样本及抽样分布

第 9 讲 · 概率论与数理统计

复旦大学经济学院　ECON130001

上一讲回顾：样本均值的波动

独立同分布的观测值

若 $E(X_i)=\mu$ 有限，大数定律给出 $\overline{X}_n \xrightarrow{P} \mu$。

若另有 $0<\sigma^2=D(X_i)<\infty$，中心极限定理给出

$$\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1).$$

有限样本

在上述条件下，$E(\overline{X}_n)=\mu$，$D(\overline{X}_n)=\sigma^2/n$。

如果只抽取少量观测值，在哪些条件下还能求出 $\overline{X}_n$ 的精确分布？

同一总体抽两次，样本均值有哪些可能?

抽样 袋中有两张卡片，分别写着 2 和 4。每次等可能地抽一张，放回后混匀。抽两次，计算两个数的平均值。

结果 四种等可能的有序样本为 $(2,2)$、$(2,4)$、$(4,2)$、$(4,4)$，对应的均值依次为 2、3、3、4。

抽样分布 样本均值取 2、3、4 的概率分别为 $1/4$、$1/2$、$1/4$。每组样本给出一个均值，这些均值的可能取值和相应概率就构成了样本均值的抽样分布。

本讲学习目标

1. 区分总体、简单随机样本和统计量，说明重复抽样时哪些量会变化。
2. 用分组表、直方图和经验分布函数描述观测值。
3. 从正态样本构造 χ^2、t、F 分布，确定各自的自由度。

同一组日收益率可以回答哪些问题?

某科技公司股票的日收益率

为讨论抽样分布，设 n 个交易日的收益率相互独立，均服从 $N(\mu, \sigma^2)$。

本讲

画直方图可以观察各日收益率的分布。若重复抽取同样大小的样本，每次都算一次均值，所得 $\overline{X}$ 服从什么分布?

统计量在重复抽样下的概率分布称为**抽样分布**。

后两讲

用样本估计平均日收益率 μ 和方差 σ^2，检验数据是否支持 $\mu > 0$，也可检验正态分布假设。

总体与样本

研究对象与观测变量

研究对象的全体称为**总体**，其中每个单位称为**个体**。用随机变量 X 描述所研究的特征，记其分布函数为 F。

简单随机样本

若 $X_1, \ldots, X_n$ 相互独立，且每个 X_i 的分布函数都是 F，则称它们为来自该总体的**简单随机样本**。

n 为样本容量。一次抽样所得的具体数值记为 $x_1, \ldots, x_n$。

统计量与未知参数

定义

若样本的函数 $g(X_1, \ldots, X_n)$ 不含未知参数，则称其为**统计量**。例如，样本均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 可直接由数据计算。

哪些量可以由数据计算?

给定一个数 μ_0，可以计算 $\overline{X} - \mu_0$。若总体均值 μ 未知，则无法仅凭样本算出 $\overline{X} - \mu$。

推断时也会用到含未知参数的量。例如，在正态模型下，$(\overline{X} - \mu)/(\sigma/\sqrt{n})$ 的分布与参数值无关。

30 名学生的成绩如何分布?

