高级计量经济学

Appendix 2: Review of Statistics

Hansen, B. (2022). Probability & Statistics for Economists. Princeton University Press.

黄嘉平

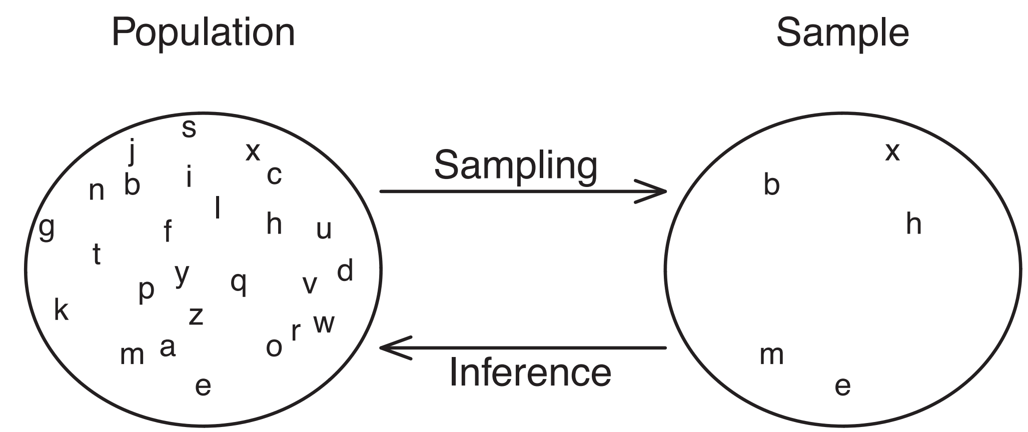
工学博士 经济学博士
深圳大学中国经济特区研究中心 讲师

办公室 粤海校区汇文楼1510
E-mail huangjp@szu.edu.cn
Website https://huangjp.com

统计学中的基础概念

随机抽样

Random sampling



- 随机变量的向量称为**随机向量（random vector）**，即 $Y = (Y_1, Y_2, \ldots, Y_k)$，其中 Y_j 为随机变量

随机向量的集合 $\{X_1, X_2, \ldots, X_n\}$ 为**独立同分布（independent and identically distributed/i.i.d.）**是指这些随机向量相互独立且服从同一个分布 F

独立同分布的随机向量集合 $\{X_1, X_2, \ldots, X_n\}$ 称为**随机样本（random sample）**

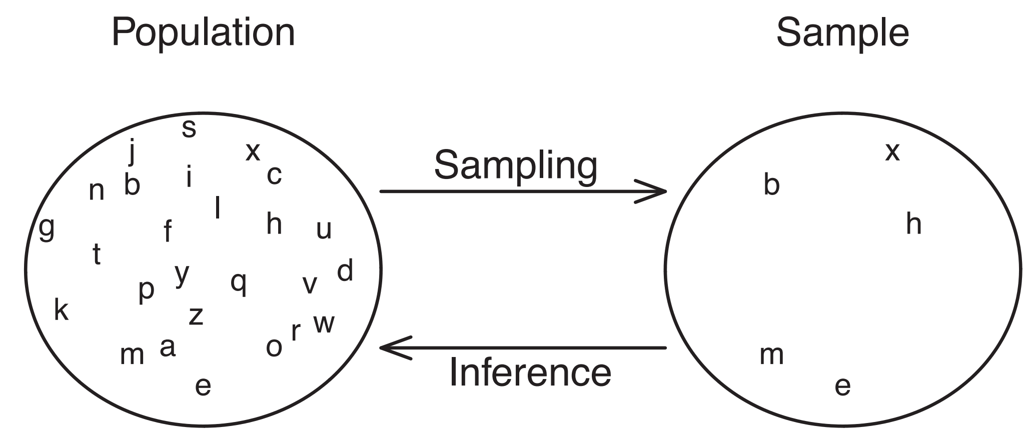
- F 称为**总体分布（population distribution）**或**总体（population）**

- 从总体中获得随机样本的过程称为**随机抽样（random sampling）**。随机抽样有两种解释：

  1. 全体观测对象切实存在。当总体中包含 N 个个体（N 可以是有限或无限）是，随机抽样是从其中随机选择 n 个个体的操作。例如问卷调查或普查　“随机”代表每个个体被选中的概率相等
  2. 全体观测对象不存在。此时通过特定的**数据生成过程（data generating process/DGP）**产生 n 个结果的操作。例如实验或观测性研究

样本与推断

Sample and inference



样本（sample）是随机抽样的结果

- 我们通常用到的**数据集（dataset）**可以看做是某个总体的样本，数据集中的每一条数据都是一个**观测值（observation）**
- 当样本中有 n 个观测值时，n 被称为**样本量（sample size）**

通过样本推测总体或数据生成过程的特征的行为称为**推断（inference）**

- 采用哪种推断方法取决于数据是如何获得的，即抽样的方法。我们主要关注随机抽样
- 随机抽样以外还有分层抽样（stratified sampling）、整群抽样（clustered sampling）、面板数据（panel data）、时间序列数据（time series data）、空间数据（spatial data）等

真实数据的例子

Current Population Survey

- Current Population Survey (CPS) 是美国政府主导的关于就业的普查
- CPS 的样本量为60000个家庭
- CPS 采用两阶段分层抽样法
  1. 将全国分成 852 个地区，并按照所在地和人口特征等对这些地区进行分层，在每层中随机抽选一个地区
  2. 以地区为单位，在地区内部随机抽选对象家庭
- 右侧的表中包含2009年3月份普查中，已婚黑人女性且工作经验超过12年的子样本。Wage 为时薪（年薪/工作时长），Education 为受教育年限

Table 6.1: Observations From CPS Data Set

Observation	Wage	Education
1	37.93	18
2	40.87	18
3	14.18	13
4	16.83	16
5	33.17	16
6	29.81	18
7	54.62	16
8	43.08	18
9	14.42	12
10	14.90	16
11	21.63	18
12	11.09	16
13	10.00	13
14	31.73	14
15	11.06	12
16	18.75	16
17	27.35	14
18	24.04	16
19	36.06	18
20	23.08	16

Hansen, B. E. *Probability & Statistics for Economists*. PUP

统计量、参数、估计量

Statistics, parameters, and estimators

总体 F 的任意函数 θ 称为**参数（parameter）**

- 例如，总体均值 $\mu = E[X]$ 是总体分布 F 的一阶矩

样本 $\{X_1, \ldots, X_n\}$ 的函数称为**统计量（statistic）**

- 例如，样本均值 $\bar{X} = \frac{1}{n}\sum_i^n X_i$ 是一个统计量

如果我们想用某一统计量来猜测参数 θ，则称这个统计量是 θ 的**估计量（estimator）**，记作 $\hat{\theta}$

- 例如，样本均值 $\bar{X}$ 是总体均值 μ 的一个估计量，可以写成 $\hat{\mu} = \bar{X}$
- 当样本已经确定时，估计量 $\hat{\theta}$ 的取值被称为**估计值（estimate）**

抽样分布

Sampling distribution

- 统计量是随机变量的函数，因此也是随机变量，并服从其自身的概率分布

统计量的概率分布称为为**抽样分布（sampling distribution）**

- 估计量 $\hat{\theta}$ 的抽样分布可以帮助我们了解更多关于参数 θ 的信息
- 以样本均值 $\bar{X}$ 为例，我们可以从以下角度了解 $\bar{X}$ 的抽样分布
  1. 偏差和方差
  2. 当总体分布是正态分布时，$\bar{X}$ 的分布
  3. 大样本下（$n \to \infty$）的分布
  4. 分布的渐进展开
  5. Bootstrap 近似

