参数估计

第 10 讲 • 概率论与数理统计

复旦大学经济学院　ECON130001

上一讲我们做了什么

造了四个统计量，并算出了它们的分布

$\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$	σ 已知
$\dfrac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$	σ 未知
$\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	推断方差
$\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$	比较两个方差

★ 留下的问题

上一讲只说了 $\overline{X}$ 「逼近」μ、S^2「逼近」σ^2。 **可凭什么是它们？**换一个式子行不行？

更要紧的是：**估出来的那个数，离真值到底有多远？** 只报一个数字，等于什么都没说。

为什么需要这一讲

场景　1801 年 1 月 1 日，皮亚齐在巴勒莫发现一颗新天体（后来的**谷神星**）。 他跟了 41 天，它在天球上只走过 **9 度**，随后转到太阳背后，看不见了。 要等半年多才可能重新出现——**但没人知道该往哪儿看。**

冲突　定一条轨道，理论上**三次观测就够**。皮亚齐留下了二十来次记录。 当时一位能干的天文学家会说：**从里面挑三次最可靠的，解方程。**
可挑哪三次？不同的三组给出的轨道差好几度—— 而在茫茫天区里差几度，就是**永远找不到**。 24 岁的高斯说：**一次都不能扔，全都要用。**

悬念　**可把二十多次观测全放进去，方程比未知数多，根本无解。「最贴近全部数据的那条轨道」——这句话听着很对， 可它在数学上到底是什么意思？「最好」凭什么算最好?**

来源
Piazzi 于 1801 年 1 月 1 日在巴勒莫天文台发现谷神星，观测约 41 天、天球弧长约 9°， 随后天体没入日光。Gauss 依其计算给出预报位置，von Zach 于 1801 年 12 月 7 日、Olbers 于 12 月 31 日先后在预报位置附近重新找到该天体。 Gauss 在 *Theoria Motus Corporum Coelestium*（1809）中论证： 若观测误差服从正态分布，则使似然最大的参数正是使残差平方和最小的参数。 thatsmaths.com ｜Rutgers · Gauss and Ceres ｜访问日期 2026-08-02。

本讲学习目标

1. 用**一致性、无偏性、有效性**三把尺子判断一个估计量好不好
2. 会用**矩估计**与**最大似然估计**两种方法，从数据里把参数解出来
3. 会构造**置信区间**：不只报一个数，还报出这个数有多准

本讲地图

- **点估计**——报一个数
  - 矩估计（让模型的矩去对上样本的矩）
  - 最大似然估计（哪个参数让手上这批数据最可能出现）
- **区间估计**——报一个范围，并说明有多大把握

什么是参数估计

问题的形状

实际工作中，我们往往**知道总体分布的大致类型**（正态、指数、二项……），**但不知道其中的参数**。参数估计就是用样本把它算出来：

$$\hat{\theta}(X_1, X_2, \cdots, X_n)$$

称为 θ 的**估计量**。

注意它是**统计量**——不含未知参数，拿到数据就能算。

经济学里的样子

消费函数 $C = a + bY$。边际消费倾向 b 就是要估的参数—— 它决定了财政刺激的乘数有多大。

AI 里的样子

一个图像识别模型内部有上亿个参数。**所谓「训练模型」，本质就是用大量数据去估计这些参数**——用的正是本讲后半的最大似然思想。

第一把尺子：一致性

三把尺子

一致性（样本够多时收敛到真值）→ **无偏性**（平均而言不偏）→ **有效性**（波动更小）

一致估计

若 $n \to \infty$ 时 $\hat{\theta}$ 依概率收敛于 θ，即对任意 $\varepsilon > 0$

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$$

则称 $\hat{\theta}$ 是 θ 的**一致估计**。

哪些是一致估计

- 样本均值 $\overline{X}$ 是期望的一致估计——**这就是第 8 讲的大数定律**
- 样本方差 S^2 是方差的一致估计
- $\frac{1}{n}\sum(X_i - \overline{X})^2$ 也是方差的一致估计

注意最后两条：分母除 $n-1$ 还是 n，都一致。 一致性是个「大样本」的要求，n 很大时差一个分母无所谓。

第二把尺子：无偏性

无偏估计

若 $E\hat{\theta} = \theta$，则称 $\hat{\theta}$ 是 θ 的**无偏估计**。

「无偏」说的是**把整个抽样过程重复无穷多次，估计值的平均正中真值** ——不是说某一次估得准。

是无偏估计

- $\overline{X}$ 是 μ 的无偏估计
- X_1 **（只取第一个观测）也是** μ **的无偏估计**
- S^2 是 σ^2 的无偏估计

不是无偏估计

- $\frac{1}{n}\sum(X_i - \overline{X})^2$ **不是** σ^2 的无偏估计

——上一讲那个 $n-1$ 的分母，就是为了这一条。

★ 请盯住左栏第二条

X_1 无偏，**却连一致都不是**：样本再多，它也只看第一个数，永远不收敛。 **无偏与一致，是两件互不蕴含的事。**

反过来 $\frac{1}{n}\sum(X_i - \overline{X})^2$ 有偏、却一致。 这说明**一把尺子不够用**。

无偏性：两个独立样本怎么合并

定理

从同一总体中抽取两个独立样本 $(X_{11},\cdots,X_{1n_1})$ 与 $(X_{21},\cdots,X_{2n_2})$，则

$$\overline{X}=\frac{1}{n_1+n_2}\left(n_1\overline{X_1}+n_2\overline{X_2}\right)$$

$$S^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$

分别是总体期望与方差的无偏估计量。

两个式子都是「按自由度加权」

均值按**样本量**加权，方差按**自由度** n_i-1 加权。 分母 n_1+n_2-2：**两组各损失一个自由度**，因为各自的均值都是从数据里算出来的。

第二个式子就是上一讲两样本 t 统计量里的**合并方差** S_p^2 ——它在本讲区间估计、下一讲假设检验里还会出现三次。

无偏性例题：同一个参数的两个无偏估计

例

设总体服从参数为 λ 的指数分布，令 $\theta = 1/\lambda$。证明 $\overline{X}$ 与 $n \cdot \min(X_1, \cdots, X_n)$ **都是** θ 的无偏估计。

