服从离散均匀分布。如果每次抛 n 枚色子，这 n 个色子的结果对应 $X_1, X_2 \dots X_n$ 。然后求 n 个随机变量的平均值 \bar{X} 。根据 (2)， \bar{X} 服从正态分布 $N\left(\mu, \frac{\sigma^2}{n}\right)$ 。

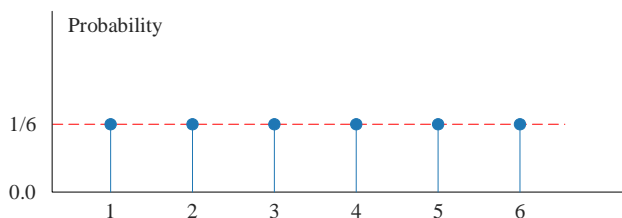
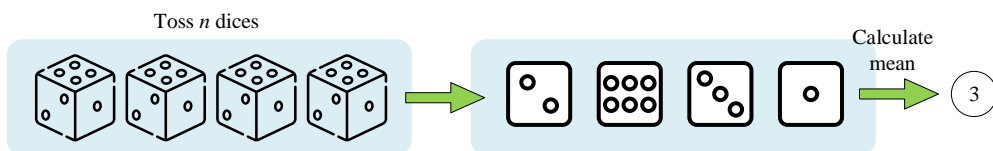


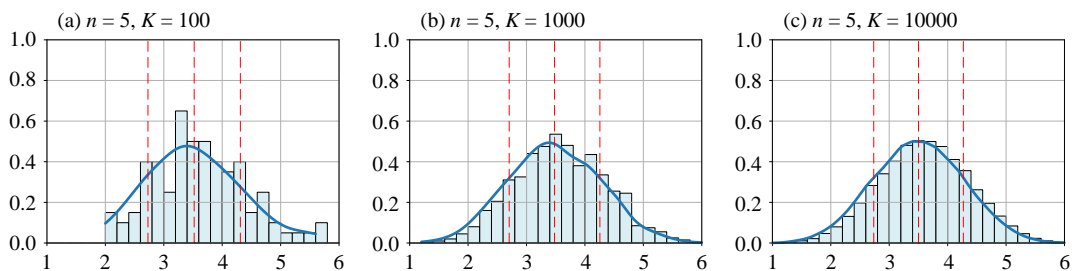
图 4. 抛一枚色子结果和对应的理论概率值

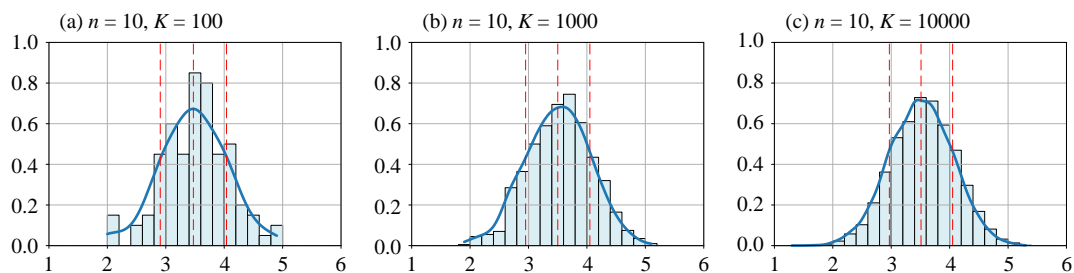
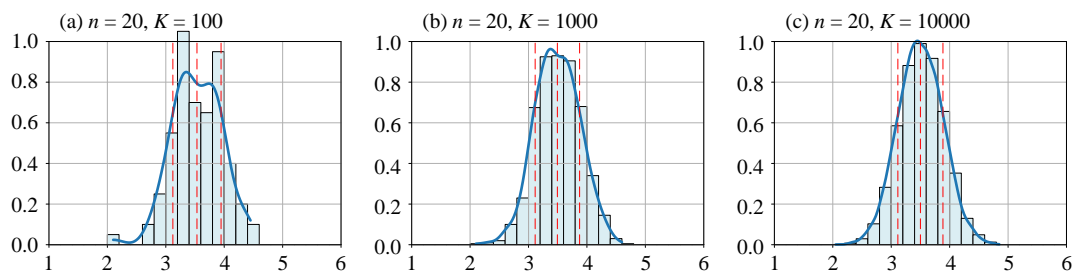
如图5所示，每次抛 n 枚色子，一共抛 K 次。下面，我们分别改变 n 和 K 进行蒙特卡罗模拟。