总体均值：$E[X_i] = \mu$
总体方差：$\mathrm{Var}[X_i] = E[(X_i - \mu)^2] = \sigma^2$
$\sigma = \sqrt{\sigma^2}$ 称为总体标准差（standard deviation）

估计偏差

Estimation bias

估计量 $\hat{\theta}$ 的**偏差（bias）**定义为 $\mathrm{bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$

- 如果偏差为 0，则称该估计量为**非偏（unbiased）**
- 在不同的总体分布 F 下，同一估计量可能为非偏也可能有偏
- $E[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$，因此当 $\mu < \infty$ 时，$\bar{X}$ 为 μ 的非偏估计量
- μ 的其他非偏估计量包括 $X_1, \dfrac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$ 等
- $\hat{\theta}$ 是 θ 的非偏估计量 $\Rightarrow$ $\hat{\beta} = a\hat{\theta} + b$ 是 $\beta = a\theta + b$ 的非偏估计量
- 如果 h 是非线性函数，则 $\hat{\beta} = h(\hat{\theta})$ 不一定是 $\beta = h(\theta)$ 的非偏估计量

估计方差

Estimation variance

估计量 $\hat{\theta}$ 的方差 $\mathrm{Var}[\hat{\theta}] = E\left[(\hat{\theta} - E[\hat{\theta}])^2\right]$ 称为**抽样方差（sampling variance）**

- $\bar{X}$ 的抽样方差：
  - $X_1, \ldots, X_n$ 为独立同分布 $\Rightarrow$ $X_1, \ldots, X_n$ 两两不相关
  - $X_1, \ldots, X_n$ 两两不相关 $\Rightarrow$ $\mathrm{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathrm{Var}[X_i]$ （尝试证明这个命题）
  - $\mathrm{Var}[\bar{X}] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}[X_i] = \frac{\sigma^2}{n}$ $(\mathrm{Var}[X_i] = \sigma^2 < \infty)$

- 如果 $\hat{\beta} = a\hat{\theta} + b$，则 $\mathrm{Var}[\hat{\beta}] = a^2\mathrm{Var}[\hat{\theta}]$

- 一般情况下，我们无法获得非线性变换 $h(\bar{X})$ 的方差的准确表达

均方误差
Mean squared error (MSE)

- 均方误差或均方误是衡量估计准确度的常用标准

估计量 $\hat{\theta}$ 的均方误差（mean squared error）定义为

$$\mathrm{MSE}\left[\hat{\theta}\right] = E\left[(\hat{\theta} - \theta)^2\right]$$

- 将定义展开可得

$$\begin{aligned}\mathrm{MSE}\left[\hat{\theta}\right] &= E\left[(\hat{\theta} - \theta)^2\right] \\ &= E\left[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)^2\right] \\ &= E\left[(\hat{\theta} - E[\hat{\theta}])^2\right] + 2E\left[\hat{\theta} - E[\hat{\theta}]\right]\left(E[\hat{\theta}] - \theta\right) + \left(E[\hat{\theta}] - \theta\right)^2\end{aligned}$$

右侧第一项为 $\mathrm{Var}[\hat{\theta}]$，第二项为零，第三项为 $(\mathrm{bias}[\hat{\theta}])^2$，因此

当 $\mathrm{Var}[\hat{\theta}] < \infty$ 时，$\mathrm{MSE}\,[\hat{\theta}] = \mathrm{Var}[\hat{\theta}] + \left(\mathrm{bias}[\hat{\theta}]\right)^2$

- 我们倾向于选择非偏且方差小的估计量，但两者往往无法兼得。此时需要根据研究需要进行取舍

对总体方差的估计

Estimation of variance

- 如果已知总体均值 μ，则 $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$ 是 σ^2 的非偏估计量

$$E[\tilde{\sigma}^2] = \frac{1}{n}\sum_{i=1}^{n}E[(X_i-\mu)^2] = \frac{1}{n}\sum_{i=1}^{n}\sigma^2 = \sigma^2$$

- 如果 μ 为未知，那么 $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$ 是 σ^2 的非偏估计量吗?

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu+\mu-\bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 + \frac{1}{n}\sum_{i=1}^{n}2(X_i-\mu)(\mu-\bar{X}) + \frac{1}{n}\sum_{i=1}^{n}(\mu-\bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 + 2(\bar{X}-\mu)(\mu-\bar{X}) + (\mu-\bar{X})^2 \\
&= \tilde{\sigma}^2 - (\bar{X}-\mu)^2
\end{aligned}$$

因此，$E[\hat{\sigma}^2] = \sigma^2 - \frac{\sigma^2}{n} = (1-\frac{1}{n})\sigma^2$

$\Rightarrow\ s^2 = \frac{n}{n-1}\hat{\sigma}^2$ 是 σ^2 的非偏估计量

s^2 通常称为样本方差（sample variance）。
也可将 $\hat{\sigma}^2$ 称为样本方差，此时 s^2 称为偏差修正样本方差（bias-corrected sample variance）。
注意：s 不是 σ 的非偏估计量

渐进分析

极限

Limit

- 样本量越大，样本中包含的信息越接近总体所包含的信息，因此我们希望获得 $n \to \infty$ 时估计量的极限特征

数列的极限：如果对于任意 $\varepsilon > 0$，存在 $n_\varepsilon < \infty$，使 $|a_n - a| \le \varepsilon$ 针对所有的 $n > n_\varepsilon$ 成立，则称 a 为数列 a_n 的**极限（limit）**，写作 $\lim_{n\to\infty} a_n = a$。这时我们说 a_n 在 $n \to \infty$ 时**收敛（converges）**于 a

随机变量的极限之概率极限：如果对于任意 $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr(|Z_n - c| \le \varepsilon) = 1$$

或等价的 $\lim_{n\to\infty} \Pr(|Z_n - c| > \varepsilon) = 0$

则称 c 为随机变量的序列 Z_n 的**概率极限（probability limit）**，写作 $\operatorname{plim}_{n\to\infty} Z_n = c$ 或 $Z_n \xrightarrow{p} c$。这时我们说 Z_n 在 $n \to \infty$ 时**依概率收敛（converges in probability）**于 c

大数定律

Law of large numbers (LLN)

弱大数定律（weak law of large nubmers, WLLN）：如果 X_i 为 i.i.d. 且 $E[X_i] < \infty$，则当 $n \to \infty$ 时