(1) $\overline{X}$

$E(\overline{X}) = E(X) = 1/\lambda = \theta$。完事。

(2) 最小值

令 $Z = \min(X_1, \cdots, X_n)$。**最小值大于** z**，当且仅当每一个都大于** z：

$$P(Z > z) = \prod_{i=1}^{n} P(X_i > z) = e^{-n\lambda z}$$

故 $F_Z(z) = 1 - [1 - F_X(z)]^n = 1 - e^{-n\lambda z}$ ——Z **服从参数** $n\lambda$ **的指数分布。**

于是 $E(Z) = \dfrac{1}{n\lambda}$，$E(nZ) = \dfrac{1}{\lambda} = \theta$。

两个都无偏，**那该用哪一个？**无偏性到这里就分不出高下了 ——第三把尺子登场。

第三把尺子：有效性

有效估计

$\hat{\theta}$ 与 $\hat{\theta}'$ 都是 θ 的无偏估计。样本容量给定时， 若 $D(\hat{\theta}) < D(\hat{\theta}')$，则称 $\hat{\theta}$ **比** $\hat{\theta}'$ **有效**。

若在 θ 的**一切**无偏估计中 $\hat{\theta}$ 的方差最小， 则称 $\hat{\theta}$ 为 θ 的**有效估计量**。

回到上一页那道题

$$D(\overline{X}) = \frac{D(X)}{n} = \frac{\theta^2}{n}$$

$$D\big(n\min(X_1, \cdots, X_n)\big) = n^2 D(Z) = n^2 \cdot \frac{\theta^2}{n^2} = \theta^2$$

差了 n 倍。 $n = 100$ 时，$\overline{X}$ 的方差只有它的百分之一。 答案很清楚：**用 $\overline{X}$。**

为什么差这么多？因为 $n\min$ **只用了一个数据点**（最小的那个）， 其余 $n-1$ 个全扔了。**用上全部信息的估计量，方差才会随 n 变小。**

——这正是开场故事里高斯的那句话：**一次观测都不能扔。**

有效性例题：加权平均比不过算术平均

例

比较总体期望的两个无偏估计 $\overline{X}=\frac{1}{n}\sum X_i$ 与 $X'=\sum a_iX_i\Big/\sum a_i$（$a_i>0$）的有效性。

两者都无偏（X' 的权重之和为 1）。方差：

$$D(\overline{X})=\frac{\sigma^2}{n},\qquad D(X')=\frac{\sum a_i^2}{\left(\sum a_i\right)^2}\sigma^2$$

用 $a^2+b^2\geq 2ab$：

$$\left(\sum_{i=1}^{n}a_i\right)^2=\sum_i a_i^2+\sum_{i\neq j}a_ia_j\leq\sum_i a_i^2+\sum_{i\neq j}\frac{a_i^2+a_j^2}{2}=n\sum_i a_i^2$$

故 $\dfrac{\sum a_i^2}{(\sum a_i)^2}\geq\dfrac{1}{n}$，即 $D(\overline{X})\leq D(X')$。

结论值得记住

在所有**线性无偏**估计里，**等权重的算术平均方差最小**；只有当 $a_1=\cdots=a_n$ 时取等号。

换句话说：**数据同质时，别自作聪明加权。**（若各观测精度不同，最优权重与方差成反比——那是计量经济学的加权最小二乘。）

小结：三把尺子

标准	说的是什么	一句话
一致性	$n \to \infty$ 时收敛到真值	底线，不是优点
无偏性	$E\hat{\theta} = \theta$	重复无穷多次，平均正中靶心
有效性	无偏估计中方差最小	枪法更稳，弹着点更集中

- 三者**互不蕴含**：X_1 无偏不一致，$\frac{1}{n}\sum(X_i - \overline{X})^2$ 有偏却一致
- 有效性只在**无偏**的估计量之间比较——先无偏，再比方差
- 用上全部数据的估计量才可能有效

矩估计：一个朴素得近乎理所当然的想法

核心思想：让模型去模仿样本

既然样本是从总体里抽出来的，样本的特征就该和总体的特征相似。 「矩」是刻画分布特征的量——一阶矩是期望，二阶矩与方差有关。

让模型的理论矩，等于从样本里算出来的矩。有几个未知参数，就列几个方程。

它的位置

由「数理统计之父」**卡尔·皮尔逊**在 19 世纪末提出，是最早的一般化参数估计方法。 今天有了更强的最大似然法，但矩估计因为思想直白、计算简便， 仍是理解参数估计的起点。