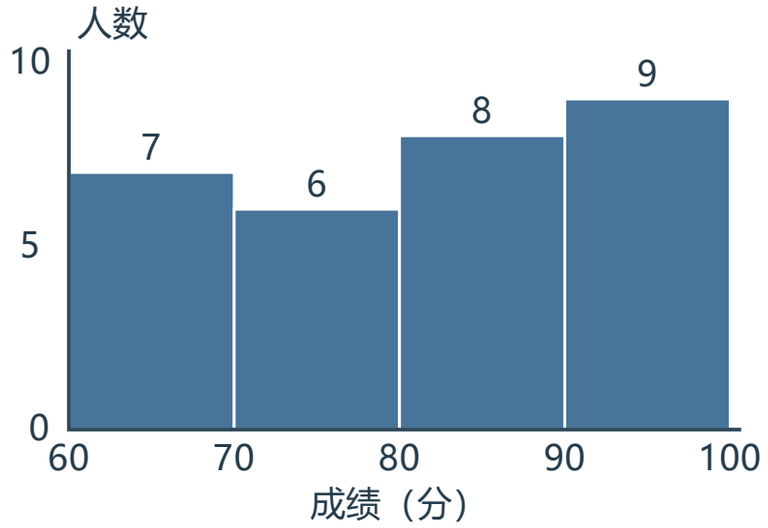
原始成绩

94 88 95 96 85 81 63 92 77 99 97 69 76 76 85 66 75 86 91 83 96 71 81 60 94 71 67 89 63 64

最低 60 分，最高 99 分。按 10 分一组统计人数。

分组统计

组限	组中值	频数	频率	累积频率
[60, 70)	65	7	23%	23%
[70, 80)	75	6	20%	43%
[80, 90)	85	8	27%	70%
[90, 100)	95	9	30%	100%



频数直方图

直方图的高度和面积

从频数到频率密度

频数直方图用柱高表示人数。若要用柱子的**面积表示该组频率**，高度应取

$$\text{频率密度} = \frac{\text{该组频数}}{\text{样本容量} \times \text{组距}}.$$

以 60 至 70 分这一组为例

7 人落在这一组，组距为 10 分。柱高为 $(7/30)/10 \approx 0.0233$，面积为 $7/30$。所有柱子的面积相加等于 1。

与理论分布的联系

对连续型总体，频率密度直方图可用来观察密度 f 的形状。把各组频率依次相加，则可观察累积比例随分数如何变化。

样本分布函数

不超过给定数值的比例

对观测值 $x_1, \ldots, x_n$，**样本分布函数**（经验分布函数）为

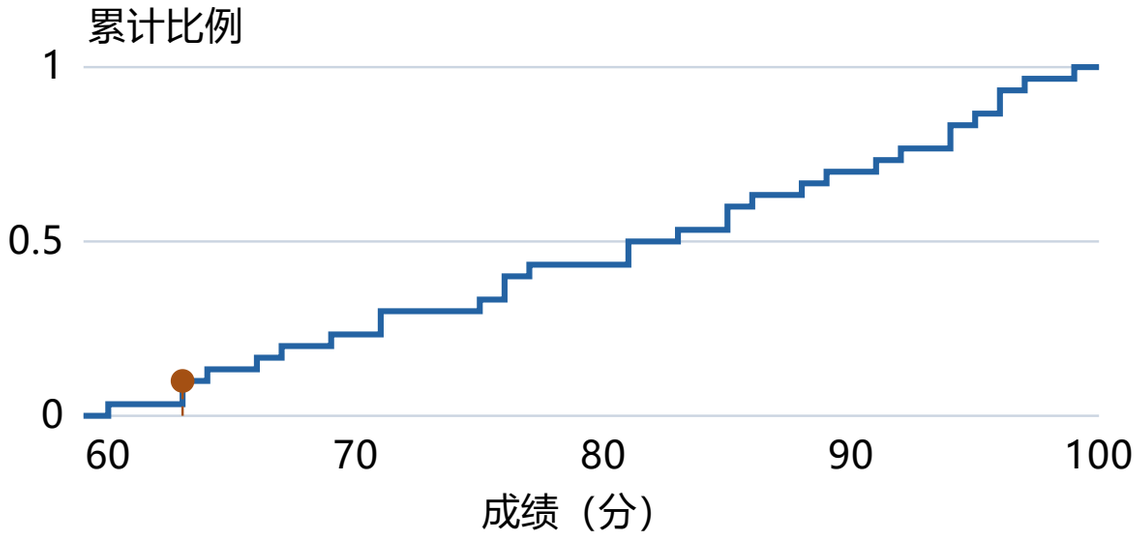
$$F_n(x) = \frac{\#\{i : x_i \le x\}}{n}.$$