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} E[X_i]$$

即样本均值依概率收敛于总体均值

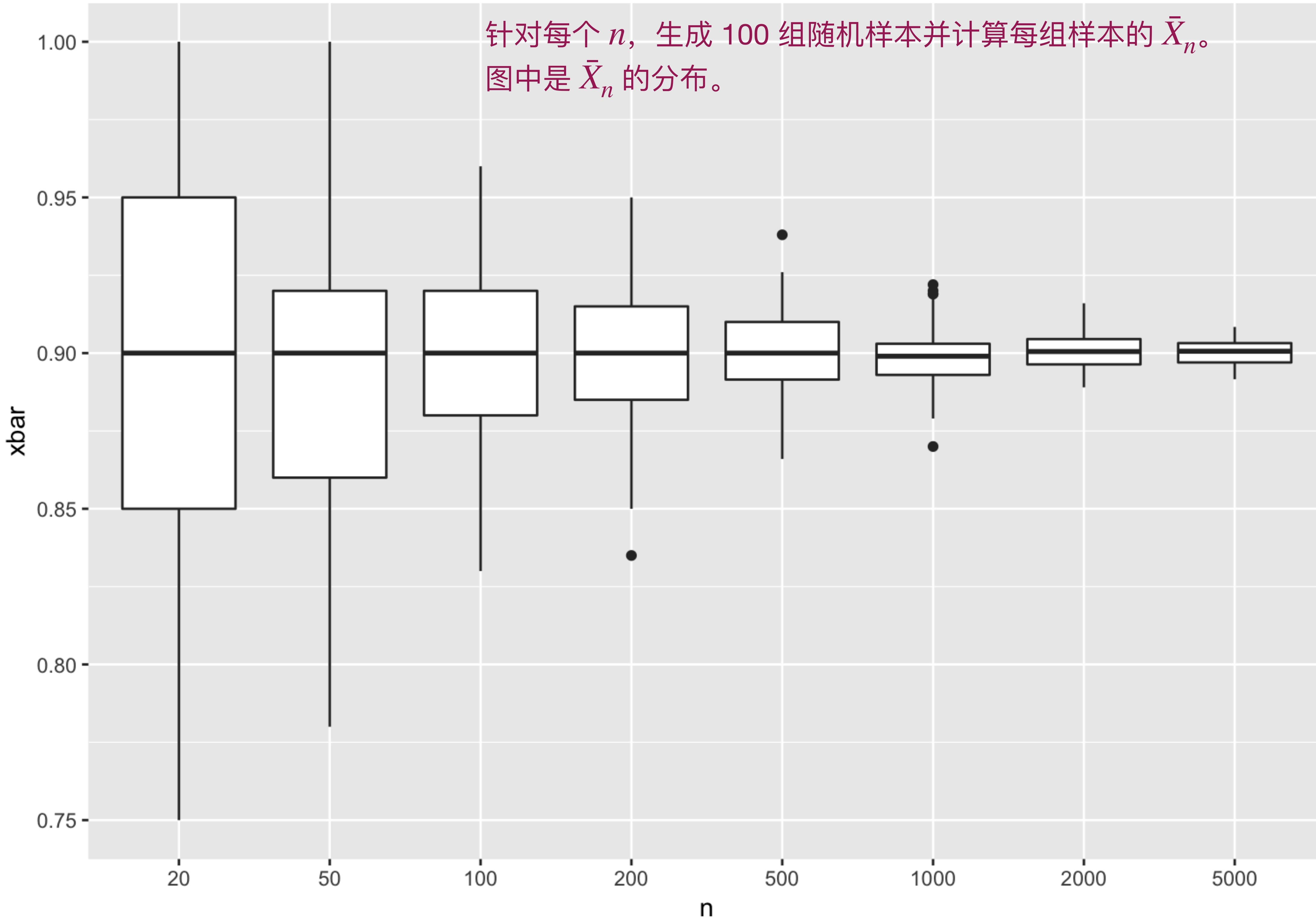
注：i.i.d. 条件可替换为独立样本且总体方差为有限

如果估计量 $\hat{\theta}$ 在 $n \to \infty$ 时依概率收敛于 θ（即 $\hat{\theta} \xrightarrow{p} \theta$），则称 $\hat{\theta}$ 是 θ 的**一致（consistent）**估计量