矩估计：定义与做法

总体矩与样本矩

总体 X 的 k **阶原点矩**：$\mu_k = E(X^k)$。

样本的 k **阶样本矩**：$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$。

由大数定律，$n \to \infty$ 时 A_k 依概率收敛于 μ_k ——**整个方法的合法性就来自这一句。**

三步走

1. 把前 k 阶总体矩写成 k 个未知参数的函数：$\mu_j = \mu_j(\theta_1, \cdots, \theta_k)$
2. **反解**出 $\theta_j = \theta_j(\mu_1, \cdots, \mu_k)$
3. 把 μ_j 换成 A_j，得 $\hat{\theta}_j = \theta_j(A_1, \cdots, A_k)$

几个未知参数，就用几阶矩。

矩估计：指数分布与均匀分布

指数分布（参数 λ）

$\mu_1 = 1/\lambda \;\Rightarrow\; \lambda = 1/\mu_1$，代入样本矩：

$$\hat{\lambda} = \frac{1}{\overline{X}}$$

一个参数，一阶矩就够。

均匀分布 $U(a,b)$

$\mu_1 = \dfrac{a+b}{2}$，$\mu_2 = \dfrac{(b-a)^2}{12} + \dfrac{(a+b)^2}{4}$，解得

$a, b = \mu_1 \mp \sqrt{3(\mu_2 - \mu_1^2)}$，故

$$\hat{a}, \hat{b} = \overline{X} \mp \sqrt{\frac{3}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

两个参数，用到二阶矩。

右边那个解不必硬记：注意 $\mu_2 - \mu_1^2$ 就是方差 $\frac{(b-a)^2}{12}$，所以 $b-a = \sqrt{12(\mu_2-\mu_1^2)}$，**再把区间中点 μ_1 往两边各推一半。**

矩估计：二项分布与正态分布

二项分布（n, p 均未知）

$\mu_1 = np$，$\mu_2 = np(1-p) + n^2p^2$

注意 $\mu_1^2 + \mu_1 - \mu_2 = np^2$，故

$$p = \frac{\mu_1^2 + \mu_1 - \mu_2}{\mu_1}, \qquad n = \frac{\mu_1^2}{\mu_1^2 + \mu_1 - \mu_2}$$

代入样本矩（记 $B^2 = \frac{1}{n}\sum(X_i - \overline{X})^2$）：

$$\hat{p} = \frac{\overline{X} - B^2}{\overline{X}}, \qquad \hat{n} = \frac{\overline{X}^2}{\overline{X} - B^2}$$

正态分布（μ, σ^2）

$\mu_1 = \mu$，$\mu_2 = \sigma^2 + \mu^2$，解得 $\mu = \mu_1$、$\sigma^2 = \mu_2 - \mu_1^2$：

$$\hat{\mu} = \overline{X}, \qquad \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

对任何分布，期望与方差的矩估计都是这两个式子。
因为一阶矩就是期望、二阶中心矩就是方差，与分布类型无关。

盯住 $\hat{\sigma^2}$ 的分母：是 n，**不是** $n-1$。 也就说——**矩估计给出的方差估计是有偏的。**

矩估计例题：公交车多久发一班

题

某路公交车每隔 θ 分钟发一班（θ 未知）。乘客在任意时刻到站， 等待时间 $X \sim U(0, \theta)$。随机调查 n 位乘客，得样本 $X_1, \cdots, X_n$，估计 θ。

三步

1. **建立关系：** $E(X) = \dfrac{0+\theta}{2} = \dfrac{\theta}{2}$
2. **令总体矩 $=$ 样本矩：** $\dfrac{\theta}{2} = \overline{X}$
3. **解出：** $\hat{\theta} = 2\overline{X}$

直觉检验

平均等待时间应该是发车间隔的一半，所以间隔就是平均等待时间的两倍。 **结果与常识严丝合缝**——矩估计的朴素之处正在于此。

先记住这道题。slide 23 我们会用最大似然法重做一遍， 得到一个完全不同的答案。哪个对？——那时再说。

矩估计的两条性质

- **矩估计一定是一致估计。**因为样本矩依概率收敛于总体矩（大数定律），而 $\hat{\theta} = \theta(A_1, \cdots, A_k)$ 是连续函数。
- **矩估计不一定是无偏估计。** $\hat{\sigma^2} = \frac{1}{n}\sum(X_i - \overline{X})^2$ 就是现成的反例。

优点与代价

优点

- 不需要知道分布的具体形式，只要矩存在
- 计算简单，解方程即可

代价

- 只用到前几阶矩，**浪费了数据里的其余信息**，往往不够有效
- 可能给出荒唐的结果（下一页那道题里就有）

最大似然估计：谁最像，就选谁

基本思想

总体分布为 $f(x;\theta)$，现有观测值 $x_1,\cdots,x_n$。问： **θ 取什么值时，我手上这批数据最有可能出现？** 就认为那个 θ 最合理。

似然函数

$$L(x_1,\cdots,x_n;\theta)=\prod_{i=1}^{n} f(x_i;\theta)$$

其中 f 是概率函数（离散）或密度函数（连续）。

★ 似然不是概率

式子长得和联合密度一模一样，但**看的方向反了**： 联合密度里 θ 固定、x 是变量；**似然函数里 x 是已经拿到的数据、固定不动， θ 才是变量。**

所以 L 的图像横轴是参数，不是数据。**它不积分成 1，也不是概率。**

最大似然估计：定义与算法

定义

若 $L(x_1, \cdots, x_n; \theta)$ 在 $\hat{\theta}$ 处达到最大值， 则称 $\hat{\theta}$ 是 θ 的**最大似然估计**（MLE）。

实际计算：先取对数

$$\ln L = \sum_{i=1}^{n} \ln f(x_i; \theta)$$

连乘变成连加，求导容易得多。$\ln$ 单调递增，所以 $\ln L$ 与 L 在同一点取最大。

标准套路：$\dfrac{\partial \ln L}{\partial \theta} = 0$，解出 $\hat{\theta}$。 多个参数就解偏导方程组。