成绩数据

有两名学生考了63分。阈值从62分移到63分时，累计比例从 $1/30$ 升至 $3/30$。

与总体分布函数的联系

对简单随机样本和固定的 x，由大数定律，$F_n(x) \xrightarrow{P} F(x)$。



不超过 63 分：3 人，累计比例 3/30

样本均值与样本方差

定义与计算（$n \geq 2$）

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

样本标准差为 $S = \sqrt{S^2}$。计算时可用 $(n-1)S^2 = \sum X_i^2 - n\overline{X}^2$。

分母为什么取 $n-1$？

离差之和 $\sum(X_i - \overline{X}) = 0$。给定前 $n-1$ 个离差，最后一个也随之确定，因此离差有 $n-1$ 个自由度。

对方差有限的简单随机样本，离差平方和的期望为 $(n-1)\sigma^2$，因此 $E(S^2) = \sigma^2$。

平移与缩放

若 $y_i = a + cx_i$，则 $\overline{y} = a + c\overline{x}$，$s_y^2 = c^2 s_x^2$。

小结：一次抽样与重复抽样

以成绩为例

- 一次抽样得到一组成绩 $x_1, \ldots, x_n$。直方图显示各分数组的人数，$F_n(x)$ 给出不超过 x 分的比例。
- 一次抽样得到的均值 $\bar{x}$ 是一个数。如果重复抽样，$\overline{X}$ 会随样本而变化。
- 下一步要研究的是重复抽样时 $\overline{X}$ 和 S^2 的分布。

正态总体下，样本均值服从什么分布?

独立正态变量的线性组合

若 $X_i \sim N(\mu_i, \sigma_i^2)$ 相互独立，则

$$\sum_{i=1}^{k} a_i X_i \sim N\left(\sum_{i=1}^{k} a_i \mu_i,\ \sum_{i=1}^{k} a_i^2 \sigma_i^2\right).$$

应用到简单随机样本

设总体为 $N(\mu, \sigma^2)$，$\sigma > 0$。令 $a_i = 1/n$，便有

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \qquad \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

在这个模型下，对每个样本容量 n 都得到精确的正态分布。

Γ 函数与 Γ 分布

从阶乘到 Γ 函数

$$\Gamma(\alpha)=\int_0^{\infty} x^{\alpha-1}e^{-x}\,\mathrm{d}x,\qquad \alpha>0.$$

$\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$，故对正整数 n 有 $\Gamma(n)=(n-1)!$。另有 $\Gamma(1/2)=\sqrt{\pi}$。

形状参数 $\alpha>0$，率参数 $\beta>0$

$$X\sim\Gamma(\alpha,\beta):\quad f(x)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}\quad(x>0).$$

$E(X)=\alpha/\beta$，$D(X)=\alpha/\beta^2$。当 $\alpha=1$ 时得到率为 β 的指数分布。

可加性

若 $X_i\sim\Gamma(\alpha_i,\beta)$ 相互独立且率参数相同，则 $\sum X_i\sim\Gamma(\sum\alpha_i,\beta)$。

χ^2 分布：独立标准正态变量的平方和

一个标准正态变量的平方

若 $Z \sim N(0,1)$，则 $Z^2 \sim \Gamma(1/2, 1/2)$。

由分布函数求密度

对 $y > 0$，$P(Z^2 \le y) = 2\Phi(\sqrt{y}) - 1$。求导得 $f(y) = e^{-y/2}/\sqrt{2\pi y}$，即 $\Gamma(1/2, 1/2)$ 的密度。

k 个独立平方项

若 $Z_1, \ldots, Z_k$ 相互独立且均服从 $N(0,1)$，由 Γ 分布的可加性，

$$Q = \sum_{i=1}^{k} Z_i^2 \sim \Gamma(k/2, 1/2) = \chi^2(k).$$

这是**自由度为 k 的卡方分布**。它取非负值，$E(Q) = k$，$D(Q) = 2k$。

样本均值与样本方差为什么独立?