- 由大数定律可知，样本均值是总体均值的一致估计量

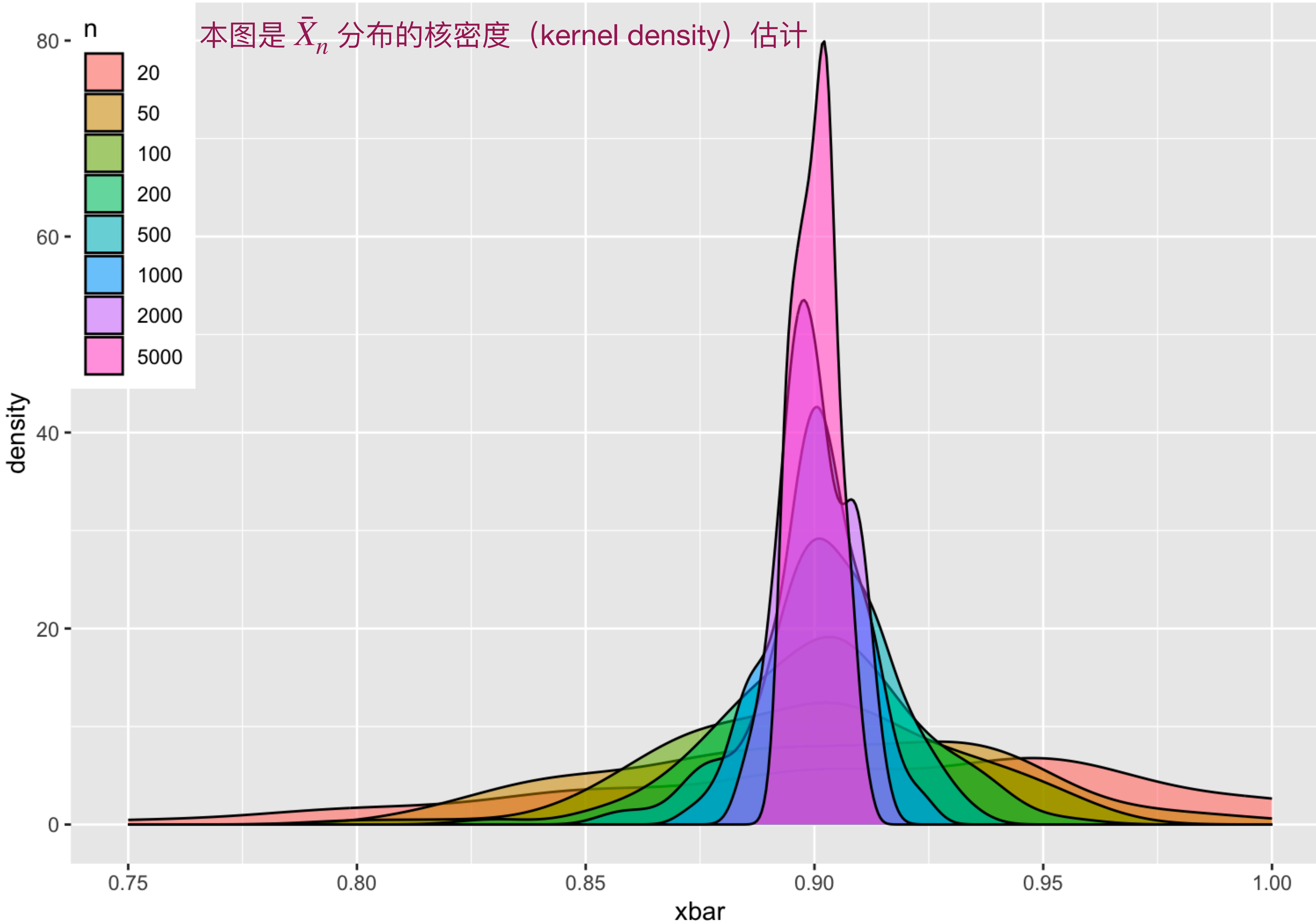
模拟大数定律

$X_i \sim \text{Bernoulli}(p = 0.9)$, $n = 20,50,100,200,500,1000,2000,5000$



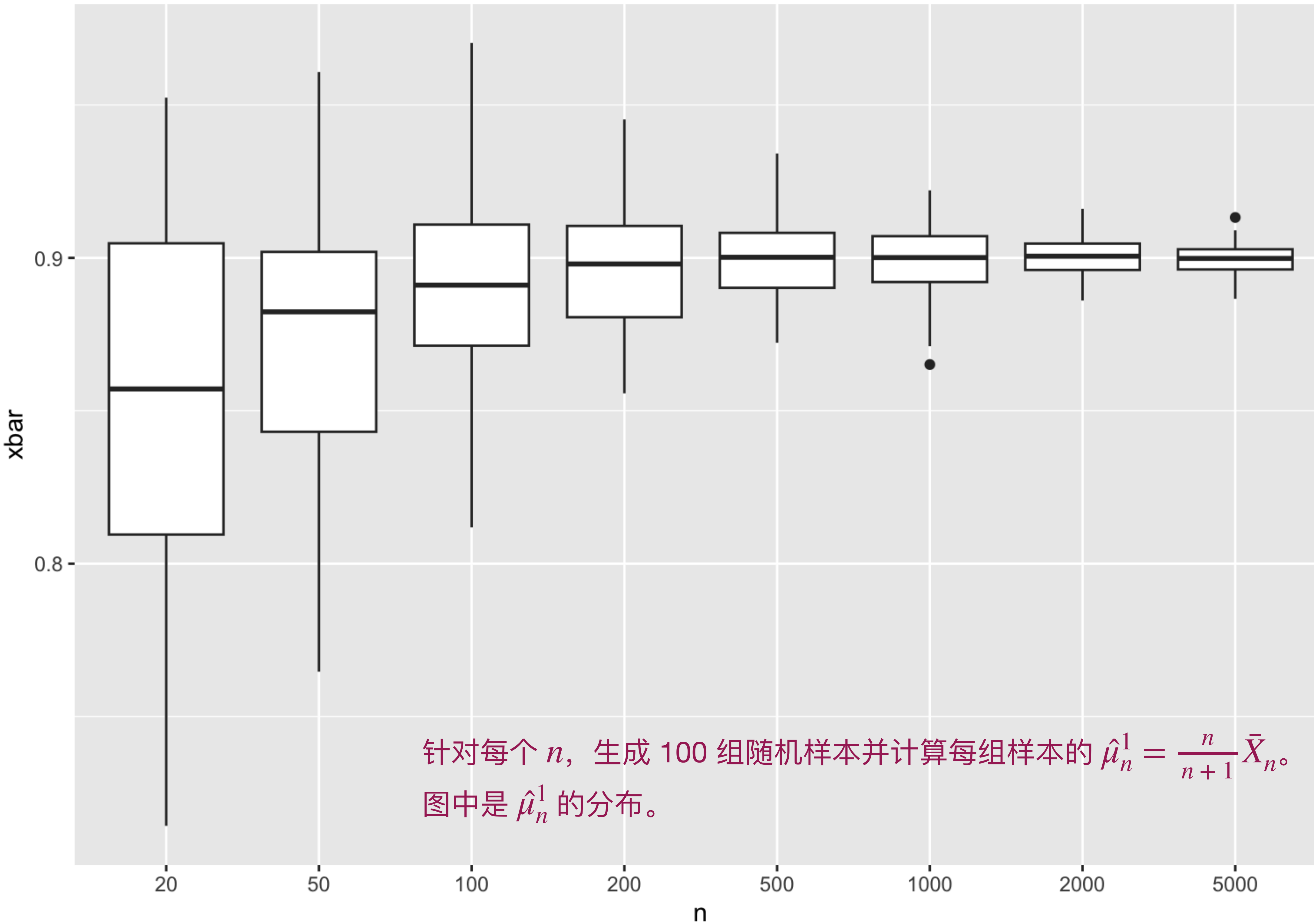
模拟大数定律

$X_i \sim \text{Bernoulli}(p = 0.9)$, $n = 20,50,100,200,500,1000,2000,5000$



模拟大数定律

$X_i \sim \text{Bernoulli}(p = 0.9)$, $n = 20,50,100,200,500,1000,2000,5000$

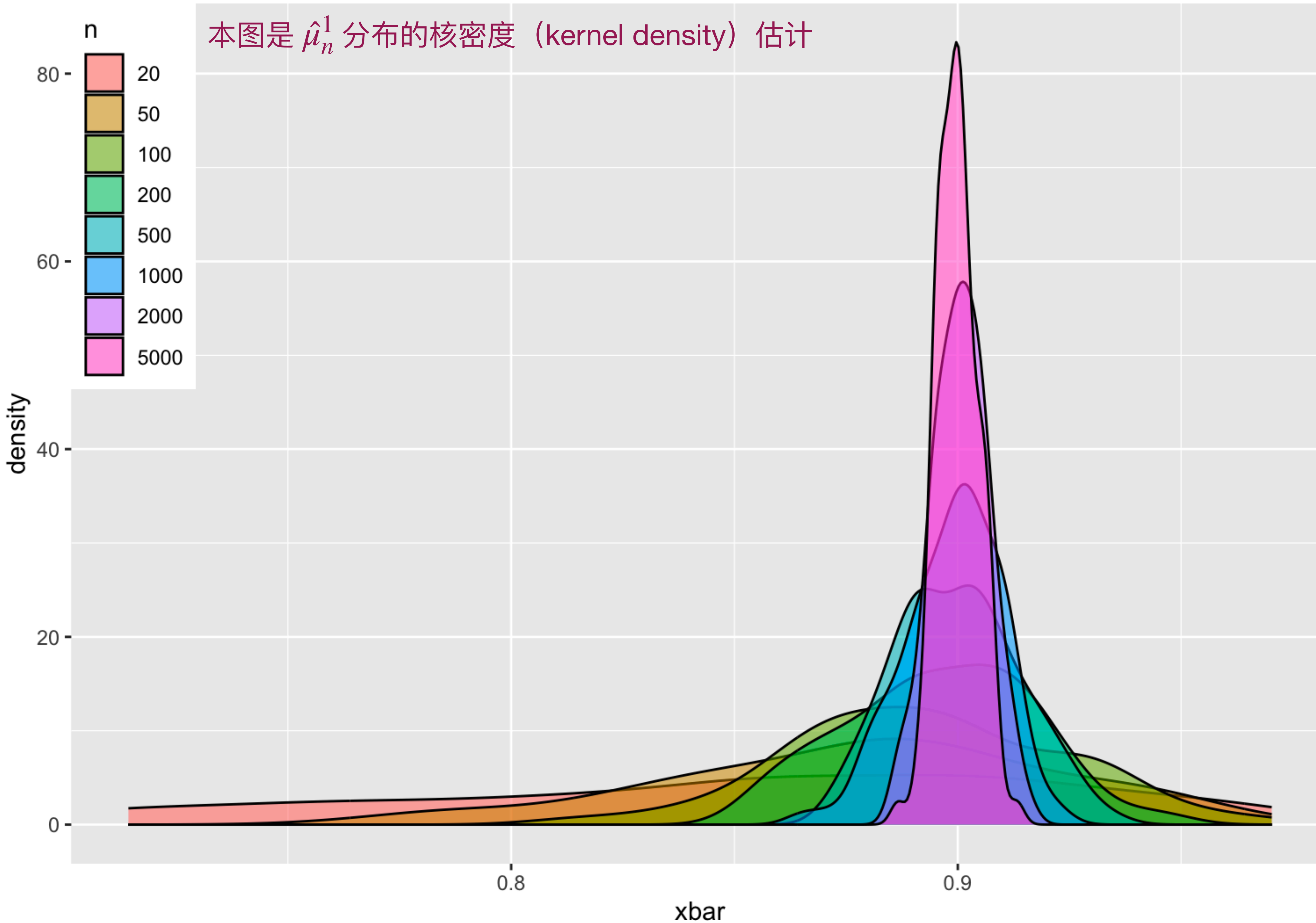


针对每个 n，生成 100 组随机样本并计算每组样本的 $\hat{\mu}_n^1 = \frac{n}{n+1}\bar{X}_n$。

图中是 $\hat{\mu}_n^1$ 的分布。

模拟大数定律

$X_i \sim \text{Bernoulli}(p = 0.9), n = 20{,}50{,}100{,}200{,}500{,}1000{,}2000{,}5000$



中心极限定理

Central limit theorem (CLT)

随机变量的极限之分布极限：Z_n 服从分布函数 $G_n(u) = \Pr(Z_n \le u)$。如果对任意 u，$G(u) = \Pr(Z \le u)$ 是连续函数且当 $n \to \infty$ 时 $G_n(u) \to G(u)$，则称 Z_n **依分布收敛**（**converges in distribution**）于 Z，写作 $Z_n \xrightarrow{d} Z$