但求导不是万能的

只有当 L 光滑、最大值在内部取到时，求导才有效。 如果 L 在边界上单调、最大值取在端点（比如均匀分布）， **令导数为零会得不到解，甚至得到错误答案**——slide 23、27 各有一例。

交互：似然函数长什么样

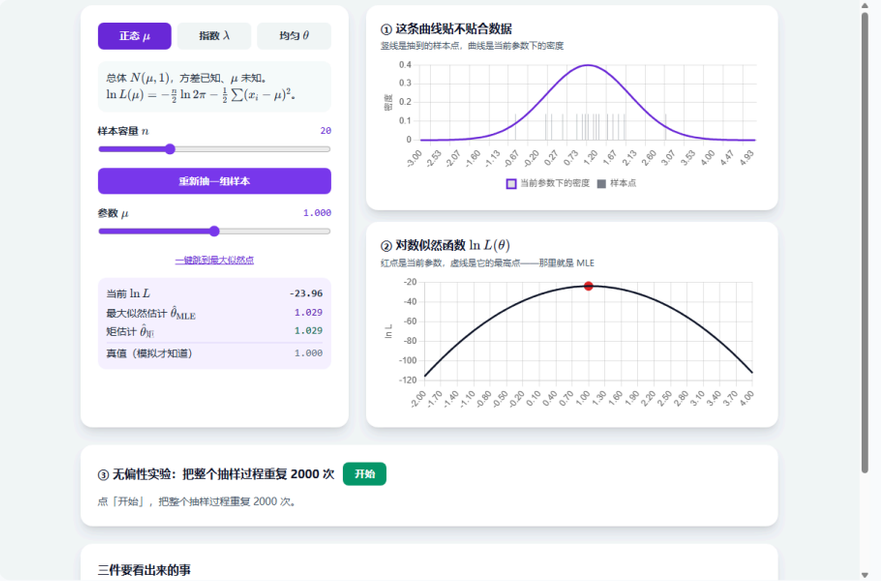
怎么用

上面一张图是**数据与当前参数下的密度**，下面一张是**对数似然函数**。 拖动参数滑块，两张图同时变。

现在做两件事

1. 在「正态 μ」页拖动 μ：**曲线贴合数据时，$\ln L$ 恰好最高** ——「最像」这个直觉是有图像的
2. 把 n 从 5 拖到 60：$\ln L$ **的峰越来越尖**。 峰越尖，参数越挪不动——**这就是「估得更准」的几何含义**

「均匀 θ」那一页留到 slide 24 再看。



交互演示：拖动参数，看数据与密度的贴合程度和对数似然曲线同步变化

https://fin-tech.fun/courseware/probability-statistics-undergrad/toolkit/mle_lab.html

MLE：指数分布与二项分布

指数分布（参数 λ）

$$\ln L = \sum_{i=1}^{n}(\ln\lambda - \lambda x_i) = n\ln\lambda - \lambda\sum_{i=1}^{n} x_i$$

$$\frac{\partial \ln L}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \implies \hat{\lambda} = \frac{1}{\overline{X}}$$

与矩估计完全一样。

二项分布（n 已知，p 未知，样本容量 k）

$$\ln L = \sum_{i=1}^{k} \ln C_n^{x_i} + \ln p \sum_{i=1}^{k} x_i + \ln(1-p)\left(kn - \sum_{i=1}^{k} x_i\right)$$

$$\frac{\partial \ln L}{\partial p} = \frac{\sum x_i}{p} - \frac{kn - \sum x_i}{1-p} = 0 \implies \hat{p} = \frac{\overline{X}}{n}$$

注意二项那一列：$\sum \ln C_n^{x_i}$ 这一项**不含 p**，求导直接消失。 **凡是与参数无关的因子，取对数后都不影响求导**——写似然时可以直接扔掉。

MLE：均匀分布——求导在这里失效

$U(a,b)$，a,b 均未知

参数必须满足 $a < \min x_i$ 且 $b > \max x_i$ ——**否则某个 x_i 落在区间外，$f(x_i)=0$，整个 $L=0$。** 在此条件下 $\ln L = -n\ln(b-a)$，它**没有驻点**：对 a 求导恒正、对 b 求导恒负。 要 $\ln L$ 最大就要 $b-a$ 最小，而它受上面那个约束限制：

$$\hat{a} = \min x_i, \qquad \hat{b} = \max x_i$$

★ 回到公交车那道题

$U(0,\theta)$ 的 MLE 是 $\hat{\theta} = \max x_i$： 「已经见过有人等了 19 分钟，那间隔至少 19 分钟」。 而矩估计给的是 $2\overline{X}$。**两个答案完全不同。**

而且 $\max x_i$ **永远小于真正的 θ**——它**系统性偏低**，是有偏估计。 可以证明 $E(\max X_i) = \dfrac{n}{n+1}\theta$，所以 $\dfrac{n+1}{n}\max x_i$ 才无偏。