正态总体的简单随机样本

若 $X_i \sim N(\mu, \sigma^2)$ 相互独立，$n \geq 2$，$\sigma > 0$，则

$$\overline{X} \text{ 与 } S^2 \text{ 独立}, \qquad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

把标准化样本分成两个正交部分

令 $Z_i = (X_i - \mu)/\sigma$。在 n 维空间中，以 $(1, \ldots, 1)/\sqrt{n}$ 为第一条坐标轴，补成一组正交单位轴。

变换后的坐标仍为相互独立的标准正态变量。第一坐标是 $\sqrt{n}(\overline{X} - \mu)/\sigma$。

其余 $n-1$ 个坐标的平方和等于 $\sum_i (X_i - \overline{X})^2/\sigma^2$，所以它与第一坐标独立，且服从 $\chi^2(n-1)$。

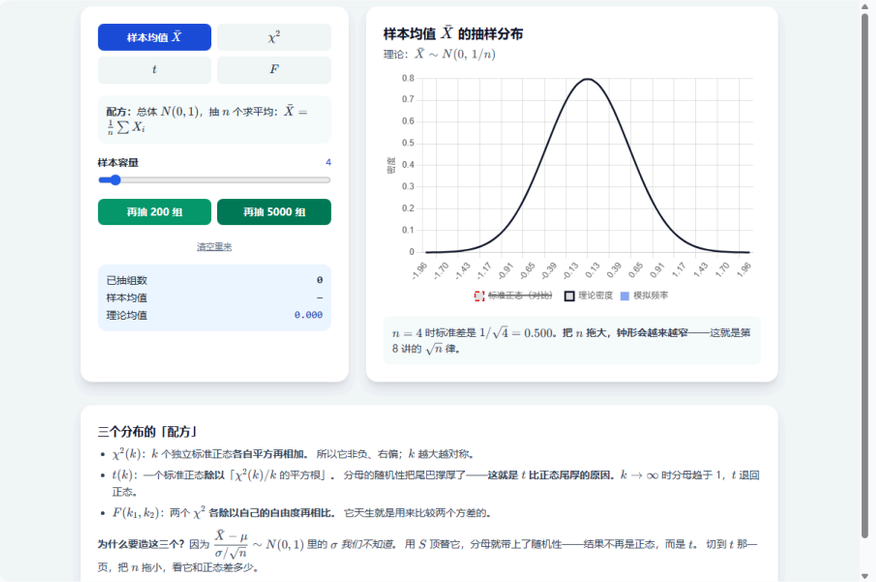
模拟：样本均值与 χ^2 分布

样本均值

从 $N(0,1)$ 重复抽样，比较 $n=4$ 和 $n=16$。两种情形下，均值都在 0 附近，标准差从 $1/2$ 减为 $1/4$。请对照横轴刻度。

χ^2 分布

在卡方页将自由度从 2 调到 20，点击「再抽 5000 组」。均值向右移动，右偏程度如何变化？



模拟：比较不同样本容量下的均值分布与不同自由度的卡方分布

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

用 S 代替 σ 后的分布

标准误也要估计

正态总体的标准差未知时，用 $S/\sqrt{n}$ 估计 $\overline{X}$ 的标准误。分母随样本变化，需要重新确定整个比值的分布。

t 分布的构造

设 $Z \sim N(0,1)$，$V \sim \chi^2(k)$，且 Z 与 V 独立，则

$$T = \frac{Z}{\sqrt{V/k}} \sim t(k).$$

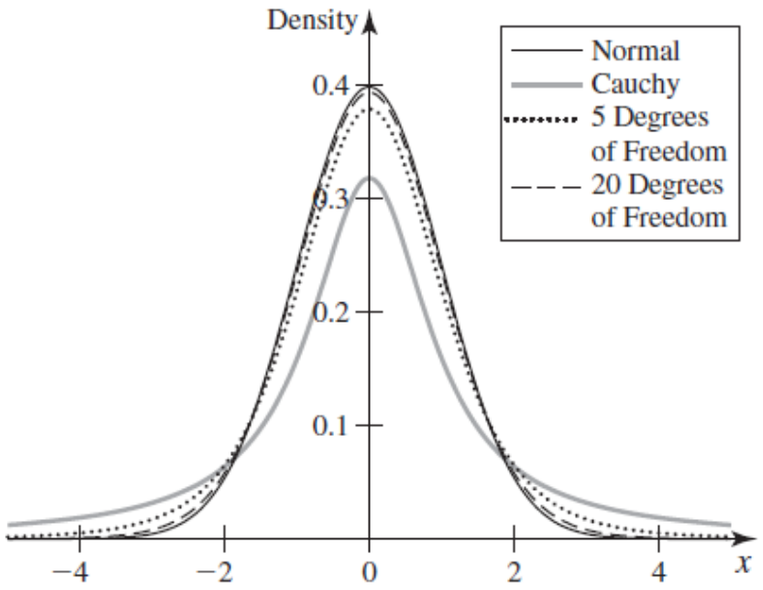
自由度为 k 的 t 密度

$$f(x) = \frac{\Gamma((k+1)/2)}{\sqrt{k\pi}\,\Gamma(k/2)}(1 + x^2/k)^{-(k+1)/2}, \quad x \in \mathbb{R}.$$

尾部形状

t 分布关于 0 对称。分母有时很小，使得 $|T|$ 可能很大，尾部比标准正态分布厚。

正态样本均值的 t 分布



自由度越大越接近正态

单样本结果