- 称 $G(u)$ 为 Z_n 的**渐进分布**（**asymptotic distribution**）或**大样本分布**（**large sample distribution**）或**极限分布**（**limit distribution**）

Lindeberg-Lévy 中心极限定理：如果 X_i 为 i.i.d. 且 $E[X_i^2] < \infty$，则当 $n \to \infty$ 时

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} Z \sim N(0, \sigma^2)$$

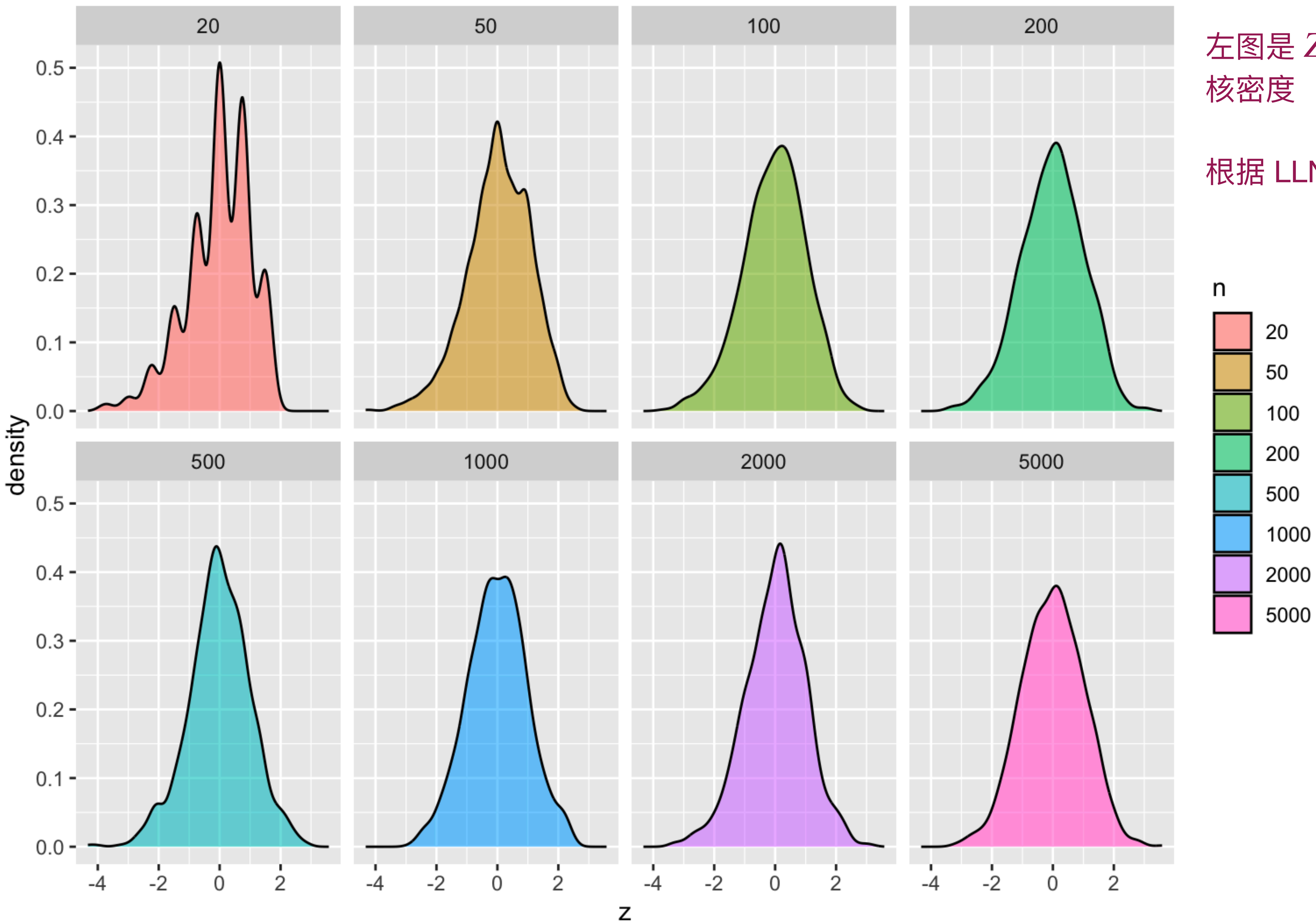
此处 $\mu = E[X_i]$，$\sigma^2 = E[(X_i - \mu)^2]$，$N(a, b^2)$ 为均值为 a 方差为 b^2 的正态分布

- 在有限样本下，$Z_n = \sqrt{n}(\bar{X}_n - \mu)$ 的均值为 0，方差为 σ^2。LLN 告诉我们 Z_n 的渐进分布是正态分布

- CLT 也可以写成 $\bar{X}_n \overset{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$，即可以用 $N\left(\mu, \frac{\sigma^2}{n}\right)$ 作为 $\bar{X}_n$ 分布的近似