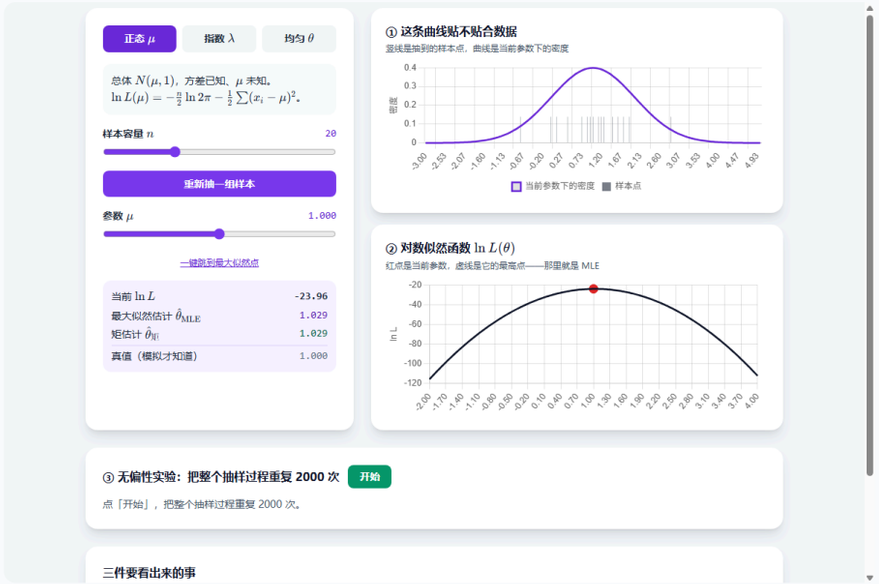
交互：无偏与有效，是两件事

切到「均匀 θ」页，做这个实验

1. 看②：对数似然在 $\theta < \max x_i$ 处**整段没有图像**（似然为 0），过了 $\max x_i$ 就**单调下降**——**最大值在端点**
2. 点③「开始」，把整个抽样过程重复 2000 次，看表

表里会看到什么

- 矩估计 $2\overline{X}$：平均值正中真值，**无偏**，但**散得开**
- MLE $\max x_i$：平均值**偏低**，且偏差不随重复次数消失；但**标准差小得多**
- 把 n 拖大再跑：偏差缩小——**它虽有偏，却是一致的**



交互演示：重复 2000 次抽样，对比 MLE 与矩估计的偏差和标准差

https://fin-tech.fun/courseware/probability-statistics-undergrad/toolkit/mle_lab.html

MLE：正态分布——回到谷神星

$N(\mu, \sigma^2)$，两个参数都未知

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

分别求偏导：

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2}\Big(\sum_{i=1}^{n} x_i - n\mu\Big) = 0, \qquad \frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

$$\hat{\mu} = \overline{X}, \qquad \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

★ 开场故事的答案

盯住第一个偏导：让它为零，等价于**让 $\sum(x_i - \mu)^2$ 最小**。 也就是说——**在正态误差下，最大似然估计就是「残差平方和最小」。**

这正是高斯给欧洲天文学界的答案：**「最贴近全部数据」 = 残差平方和最小**。 对最简单的情形（反复测同一个量），它给出的就是**算术平均**。 von Zach 按高斯的预报，在 1801 年 12 月 7 日重新找到了谷神星。

把这个想法推广到 $C = a + bY$ 这样的关系上，就是**最小二乘法**——计量经济学的第一课。

MLE 应用：训练一个最小的分类模型

场景：根据年龄预测是否购买

$Y = 1$ 表示购买。用**逻辑回归**描述：

$$P(Y = 1 \mid x) = \frac{1}{1 + e^{-\theta x}}$$

θ 就是要用数据估计的未知参数。

用 MLE 估 θ

收集到 3 个样本：(年龄 20, 不买)、(年龄 35, 买)、(年龄 50, 买)。似然函数就是 「**观测到这三个结果**」的概率：

$$L(\theta) = \left(1 - \frac{1}{1 + e^{-20\theta}}\right) \cdot \frac{1}{1 + e^{-35\theta}} \cdot \frac{1}{1 + e^{-50\theta}}$$

找 $\hat{\theta}$ 使 L 最大——**这个过程就叫「训练模型」。** 实践中对 $\ln L$ 求最大，并用梯度下降数值求解。

深度学习里那个「交叉熵损失函数」，**就是负对数似然。** 最小化损失 = 最大化似然。**本页这个式子和一个上亿参数的模型，用的是同一个原理。**

点估计例题

例 1

$f(x,\theta)=(\theta+1)x^{\theta}$，$0<x<1$，$\theta>-1$ 未知。 求 θ 的矩估计与最大似然估计。

矩估计： $\mu_1=EX=\int_0^1(\theta+1)x^{\theta+1}\mathrm{d}x=\dfrac{\theta+1}{\theta+2}$，

反解 $\theta=\dfrac{2\mu_1-1}{1-\mu_1}$：

$$\hat{\theta}=\frac{2\overline{X}-1}{1-\overline{X}}$$

MLE： $\ln L=n\ln(\theta+1)+\theta\sum\ln x_i$，令 $\dfrac{\mathrm{d}\ln L}{\mathrm{d}\theta}=0$：

$$\hat{\theta}=-\frac{n}{\sum_{i=1}^{n}\ln x_i}-1$$

例 2

$f(x,\theta)=2e^{-2(x-\theta)}$，$x>\theta$，$\theta>0$ 未知。求 θ 的 MLE。

$$\ln L=n\ln 2-2\sum_{i=1}^{n}x_i+2n\theta$$

它是 θ 的**单调递增函数**，令导数为零无解。 但约束是 $\theta\le\min x_i$（否则某个 x_i 落在支撑外，$L=0$），故

$$\hat{\theta}=\min x_i$$

又一次：最大值在端点。 凡是**参数出现在分布支撑的边界上**（$x>\theta$、$0<x<\theta$ 这类）， 就要先画出 $\ln L$ 随 θ 变化的样子，别急着求导。