图 5. 每次抛 n 枚色子，一共抛 K 次

如图6所示，当 $n = 5$ 时，也就是每次抛 5 枚色子，随着 K 增大，我们很容易看出平均值 \bar{X} 趋向于正态分布。

根据 (2)，增大 n 会导致标准误 SE 会不断减小，对比图6、图7、图8，容易发现随着 n 增大，直方图逐渐变“瘦”，也就是说 SE 减小。

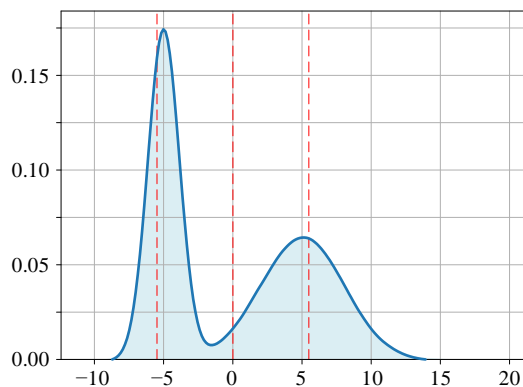
图 6. 每次抛 $n = 5$ 枚色子

图 7. 每次抛 $n = 10$ 枚色子图 8. 每次抛 $n = 20$ 枚色子

Bk5_Ch16_01.py 绘制图 6、图 7、图 8。

连续

第二个例子是连续随机变量。图 9 所示为随机数分布，这个分布有双峰，显然不是一个正态分布。如图 10 所示，试验中，每次抽取 $n = 10$ 个样本，随着试验次数 K 不断增大，平均值 \bar{X} 逐渐趋向于正态分布。图 11 中，这个趋势更加明显。图 12 所示为标准误 SE 随着 n 增大不断减小。



本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 9. 随机数分布

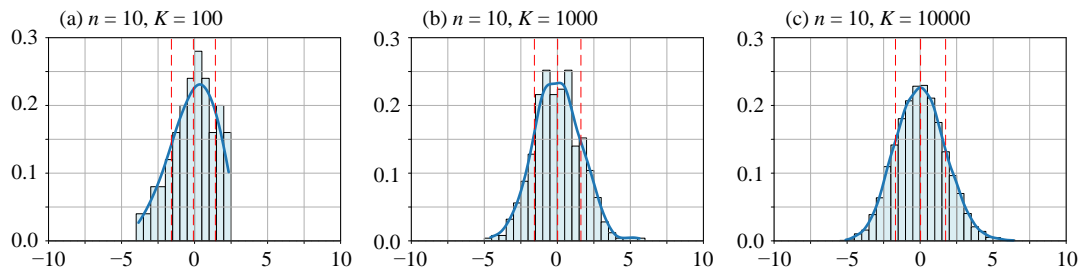


图 10. 每次抽取 10 个样本

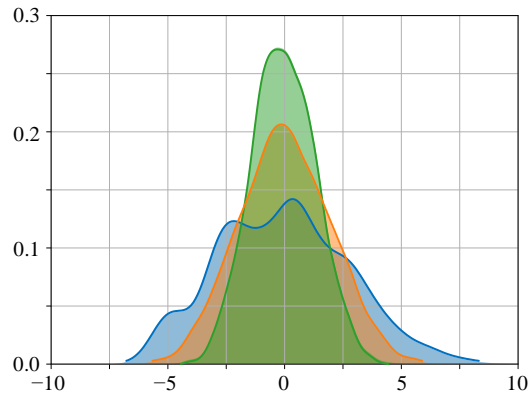
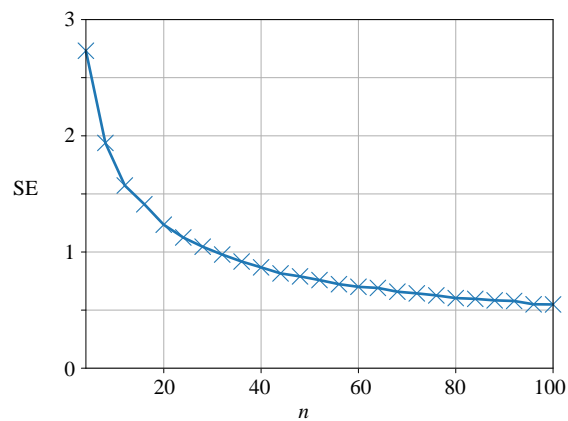


图 11. 随着试验次数增大，均值分布逐渐趋向正态

图 12. 标准误差随 n 变化

Bk5_Ch16_02.py 绘制图 9、图 10、图 11、图 12。

16.4 最大似然：鸡兔比例

白话说，最大似然估计 MLE 就是找到让似然函数取得最大值的参数。

鸡兔同笼

我们先看一个简单的例子。

试想，一个农场散养大量“走地”鸡、兔。假设农场的鸡兔比例真实值为 θ ，但是农夫自己并不清楚。为了搞清楚农场鸡兔比例，农夫决定随机抓 n 只动物。 $X_1, X_2 \dots X_n$ 为每次抓取动物的结果。 X_i 的样本空间为 $\{0, 1\}$ ，其中 0 代表鸡，1 代表兔。