模拟中心极限定理

$X_i \sim \text{Bernoulli}(p=0.9)$, 针对每个 n 生成 1000 组随机样本

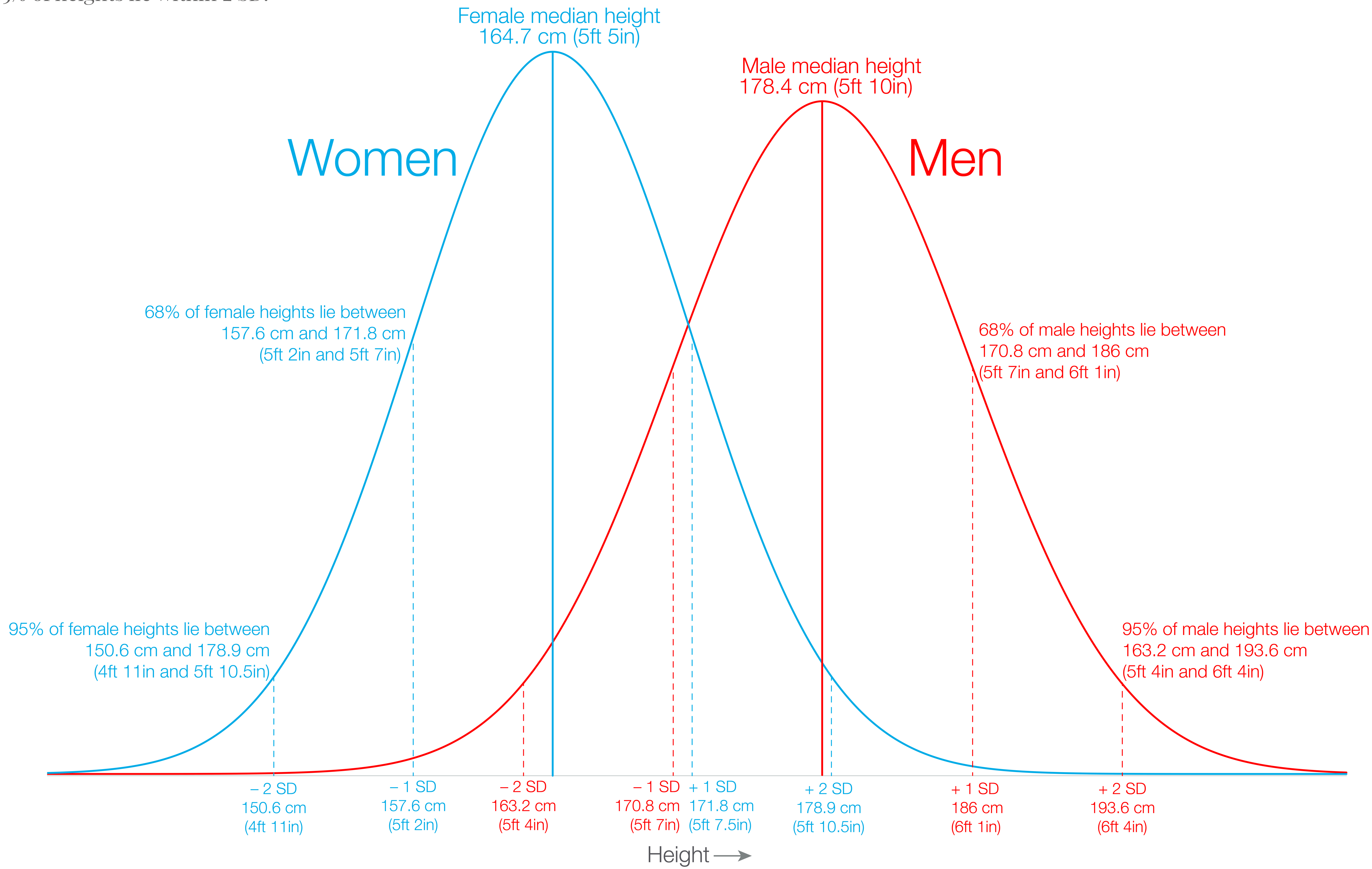


左图是 Z_n/σ 的分布的核密度（kernel density）估计

根据 LLN，渐进分布为 $N(0,1)$

The distribution of male and female heights

The distribution of adult heights for men and women based on large cohort studies across 20 countries in North America, Europe, East Asia and Australia. Shown is the sample-weighted distribution across all cohorts born between 1980 and 1994 (so reaching the age of 18 between 2008 and 2012).
Since human heights within a population typically form a normal distribution:
– 68% of heights lie within 1 standard deviation (SD) of the median height;
– 95% of heights lie within 2 SD.



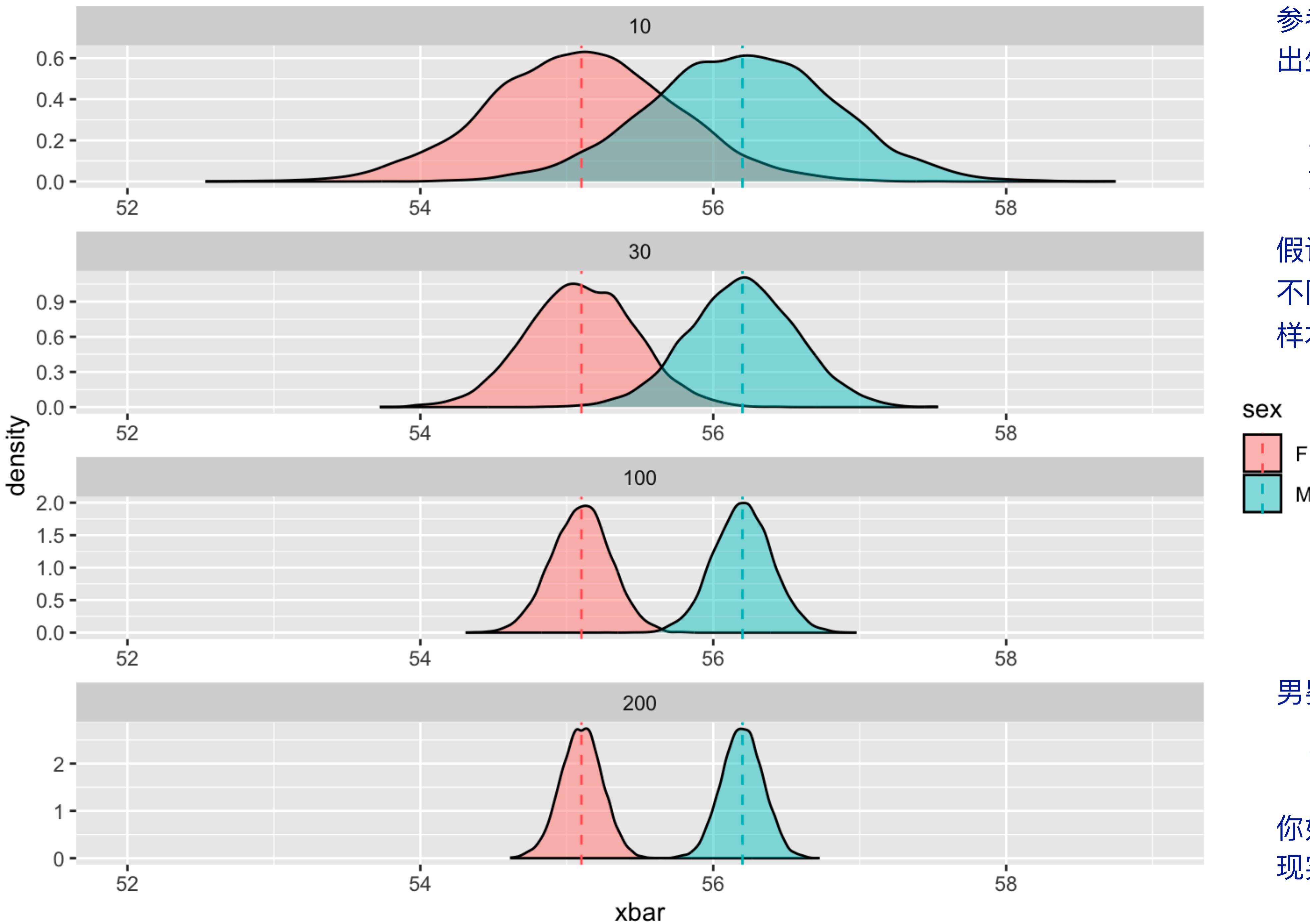
Note: this distribution of heights is not globally representative since it does not include all world regions due to data availability.