小结：点估计

	矩估计	最大似然估计
思想	模型的矩 = 样本的矩	让手上这批数据最可能出现
做法	列方程组，反解，代入 A_k	写 $\ln L$，求导置零（或看端点）
要求	只要矩存在，不必知道分布形式	必须知道 $f(x;\theta)$ 的具体形式
性质	一定一致，不一定无偏	一般一致，常常有偏，但更有效

- **似然不是概率**：横轴是参数，不是数据
- **参数落在支撑边界上时，求导会失效**——$U(0,\theta)$、$x > \theta$ 都是
- 正态总体下 MLE 就是最小二乘，$\hat{\mu} = \overline{X}$

区间估计：为什么只报一个数不够

点估计的软肋

说「这只基金的月均回报率是 1.5%」，听起来很确定。 可这个 1.5% 是从 16 个月的数据里算出来的——**换 16 个月，它会变成别的数。** 只报一个数，等于把估计的不确定性藏了起来。

置信区间

找两个**统计量** $\underline{\theta}(X_1,\cdots,X_n)$、$\overline{\theta}(X_1,\cdots,X_n)$，使

$$P(\underline{\theta}<\theta<\overline{\theta})=1-\alpha$$

则称 $(\underline{\theta},\overline{\theta})$ 为**置信区间**，$1-\alpha$ 为**置信度**。常取 $\alpha=5\%$ 或 1%。

★ 这个概率说的是谁在动

θ 是**固定的常数**，不是随机变量； 随机的是**区间的两个端点**——它们由样本算出。

所以「95% 置信区间」的正确读法是：**按这个方法反复抽样、反复造区间， 其中约 95% 的区间会套住真值。**
不是「θ 有 95% 的概率落在这个区间里」。

怎么造一个置信区间：枢轴量

三步

1. 构造 $G(X_1, \cdots, X_n, \theta)$，使 G **的分布已知、且不依赖任何未知参数。** 这样的 G 称为**枢轴量**（Pivot）
2. 给定 $1-\alpha$，选常数 a, b 使 $P(a < G < b) = 1-\alpha$
3. 把 $a < G < b$ **变形**成 $\underline{\theta} < \theta < \overline{\theta}$

枢轴量从哪儿来？——上一讲那张表

整个下半场只做一件事：**把第 9 讲那四个统计量当枢轴量，逐个反解。**

$\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$	→ μ 的区间（σ 已知）
$\dfrac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$	→ μ 的区间（σ 未知）
$\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	→ σ^2 的区间
$\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$	→ σ_1^2/σ_2^2 的区间

期望的区间估计：σ 已知，正态总体

由 $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$，找临界值 $u_{\alpha/2}$ 使

$$P\left(\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right| < u_{\alpha/2}\right) = 1-\alpha$$

变形（把 μ 解到中间）得

$$P\left(\overline{X} - \frac{\sigma}{\sqrt{n}}u_{\alpha/2} < \mu < \overline{X} + \frac{\sigma}{\sqrt{n}}u_{\alpha/2}\right) = 1-\alpha$$

其中 $u_{\alpha/2}$ 由 $\Phi_0(u_{\alpha/2}) = 1-\alpha/2$ 查表：$\alpha = 5\%$ **时** $u_{\alpha/2} = 1.96$；$\alpha = 1\%$ **时** $u_{\alpha/2} = 2.58$。

例：灯泡寿命

随机抽 10 个灯泡，平均寿命 1200 小时。总体正态、方差为 8。求 $\alpha = 5\%$ 的置信区间。

$$\left(1200 \mp 1.96 \cdot \frac{\sqrt{8}}{\sqrt{10}}\right) = (1198.25,\ 1201.75)$$

大样本：不要求总体正态

n 很大时（方差已知或未知）

由中心极限定理，$\overline{X}$ 近似正态，故

$$\left(\overline{X}-\frac{\sigma}{\sqrt{n}}u_{\alpha/2},\ \overline{X}+\frac{\sigma}{\sqrt{n}}u_{\alpha/2}\right)$$

仍可用。**σ 未知时用 S 顶替即可**——n 大时这点差别可以忽略。

注意这里不要求总体正态，靠的是第 8 讲的中心极限定理。

案例：评估一只基金的月均回报率

月回报率 $\sim N(\mu,\sigma^2)$，已知波动率 $\sigma=2\%$。抽 $n=16$ 个月，样本均值 $\bar{x}=1.5\%$。 求 μ 的 95% 置信区间。

$z_{0.025}=1.96$，半宽 $=1.96\times\frac{2}{\sqrt{16}}=0.98$：

$$(1.5-0.98,\ 1.5+0.98)=(0.52\%,\ 2.48\%)$$

整个区间都在 0 以上——这才是分析师真正关心的结论： 有较强把握认为这只基金确实在赚钱。

如果区间是 $(-0.3\%,\ 3.3\%)$ 呢？点估计还是 1.5%， 但**「赚钱」这个结论就站不住了。这就是为什么必须报区间。**

期望的区间估计：σ 未知，小样本

正态总体、σ^2 未知，用 t 作枢轴量：

$$\left(\overline{X} \mp \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1)\right)$$