⚠ 注意，抓取动物过程，我们忽略这对农场整体动物总体比例的影响。

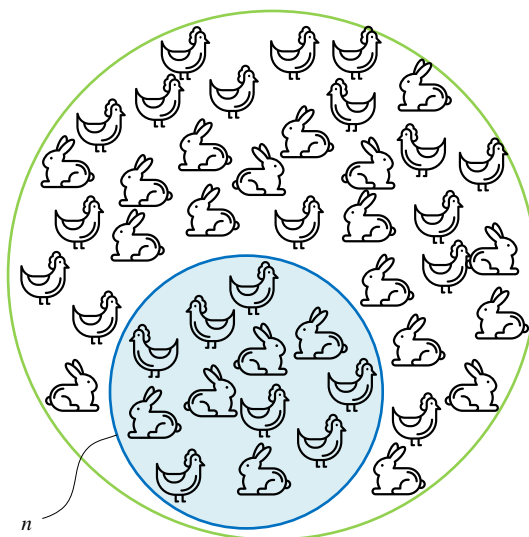


图 13. 农场有数不清的散养鸡兔

未知参数 θ

$X_1, X_2 \dots X_n$ 为 IID 的伯努利分布 $\text{Bernoulli}(\theta)$ ， X_i 的概率分布为：

$$f_{X_i}(x_i; \theta) = \theta^{x_i} (1 - \theta)^{1-x_i} \quad (4)$$

似然函数、对数似然函数一般用 θ (theta) 作为未知量。

⚠ 注意，上式本应该是概率质量函数，但是为了方便我们还是用 $f()$ 。

▲ 再次强调，本书前文提到过，为了避免混淆，本书用“|”引出条件概率中的条件，用分号“;”引出概率分布的参数。

似然函数

在统计学中，**似然函数** (likelihood function) 通常是通过观测数据的联合分布来定义。由于假设每个观测值都是独立同分布，所以上述联合概率可以被分解为每个观测值的边缘概率的乘积，即似然函数 $L(\theta)$ 为：

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n f_{x_i}(x_i; \theta) \\ &= \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} \\ &= \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i} \end{aligned} \quad (5)$$

简单来说，似然函数通常被表示为概率密度函数或概率质量函数的连乘积形式，这个连乘积表示观测数据的联合概率密度或概率质量函数。

令：

$$s = \sum_{i=1}^n x_i \quad (6)$$

s 代表 n 次抽取中兔子的总数。

这样 (5) 可以写成：

$$L(\theta) = \theta^s (1-\theta)^{n-s} \quad (7)$$

假设一次抓 20 只动物，其中 8 只兔子，则似然函数 $L(\theta)$ 为：

$$L(\theta) = \theta^8 (1-\theta)^{12} \quad (8)$$

图 14 (a) 所示为上述似然函数图像。显然，这个似然函数和横轴围成图形的面积不是 1。



本书第 20 章将介绍方法“归一化”似然函数。

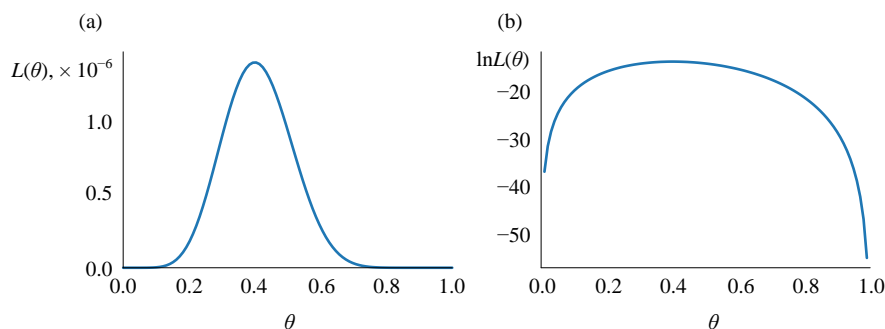


图 14. 似然函数、对数似然函数

MLE 优化问题为：

$$\arg \max_{\theta} \prod_{i=1}^n f_{X_i}(x_i; \theta) \quad (9)$$

对数似然函数

对数似然函数 (log-likelihood function) 就是对似然函数取对数，它可以将似然函数的连乘形式转换为加和形式：

$$\ln L(\theta) = s \ln \theta + (n - s) \ln(1 - \theta) \quad (10)$$

当 $n = 20$, $s = 8$ 时, (10) 为：

$$\ln L(\theta) = 8 \times \ln \theta + 12 \times \ln(1 - \theta) \quad (11)$$

图 14 (b) 所示为上述对数似然函数图像。



《数学要素》第 12 章提过，对数运算可以将连乘 Π 变成连加 Σ 。

在概率计算中，概率值累计乘积会经常出现数值非常小的正数情况。由于计算机的精度是有限的，无法识别这一类数据。而取对数之后，更易于计算机的识别，从而避免**浮点数下溢** (floating point underflow)。浮点数下溢，也叫**算术下溢** (arithmetic underflow)，指的是计算机浮点数计算的结果小于可以表示的最小数。

在最大化似然函数时，由于对数函数是单调递增的，因此最大化对数似然函数的值等价于最大化原始似然函数的值。此外，对数似然函数在计算导数时也更加方便，因为它将连乘变为加和形式，从而可以更容易地进行求导。因此，对数似然函数常常被用于最大似然估计和贝叶斯推断等统计学方法中。