Data source: Jelenkovic et al. (2016). Genetic and environmental influences on height from infancy to early adulthood: An individual-based pooled analysis of 45 twin cohorts.

This is a visualization from OurWorldinData.org, where you find data and research on how the world is changing.

两个总体的抽样分布

男婴与女婴的体长分布



参考 WHO 婴儿体长分布标准，
出生后 6 个月的婴儿体长为：

男婴：$(\mu, \sigma) = (56.2,\ 2)$
女婴：$(\mu, \sigma) = (55.1,\ 2)$

假设二者都服从正态分布，左图为不同样本量（$n = 10{,}30{,}100{,}200$）时样本均值 $\bar{X}_n$ 的抽样分布的核密度估计

男婴与女婴的实际身长均值之差为

1.1 cm > 0

你如何理解统计结果的统计学意义与现实意义？

假设检验和置信区间

假设检验的思维方式

The Idea of Hypothesis Testing

考虑回归模型

$$y_t = \beta + u_t\ ,\quad u_t \sim \mathrm{IID}(0,\sigma^2)$$

此时 OLS 估计量满足 $\hat{\beta} = \frac{1}{n}\sum_{t=1}^{n} y_t$，$\mathrm{Var}[\hat{\beta}] = \frac{1}{n}\sigma^2$。

我们想知道总体中的 β 是否满足某种限制条件，例如

$$H_0:\ \beta = \beta_0 \quad \text{vs} \quad H_1:\ \beta \neq \beta_0$$

H_0 是零假设（null hypothesis），H_1 是备择假设（alternative hypothesis）

我们要针对 H_0 进行检验，就是要找到一个检验统计量（test statistic），并使其满足：

1. 当零假设正确时，我们知道该统计量服从分布 A；
2. 当零假设错误（即备择假设正确）时，我们知道该统计量服从分布 $B \neq A$。

如果零假设正确时很难获得样本中检验统计量的取值的话，我们就有理由相信零假设是错误的。当理由充分时，我们即可拒绝（reject）零假设，而偏向于接受（accept）备择假设。

统计量的分布

Distribution of Test Statistic

首先我们假定零假设 $H_0: \ \beta = \beta_0$ 成立（这里的 β_0 就是总体中 β 的真实值）。同时我们假设 u_t 服从正态分布且 σ 已知。

一个常用的检验统计量是

$$z = \frac{\hat{\beta} - \beta_0}{\sqrt{\mathrm{Var}[\hat{\beta}]}} = \frac{\sqrt{n}}{\sigma}(\hat{\beta} - \beta_0)$$

因为 $\hat{\beta}$ 非偏，则在零假设下 $E[z] = 0$。z 的方差为

$$\mathrm{Var}[z] = E[z^2] = \frac{n}{\sigma^2}E[(\hat{\beta} - \beta_0)^2] = \frac{n}{\sigma^2}\frac{\sigma^2}{n} = 1$$

如果 u_t 服从正态分布，z 也服从正态分布，所以 $z \sim N(0,1)$。

如果备择假设 $H_1: \ \beta \neq \beta_0$ 成立，那针对任意 $\beta = \beta_1 \neq \beta_0$，则有 $\hat{\beta} = \beta_1 + \hat{\gamma}$。可知 $\hat{\gamma}$ 服从 $N(0, \sigma^2/n)$，因此 $z \sim N(\lambda,1)$，$\lambda = \frac{\sqrt{n}}{\sigma}(\beta_1 - \beta_0)$。

拒绝域和接受域

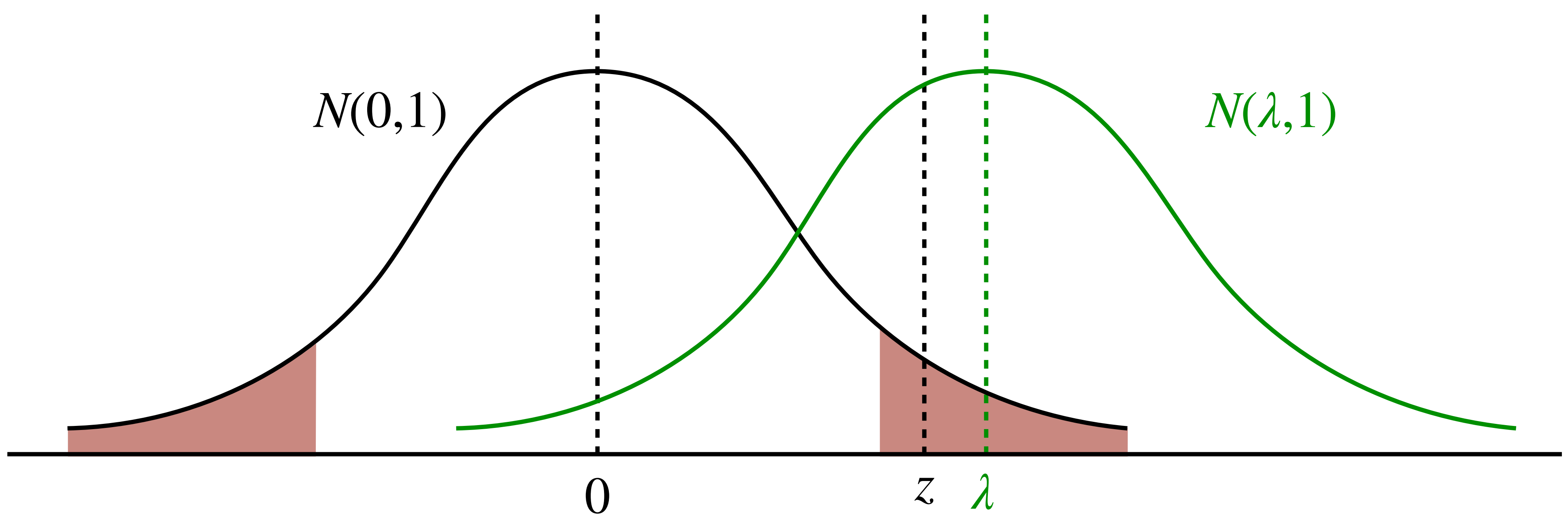
Rejection and Acceptance Regions

$$H_0: \ \beta = \beta_0 \quad \Rightarrow \quad z \sim N(0,1)$$

$$H_1: \ \beta \neq \beta_0 \quad \Rightarrow \quad \text{对任意的}\, \beta = \beta_1, \ z \sim N(\lambda, 1), \ \lambda = \frac{\sqrt{n}}{\sigma}(\beta_1 - \beta_0)$$

当 n 足够大时，则在备择假设下，大概率可以观测到 z 的取值显著不为零。如果我们确实观测到 $|z| \gg 0$，就可以拒绝零假设。

我们需要事先规定一个拒绝零假设的规则。通常我们设定一个拒绝域（rejection region），当 z 的取值在该域中时，就拒绝零假设。检验（test）= 检验统计量 + 拒绝规则。拒绝域以外的区域是接受域（acceptance region）。