对 $N(\mu, \sigma^2)$ 的简单随机样本，$n \geq 2$，

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

为什么是 $n-1$ 个自由度?

令 $Z = \sqrt{n}(\overline{X} - \mu)/\sigma$，$V = (n-1)S^2/\sigma^2$。第 16 页说明 Z 与 V 独立。将它们代入 $T = Z/\sqrt{V/(n-1)}$ 即得结论。

如何使用

检验给定的均值 μ_0 时，用 μ_0 代入就能从数据计算检验统计量。

随着自由度增大，t 分布的分位数趋近于标准正态分位数。

两组均值之差的 t 分布

共同方差与合并估计

从两个正态总体分别独立抽取简单随机样本，容量 $n_1, n_2 \geq 2$，总体均值为 μ_1, μ_2，方差均为 σ^2。按自由度合并两组样本方差：

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

用合并标准差估计均值差的标准误

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t(n_1 + n_2 - 2).$$

两组的离差自由度分别为 $n_1 - 1$ 和 $n_2 - 1$。

应用

比较两个推荐版本下用户的平均使用时长时，若上述模型适用，可用此式检验 $\mu_1 = \mu_2$。

F 分布与两个样本方差的比

两个独立卡方变量各除以自由度后的比

若 $U \sim \chi^2(k_1)$、$V \sim \chi^2(k_2)$ 相互独立，则

$$W = \frac{U/k_1}{V/k_2} \sim F(k_1, k_2), \qquad \frac{1}{W} \sim F(k_2, k_1).$$

两个独立正态样本

设两组容量分别为 $n_1, n_2 \geq 2$，总体方差分别为 $\sigma_1^2, \sigma_2^2 > 0$。代入两组独立的标准化离差平方和，得

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

比较收益率波动

若两组股票收益率满足上述假设，且两总体方差相等，则 S_1^2/S_2^2 本身服从这个 F 分布。

模拟：t 分布的尾部与 F 分布的形状

比较 t 与标准正态

先观察自由度为 2 的 t 分布。比较密度曲线与红色虚线在中心和尾部的高度。

再将自由度调到 40，观察两条曲线如何接近。

比较 $F(2,2)$ 与 $F(40,40)$

在模拟器中分别选择 $F(2,2)$ 和 $F(40,40)$，哪一条曲线在 1 附近更集中?