优化问题

有了对数似然函数，(18) 中的 MLE 优化问题可以写成：

$$\arg \max_{\theta} \sum_{i=1}^n \ln f_{X_i}(x_i; \theta) \quad (12)$$

(10) 中 $\ln L(\theta)$ 对 θ 求偏导为 0，构造等式：

$$\frac{d \ln L}{d \theta} = \frac{s}{\theta} - \frac{n-s}{1-\theta} = 0 \quad (13)$$

求解上式得到：

$$\hat{\theta}_{MLE} = \frac{s}{n} \quad (14)$$



我们将在本书第 21 章用贝叶斯派统计推断重新求解这个问题。

16.5 最大似然：以估算均值、方差为例

设 $X \sim N(\mu, \sigma^2)$ ， μ 和 σ^2 为未知参数。

X_1, X_2, \dots, X_n 来自 X 的 n 个样本，显然 X_1, X_2, \dots, X_n 独立同分布。 x_1, x_2, \dots, x_n 是 X_1, X_2, \dots, X_n 的观察值。下面介绍利用最大似然方法求解 μ 和 σ^2 的估计量。

X_i 的概率密度函数为：

$$f_{X_i}(x_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left(-\frac{1}{2 \sigma^2} \left(x - \mu \right)^2 \right) \quad (15)$$

未知参数 θ

令 $\theta_1 = \mu, \theta_2 = \sigma^2$ ， X_i 的概率密度函数则写成：

$$f_{X_i}(x_i; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi \theta_2}} \exp \left(-\frac{1}{2\theta_2} (x_i - \theta_1)^2 \right) \quad (16)$$

似然函数

似然函数 $L(\theta_1, \theta_2)$ 为 $f_{X_i}(x_i; \theta_1, \theta_2)$ 的连乘：

$$\begin{aligned}
L(\theta_1, \theta_2) &= f_{x_1}(x_1; \theta_1, \theta_2) \cdot f_{x_2}(x_2; \theta_1, \theta_2) \cdots f_{x_n}(x_n; \theta_1, \theta_2) \\
&= \prod_{i=1}^n f_X(x_i; \theta_1, \theta_2) \\
&= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\theta_2}} \exp\left(\frac{-1}{2\theta_2}(x_i - \theta_1)^2\right)
\end{aligned} \tag{17}$$

对数似然函数

对 (17) 取对数得到 $\ln L(\theta_1, \theta_2)$:

$$\ln L(\theta_1, \theta_2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\theta_2) - \frac{1}{2\theta_2} \left(\sum_{i=1}^n (x_i - \theta_1)^2 \right) \tag{18}$$

优化问题

为了最大化 (18) 中 $\ln L(\theta_1, \theta_2)$, 对 θ_1 、 θ_2 求偏导为 0, 构造等式:

$$\begin{aligned}
\frac{\partial \ln L}{\partial \theta_1} &= \frac{1}{\theta_2} \left(\sum_{i=1}^n (x_i - \theta_1) \right) = 0 \\
\frac{\partial \ln L}{\partial \theta_2} &= -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2} \left(\sum_{i=1}^n (x_i - \theta_1)^2 \right) = 0
\end{aligned} \tag{19}$$

可以求得:

$$\begin{aligned}
\hat{\theta}_1 &= \frac{\sum_{i=1}^n x_i}{n} = \bar{X} \\
\hat{\theta}_2 &= \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n}
\end{aligned} \tag{20}$$

“戴帽子”的 $\hat{\theta}_1$ 、 $\hat{\theta}_2$ 为对真实 θ_1 、 θ_2 的估计。注意, 上式中 $\hat{\theta}_2$ 并不是对方差的无偏估计。

具体值

给定样本为 $\{-2.5, -5, 1, 3.5, -4, 1.5, 5.5\}$, 下面用 MLE 估算其均值和方差。

将样本代入 (18), 得到对数似然函数:

$$\ln L(\theta_1, \theta_2) = -6.432 - 3.5 \ln \theta_2 - \frac{7\theta_1^2 + 93}{2\theta_2} \tag{21}$$

$\ln L(\theta_1, \theta_2)$ 对 θ_1 、 θ_2 求偏导为 0, 构造等式:

$$\begin{aligned}\frac{\partial \ln L}{\partial \theta_1} &= -\frac{7\theta_1}{\theta_2} = 0 \\ \frac{\partial \ln L}{\partial \theta_2} &= \frac{7\theta_1^2 - 7\theta_2 + 93}{2\theta_2^2} = 0\end{aligned}\quad (22)$$

求解上式得到：

$$\begin{aligned}\hat{\theta}_1 &= 0 \\ \hat{\theta}_2 &= 13.2857\end{aligned}\quad (23)$$