与上一个公式只差两处：$\sigma \to S$，$u \to t$。

例：灯泡（这次方差未知）

$n=10$，$\bar{x}=1200$，$S^2=8$，$t_{0.025}(9)=2.262$：

$$\left(1200 \mp 2.262 \cdot \frac{\sqrt{8}}{\sqrt{10}}\right) = (1198,\ 1202)$$

比 σ 已知时宽了——因为多估了一个 σ，代价就是区间变宽。

例：1 月最高温度

抽 4 天：10, 5, 2, 7。求 $\alpha=1\%$ 的置信区间，$t_{0.005}(3)=5.841$。

$\bar{x}=6$，$S^2=\frac{16+1+16+1}{3}=11.33$：

$$\left(6 \mp 5.841 \cdot \frac{\sqrt{11.33}}{\sqrt{4}}\right) = (-3.8,\ 15.8)$$

区间宽到几乎没有信息。4 个数据、要求 99% 的把握， 换来的就是这样一个结论。**数据太少时，诚实的报告本来就该这么宽。**

交互：95% 到底是什么意思

当堂要看清的一件事

反复抽样、反复造区间。**每个区间都不一样，真值那条竖线一动不动。**

1. 抽 100 次，数一数**有几条区间没套住真值**——大约 5 条
2. 把置信度调到 99%：区间变宽，漏掉的更少
3. 把样本量调大：区间变窄，漏掉的比例**不变**（还是 5%）

第 3 条是关键：**置信度管的是「漏掉的比例」，样本量管的是「区间的宽度」。 两件事互不干涉。**

交互演示：反复造区间，数一数有多少条没套住真值

https://fin-tech.fun/courseware/probability-statistics-undergrad/toolkit/confidence_interval.html

两个正态总体期望差的区间估计

情形一：大样本（方差已知或未知）

$$\left(\overline{X}-\overline{Y}\mp\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\,u_{\alpha/2}\right)$$

方差未知时用 S_1^2, S_2^2 顶替。**根号里是相加**——第 9 讲说过，差的方差也是相加。

情形二：小样本，且已知两总体方差相等（数值未知）

$$\left(\overline{X}-\overline{Y}\mp S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\,t_{\alpha/2}(n_1+n_2-2)\right)$$

其中 $S_p^2=\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ ——**就是 slide 8 那个合并方差。**

用在哪：**A/B 测试**。两版推荐算法，两组用户的人均点击数。 如果 $\mu_1-\mu_2$ 的置信区间**整个在 0 以上**，就有把握说版本一更好； **如果区间跨过 0，「更好」这个结论就下不了**——哪怕两组的样本均值确实差了一截。

方差的区间估计

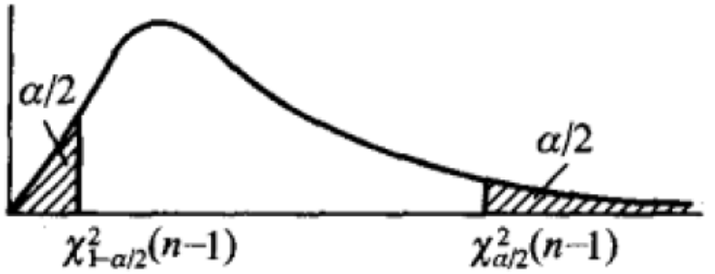
正态总体，用 $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ 作枢轴量：

$$P\left(a < \frac{(n-1)S^2}{\sigma^2} < b\right) = 1-\alpha$$

把 σ^2 解到中间要取倒数，不等号方向翻转，故

$$\left(\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a}\right)$$

取 $a = \chi^2_{1-\alpha/2}(n-1)$、$b = \chi^2_{\alpha/2}(n-1)$，其中 χ^2_α 由 $P(\chi^2 \geq \chi^2_\alpha) = \alpha$ 查表。



χ^2 分布临界值

与前面两个不同：χ^2 **不对称，所以两端的临界值不是一对相反数**，要分别查两次表。学生最常见的错误是只查一个然后加负号。

方差区间估计的两道例题

例：糖果的重量

抽 30 袋，样本标准差 10 克，总体正态。求 $\alpha = 5\%$ 时 σ^2 的置信区间。

$(n-1)S^2 = 29 \times 100 = 2900$，查表 $\chi^2_{0.025}(29) = 45.72$、$\chi^2_{0.975}(29) = 16.05$：

$$\left(\frac{2900}{45.72},\ \frac{2900}{16.05}\right) = (63.4,\ 180.7)$$

例：炮口速度

取 9 发炮弹试验，样本标准差 $S = 11$ m/s，总体正态。 求 $\alpha = 5\%$ 时**标准差** σ 的置信区间。 已知 $\chi^2_{0.025}(8) = 17.5$、$\chi^2_{0.975}(8) = 2.18$。

$(n-1)S^2 = 8 \times 121 = 968$：

$$\sigma^2 \in \left(\frac{968}{17.5},\ \frac{968}{2.18}\right) = (55.31,\ 444.04)$$

再对两个端点开平方：

$$\sigma \in (7.44,\ 21.07)$$

右题最后一步是**唯一的技术点**：要 σ 的区间，就把 σ^2 区间的两个端点各开平方。 这一步合法，是因为开平方是**严格单调**的变换，不改变「落在区间内」这件事。

区间从 7.44 跨到 21.07，**宽得惊人**——9 个数据估方差，就是这个精度。 **估方差比估均值要费数据得多。**

两个正态总体方差之比

期望未知的情形。由第 9 讲

$$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$$

得

$$P\left(F_{1-\alpha/2} < \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} < F_{\alpha/2}\right) = 1-\alpha$$

故 $\dfrac{\sigma_1^2}{\sigma_2^2}$ 的置信区间为

$$\left(\frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{\alpha/2}(n_1-1,n_2-1)},\ \frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{1-\alpha/2}(n_1-1,n_2-1)}\right)$$