模拟：比较不同自由度的 t 分布和 F 分布

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

例题（一）：样本均值的概率计算

单个样本均值

从 $N(52, 6.3^2)$ 独立抽取 36 个观测值。均值落在 50.8 至 53.8 之间的概率是多少？

标准误为 $6.3/\sqrt{36} = 1.05$。令 Φ 为标准正态分布函数。

$$\begin{aligned} P &= \Phi(1.8/1.05) \\ &\quad - \Phi(-1.2/1.05) \\ &\approx 0.8302. \end{aligned}$$

两个样本均值之差

两个独立样本均来自 $N(20, 3)$，容量分别为 10、15。均值之差 $\Delta = \overline{X} - \overline{Y}$ 的绝对值超过 0.3 的概率是多少？

$\Delta \sim N(0, 3/10 + 3/15) = N(0, 0.5)$。

$$\begin{aligned} P(|\Delta| > 0.3) &= 2\Phi(-0.3/\sqrt{0.5}) \\ &\approx 0.6714. \end{aligned}$$

例题（二）：构造卡方变量与 t 变量

设 $X_1, \ldots, X_6$ 相互独立且均服从 $N(0,1)$

（1）令 $Y = (X_1 + X_2 + X_3)^2 + (X_4 + X_5 + X_6)^2$。求 C 使 CY 服从卡方分布。

（2）令 $T = C(X_1 + X_2)/\sqrt{X_3^2 + X_4^2 + X_5^2}$。求正数 C 使 T 服从 t 分布。

先辨认相互独立的部分

（1）两个三项和独立且均服从 $N(0,3)$。各除以 $\sqrt{3}$ 后再平方相加，得 $C = 1/3$，$CY \sim \chi^2(2)$。

（2）用 $(X_1 + X_2)/\sqrt{2}$ 作分子，用 $\sqrt{(X_3^2 + X_4^2 + X_5^2)/3}$ 作分母，得 $C = \sqrt{3/2}$，$T \sim t(3)$。

用卡方分布计算方差的概率

从正态总体独立抽取 16 个观测值。令 $V = 15S^2/\sigma^2 \sim \chi^2(15)$，则 $P(S^2/\sigma^2 \le 2.041) = P(V \le 30.615) \approx 0.9901$。

课后练习：样本方差的抽样分布

仍用第3页的两张卡片

独立、有放回地抽两次，分别记为 X_1, X_2。求样本方差 S^2 的可能取值及其概率。

按两个观测值是否相同分类

当 $n=2$ 时，$S^2=(X_1-\overline{X})^2+(X_2-\overline{X})^2$。

两次抽到相同的数时，两个离差均为0，故 $S^2=0$。抽到不同的数时，均值为3，两个离差为 -1 和 1，故 $S^2=2$。

$$P(S^2=0)=\tfrac{1}{2},\qquad P(S^2=2)=\tfrac{1}{2}.$$

正态样本的抽样分布

单样本来自正态总体，两样本情形还要求两组独立

标准化后的量	精确分布	推断对象
$\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	$N(0,1)$	均值（σ 已知）
$\dfrac{\overline{X}-\mu}{S/\sqrt{n}}$	$t(n-1)$	均值（用 S 估计 σ）
$\dfrac{(n-1)S^2}{\sigma^2}$	$\chi^2(n-1)$	方差
$\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$	$F(n_1-1,n_2-1)$	两个方差之比

下一讲：从抽样分布到参数估计

日收益率的均值

观测到一组日收益率后，可以用样本均值 $\bar{x}$ 估计平均日收益率 μ。

若总体方差未知，能否利用 $(\overline{X}-\mu)/(S/\sqrt{n})$ 的抽样分布，给出 μ 的置信区间？