并计算得到对数似然函数的最大值：

$$\max \{\ln L(\theta_1, \theta_2)\} = -18.98598 \quad (24)$$

图 15 所示为 $\ln L(\theta_1, \theta_2)$ 曲面，× 对应对数似然函数最大值点位置。



本书第 24 章中，我们将用到 MLE 估算线性回归参数。

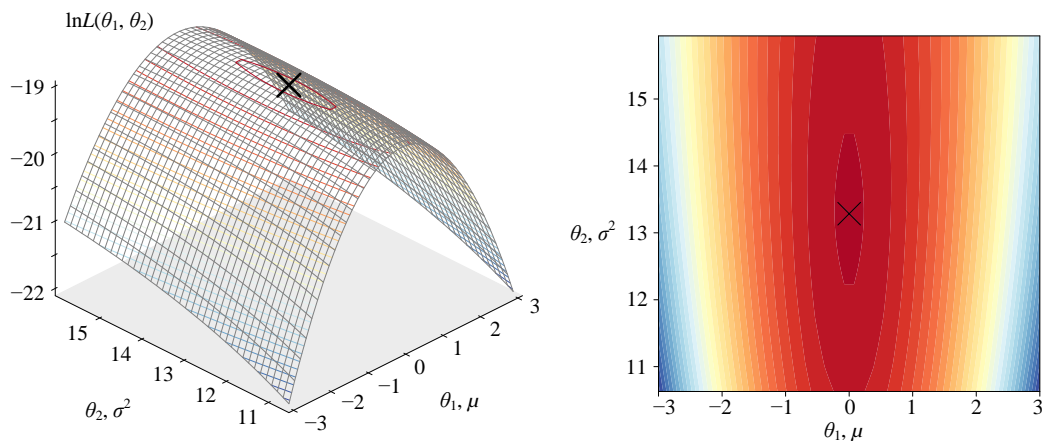


图 15. $\ln L(\theta_1, \theta_2)$ 曲面和最大值点位置



Bk5_Ch16_03.py 绘制图 15。

16.6 区间估计：总体方差已知，均值估计

不同于点估计仅估出一个数值，**区间估计** (interval estimate) 在推断总体参数时，根据统计量的抽样分布特征，估算出总体参数的一个区间范围，并且估算出总体参数落在这一区间的概率。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

区间估计在点估计的基础上附加**误差限** (margin of error) 来构造**置信区间** (confidence interval)，置信区间对应的概率，被称为**置信度** (confidence level)。

本节介绍总体方差 σ^2 已知，计算给定置信水平下均值的区间估计。

双边置信区间

对于样本数据 $\{x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(n)}\}$ ，计算**样本平均值** (sample mean 或 empirical mean)：

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x^{(i)} \quad (25)$$

如果总体的方差已知，总体平均值 μ 的 $1 - \alpha$ 水平的**双边置信区间** (two tailed confidence interval) 可以表达为：

$$\left(\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right) \quad (26)$$

其中：

\bar{X} 为**样本均值** (sample mean)。

n 为**样本数量** (sample size)。

α 为**显著性水平** (significance level)，代表的意义是在一次试验中小概率事物发生的可能性大小。 α 通常取 0.1 或 0.05。

$1 - \alpha$ 为**置信水平** (confidence level)，表示真值在置信区间内的可信程度。

$z_{1-\alpha/2}$ 叫**临界值** (critical value)，本质上就是 Z 分数。 $z_{1-\alpha/2}$ 可以通过标准正态分布的逆累积概率密度分布函数计算。

σ 为**总体的标准差** (volatility of the population)。

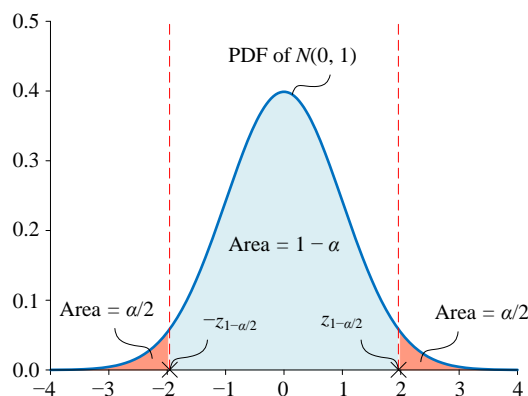
如图 16 所示， $1 - \alpha$ 为置信水平意味着：

$$\Pr \left(\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha \quad (27)$$

求解 $z_{1-\alpha/2}$ 的方法为：

$$z_{1-\alpha/2} = F_{N(0,1)}^{-1} \left(1 - \frac{\alpha}{2} \right) = -F_{N(0,1)}^{-1} \left(\frac{\alpha}{2} \right) \quad (28)$$

$F_{N(0,1)}^{-1}(\cdot)$ 是标准正态分布的**逆累计分布函数** (inverse cumulative distribution function, ICDF)。这和本书前文介绍的百分点函数 PPF 本质上一致。

图 16. 标准正态分布和 $1 - \alpha$ 置信水平

95%置信水平

总体方差已知，95% ($1 - \alpha = 1 - 5\%$) 置信水平的双边置信区间约为：

$$\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right) \quad (29)$$

也就是说：

$$\Pr \left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right) \approx 0.95 \quad (30)$$