怎么读这个区间

比较两支股票的风险时，看这个区间**是否包含 1**： 包含 1，就说不出谁的波动更大；整个在 1 以上，才有把握说第一支风险更高。

查表提示：多数表只给 $F_{\alpha/2}$，$F_{1-\alpha/2}(n_1,n_2) = 1/F_{\alpha/2}(n_2,n_1)$ ——**第 9 讲那条「倒过来自由度换位」的性质，就是为这里准备的。**

区间估计例题：一道题串起三个公式

题

从一批钉子中抽 16 枚，样本均值 2.125，样本标准差 0.017，长度服从正态分布，$\alpha = 0.1$。 求：(1) 已知 $\sigma = 0.01$ 时 μ 的置信区间；(2) σ 未知时 μ 的置信区间； (3) σ 未知时 σ^2 的置信区间。

已知 $u_{0.05} = 1.65$，$t_{0.05}(15) = 1.753$，$\chi^2_{0.05}(15) = 25$，$\chi^2_{0.95}(15) = 7.26$

(1) $\left(\overline{X} \mp u_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 2.125 \mp 1.65 \times \frac{0.01}{4}$

$= (2.121,\ 2.129)$

(2) $\left(\overline{X} \mp t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 2.125 \mp 1.753 \times \frac{0.017}{4}$

$= (2.117,\ 2.130)$

(3) $\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right) = \left(\frac{0.004335}{25}, \frac{0.004335}{7.26}\right)$

$= (0.00017,\ 0.000597)$

对比 (1) 与 (2)：**同一批数据，σ 未知时区间明显更宽** （半宽从 0.0041 变成 0.0074）。多估一个参数，精度就要付出代价—— $n = 16$ 时 t 比 u 大了 6%，而 S 又比 σ 大了 70%。

两个补充：比例的区间与单侧区间

$(0\text{–}1)$ 分布参数 p 的区间估计

$E(X)=p$，$D(X)=p(1-p)$。大样本下 $\overline{X}$ 近似 $N\big(p,\ p(1-p)/n\big)$，故

$$P\left(\frac{|\overline{X}-p|}{\sqrt{p(1-p)/n}}<u_{\alpha/2}\right)\approx 1-\alpha$$

难点在于 p 同时出现在分子和分母。两边平方整理成 $ap^2+bp+c<0$，其中

$$a=n+u_{\alpha/2}^2,\quad b=-(2n\overline{X}+u_{\alpha/2}^2),\quad c=n\overline{X}^2$$

解二次不等式得区间

$$\left(\frac{-b-\sqrt{b^2-4ac}}{2a},\ \frac{-b+\sqrt{b^2-4ac}}{2a}\right)。$$

单侧置信区间

若统计量 $\underline{\theta}$ 满足 $P(\theta>\underline{\theta})\geq 1-\alpha$，则 $(\underline{\theta},+\infty)$ 是 θ 的**单侧置信区间**（给出**置信下限**）。

若 $\overline{\theta}$ 满足 $P(\theta<\overline{\theta})\geq 1-\alpha$，则 $(-\infty,\overline{\theta})$ 给出**置信上限**。

什么时候用单侧？只关心一个方向的时候。 买设备只关心寿命的**下限**，控制污染只关心排放的**上限**。

把 α 全押在一边，同样的置信度下**这一侧的界限比双侧更紧** ——查表时用 u_α 而不是 $u_{\alpha/2}$。

本讲总结

① 回到 1801 年的那颗星

高斯没有去挑「最可靠的三次观测」。他把二十多次全用上，问了一个不同的问题：**参数取什么值时，我看到的这批数据最有可能出现？**

在正态误差下，这个问题的答案是**残差平方和最小**；对反复测量同一个量的情形，答案就是**算术平均**。von Zach 按这个答案，把丢了十个月的谷神星找了回来。

「最好的估计」不是一个可以靠直觉认定的东西——它需要一个准则。本讲给了三把尺子和两种造法。

本讲总结（续）

② 核心结论

- **点估计**——三把尺子：一致（底线）→ 无偏 → 有效，**互不蕴含**
- 矩估计：模型的矩 = 样本的矩。一定一致，不一定无偏
- MLE：让手上的数据最可能出现。写 $\ln L$，求导置零**或看端点**
- 正态总体：$\hat{\mu} = \overline{X}$，MLE 即最小二乘

- **区间估计**——四个枢轴量对应四种区间
- μ（σ 已知或大样本）：$\overline{X} \mp \frac{\sigma}{\sqrt{n}} u_{\alpha/2}$
- μ（σ 未知小样本）：$\overline{X} \mp \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1)$
- σ^2：$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}} \right)$
- σ_1^2/σ_2^2：用 F，注意查表要换自由度位置

③ 易错提醒

- **置信区间的随机性在端点，不在 θ**：95% 说的是「反复造区间，95% 会套住真值」
- 参数在**支撑边界**上时（$U(0,\theta)$、$x > \theta$），求导求 MLE 会失效
- χ^2 不对称，两端临界值**要查两次表**； 要 σ 就把 σ^2 区间的端点各开平方

④ 下一讲

区间估计已经在悄悄回答判断题了：基金的回报率区间**整个在 0 以上**， 我们就说它赚钱；**跨过 0**，就说不出口。

可这个界限到底划在哪？**下一讲：假设检验**——把这句话变成一套 有明确规则、也有明确犯错概率的判断程序，**并看清它会以哪两种方式出错。**