再次强调区间估计得到的是总体参数落在某一区间的概率。图 17 (a) 所示为 100 次估算得到的 95% 置信水平的双尾置信区间。图中，黑色竖线为总体均值所在位置。

\times 代表每次估算样本均值所在位置。当总体均值落在双尾置信区间时，区间为蓝色；否则，区间为红色。图 17 (a) 给出的 100 个区间中，有 88 个双尾区间包含真实的总体均值；12 个双尾区间不包含真实的总体均值。图 17 (b) 为每次抽取得到的样本数据分布山脊图。

增大每次抽样样本数量 n ，左侧置信区间不断收窄，而右侧分布范围不断变宽，两者并不矛盾。请大家思考背后原因。

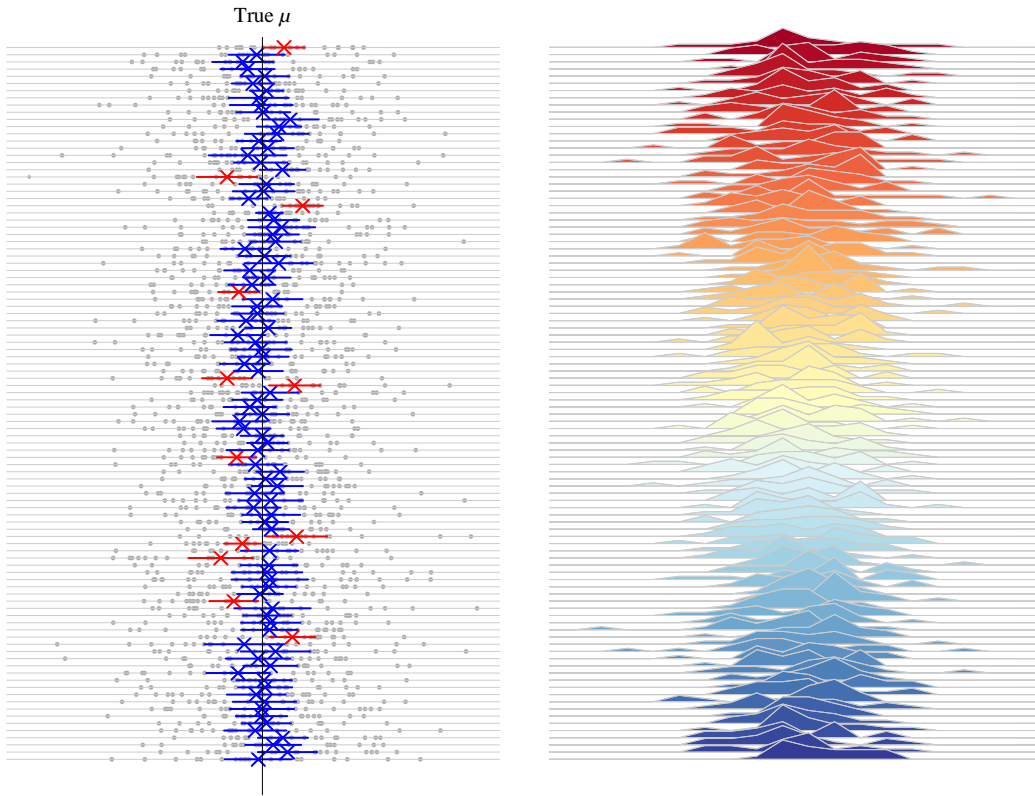


图 17. 100 次估算得到的 95% 质心水平的双尾置信区间，每次数据的分布的山脊图

单边置信区间

除了双边置信区间，统计上还常用**单边置信区间** (one-tailed confidence interval)。单边置信区间可以“左尾”，即取值范围从负无穷到平均值 \bar{X} 右侧的临界值：

$$\left(-\infty, \bar{X} + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}\right) \quad (31)$$

这意味着：

$$\Pr\left(\mu < \bar{X} + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \quad (32)$$

单边置信区间也可以是“右尾”，取值范围从 \bar{X} 左侧的临界值到正无穷：

$$\left(\bar{X} - z_{1-\alpha} \frac{\sigma}{\sqrt{n}}, +\infty\right) \quad (33)$$

这意味着：

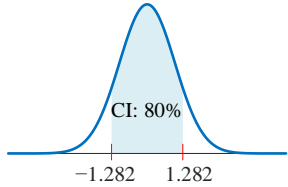
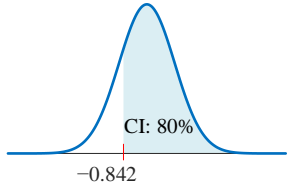
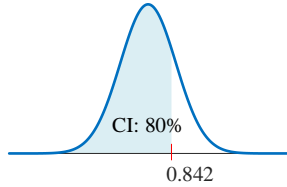
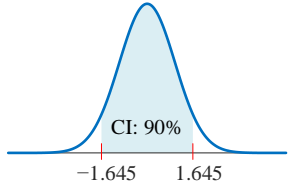
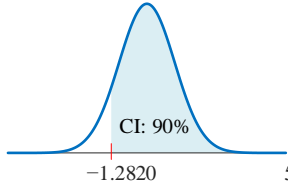
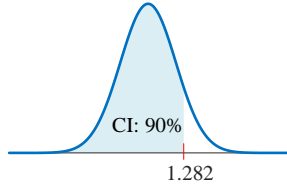
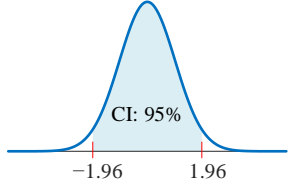
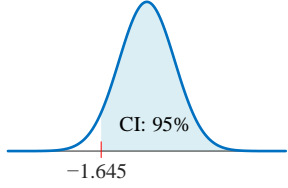
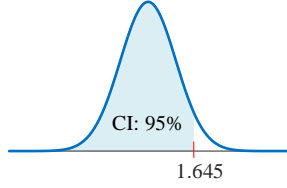
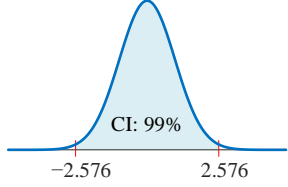
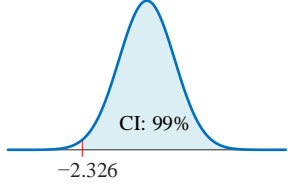
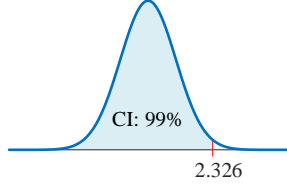
$$\Pr\left(\mu > \bar{X} - z_{1-\alpha} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \quad (34)$$

举个例子，总体方差已知，95% ($1 - \alpha = 1 - 5\%$) 水平的单侧置信区间分别为：

$$\left(-\infty, \bar{X} + 1.645 \frac{\sigma}{\sqrt{n}}\right), \left(\bar{X} - 1.645 \frac{\sigma}{\sqrt{n}}, +\infty\right) \quad (35)$$

表 1 所示为不同显著性水平的双尾、左尾、右尾置信区间。

表 1. 不同显著性水平的置信区间

显著性水平 置信水平	双尾	左尾	右尾
$\alpha = 20\%$ $1 - \alpha = 80\%$			
$\alpha = 10\%$ $1 - \alpha = 90\%$			
$\alpha = 5\%$ $1 - \alpha = 95\%$			
$\alpha = 1\%$ $1 - \alpha = 99\%$			



Bk5_Ch16_04.py 绘制表 1 中图像。

16.7 区间估计：总体方差未知，均值估计

如果总体方差 σ^2 未知，就不能用上一节的估算方法。

首先，计算样本方差 s^2 ：

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x^{(i)} - \bar{X})^2 \quad (36)$$

样本均方差 s 为：

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x^{(i)} - \bar{X})^2} \quad (37)$$

如果总体的方差未知，总体平均值 μ 的 $1 - \alpha$ 置信水平的**双边置信区间** (two tailed confidence interval) 为：

$$\left(\bar{X} - t_{1-\alpha/2}(n-1) \frac{s}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2}(n-1) \frac{s}{\sqrt{n}} \right) \quad (38)$$

其中， n 为样本数量； $t_{1-\alpha/2}(n-1)$ 为自由度 $n-1$ ，CDF 值为 $1 - \alpha/2$ 的学生- t 的逆累计分布值。图 18 所示为自由度为 5 时， $1 - \alpha$ 置信水平双尾置信区间对应位置。

自由度较小时，学生- t 分布有明显的厚尾现象。由于厚尾现象的存在，同样的置信区间，学生 t 分布的临界值的绝对值要大于标准正态分布。但是当自由度 $df = n - 1$ 不断提高，学生- t 分布逐渐接近标准正态分布。

图 19 所示为总体方差未知，总体平均值 μ 的 $1 - \alpha$ 置信水平的右尾/左尾置信区间。

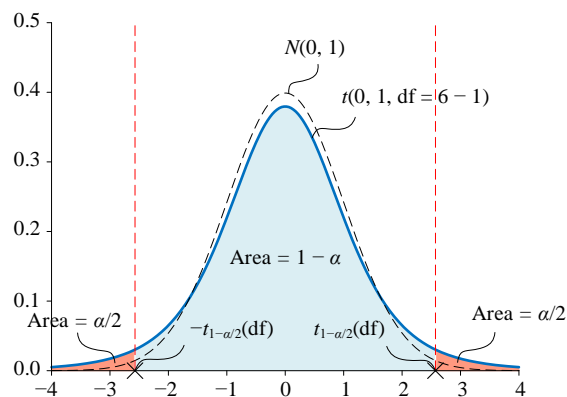
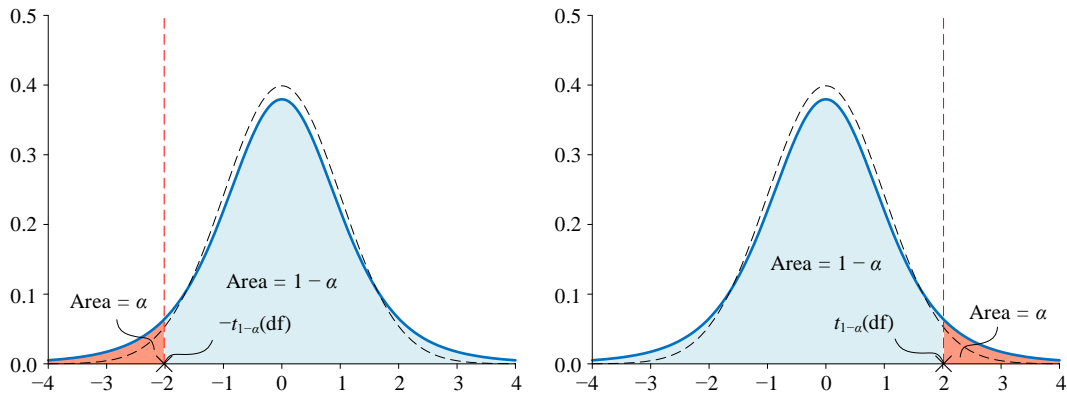
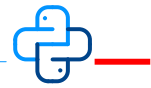


图 18. 总体方差未知，总体平均值 μ 的 $1 - \alpha$ 置信水平的双尾置信区间

图 19. 总体方差未知，总体平均值 μ 的 $1 - \alpha$ 置信水平的右尾/左尾置信区间

Bk5_Ch16_05.py 绘制图 18 和图 19。

16.8 区间估计：总体均值未知，方差估计

总体均值未知情况下， σ^2 的无偏估计为 s^2 ：

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x^{(i)} - \bar{X})^2 \quad (39)$$

方差 σ^2 的 $1 - \alpha$ 水平的**双边置信区间** (two tailed confidence interval) 为：

$$\left(\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2(n-1)}, \frac{(n-1)s^2}{\chi_{\alpha/2}^2(n-1)} \right) \quad (40)$$

其中， n 为样本数量； $\chi_{\alpha/2}^2(n-1)$ 为自由度 $n-1$ 的卡方分布。我们还会在本书第 23 章有关马氏距离的内容用到卡方分布。

(40) 意味着：

$$\Pr \left(\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2(n-1)} < \sigma^2 < \frac{(n-1)s^2}{\chi_{\alpha/2}^2(n-1)} \right) = 1 - \alpha \quad (41)$$

上式开方，得到标准差 σ 的 $1 - \alpha$ 水平的**双边置信区间**可以表达为：

$$\left(\frac{\sqrt{n-1}s}{\sqrt{\chi_{1-\alpha/2}^2(n-1)}}, \frac{\sqrt{n-1}s}{\sqrt{\chi_{\alpha/2}^2(n-1)}} \right) \quad (42)$$

图 20 所示为总体均值未知，方差估计的 $1 - \alpha$ 置信水平的双尾置信区间。

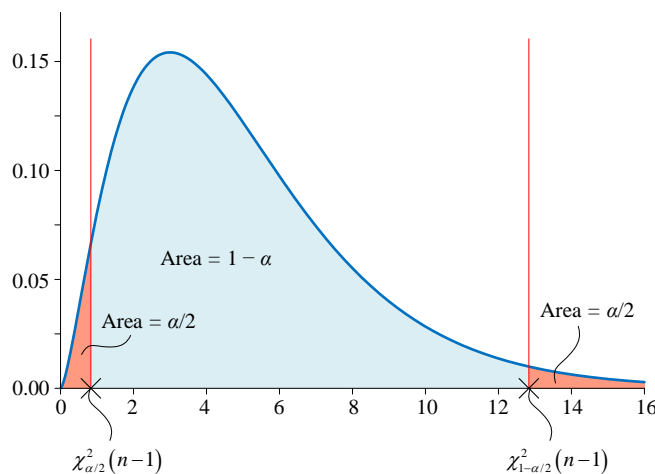


图 20. 总体均值未知，方差估计的 $1 - \alpha$ 置信水平的双尾置信区间



Bk5_Ch16_06.py 绘制图 20。

本书首先比较了统计推断的两大流派——频率学派、贝叶斯学派。频率学派认为概率是事件发生的频率，以样本为基础进行推断；而贝叶斯学派则将概率视为主观信念的度量，以先验知识为基础进行推断。两者的不同在于对概率的定义和解释方式，但两者也可以相互补充。

然后，我们简单地了解了常用的频率学派数学工具。再次说明，《统计至简》一册轻频率学派，重贝叶斯学派。这是因为机器学习、深度学习中贝叶斯学派的思想、方法、工具戏份十足。

下一章聊一聊另外一个机器学习中常用的频率学派工具——概率密度估计。



有关如何用 Python 完成假设检验，请大家自学 Stanford 这门统计学课程相关内容。

https://web.stanford.edu/class/stats110/notes/Chapter6/Large_sample.html

这门课程网站还有大量 Python 和概率统计相结合的实例，很适合初学者参考。