检验可能出错

因为样本的随机性，所有检验都可能出现错误。

通常存在两种检验错误：

	检验结果：接受零假设	检验结果：拒绝零假设
零假设成立	正确	Type I error
备择假设成立	Type II error	正确

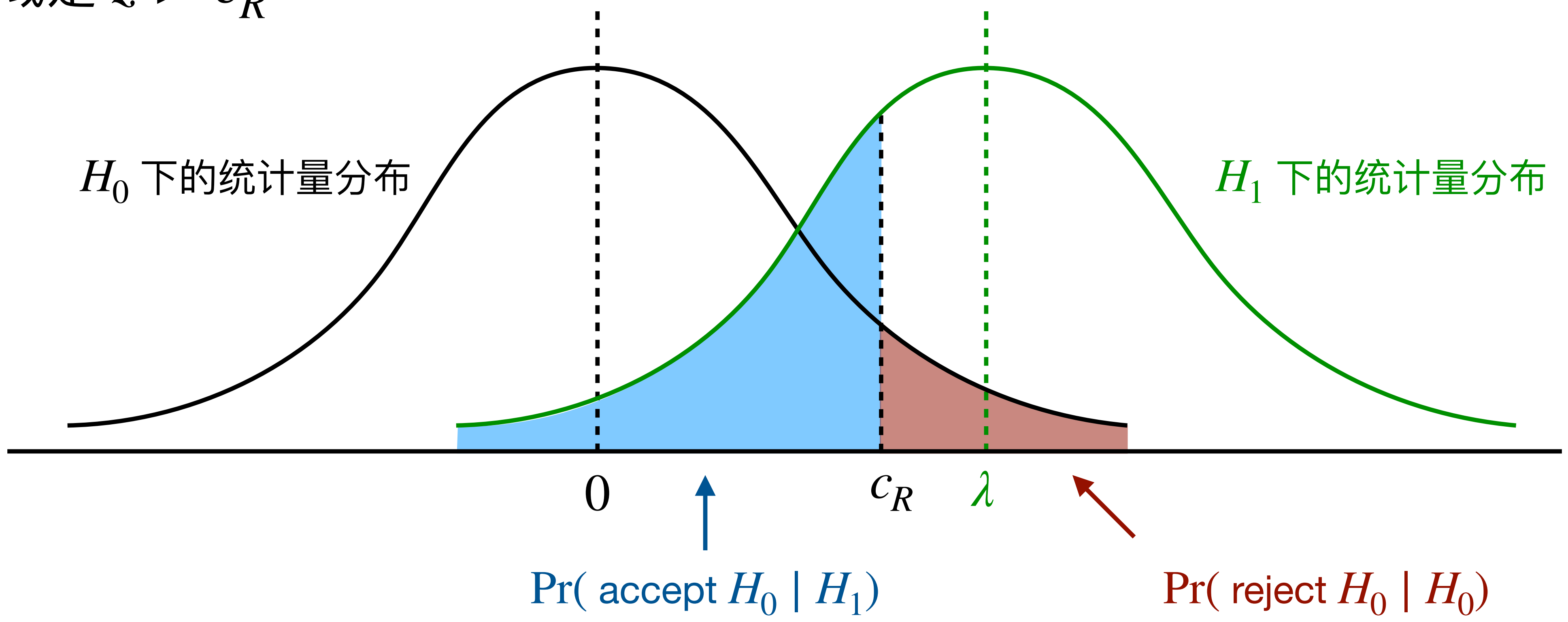
检验的 size：$\Pr($ type I error $) = \Pr($ reject $H_0 \mid H_0)$

检验的功效（power）：$1 - \Pr($ type II error$) = 1 - \Pr($ accept $H_0 \mid H_1)$

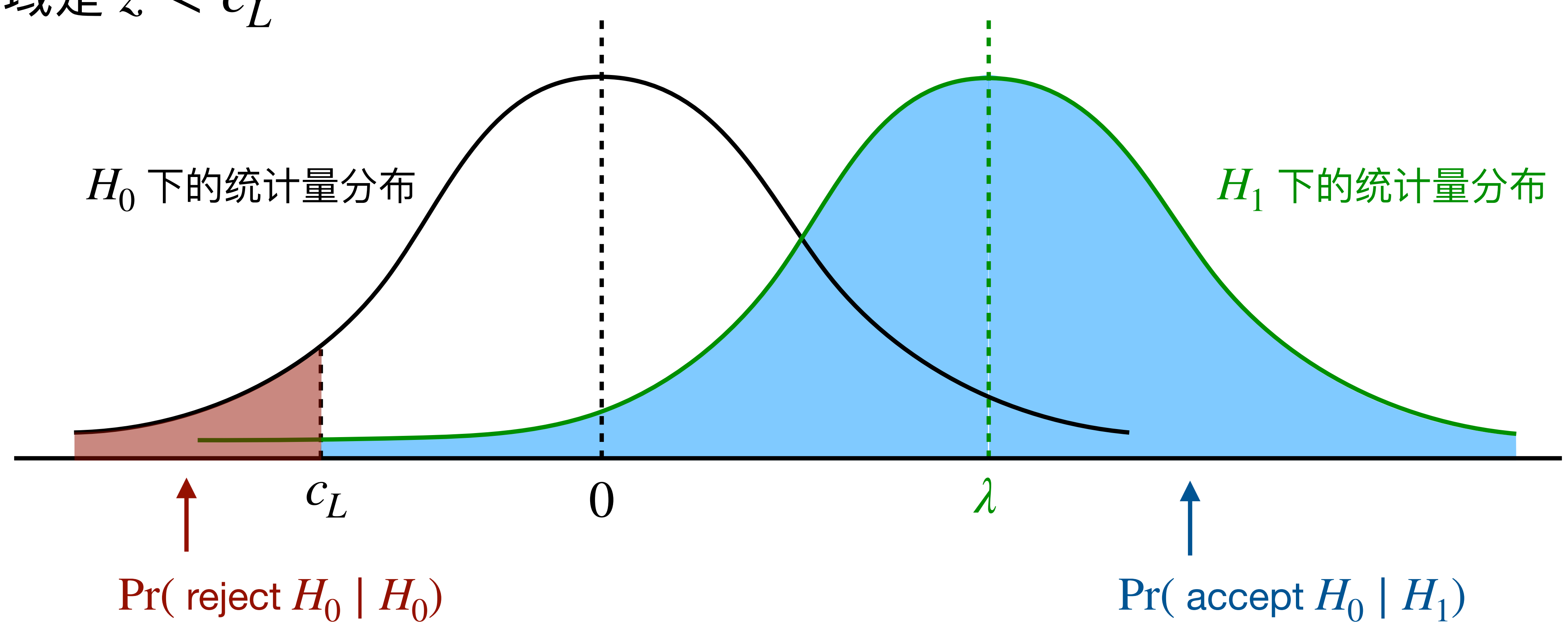
检验的拒绝域越小，size 也就越小，但同时功效也越小，因此无法同时减小两种错误的概率。我们通常选择一个可以接受的 size，然后选择功效最大的检验方法（the Neyman-Pearson approach）。

研究者可以接受的最大 size 值（允许 type I error 发生的最大概率）称为检验的显著性水平（level of significance），通常记作 α。

当拒绝域是 $z > c_R$



当拒绝域是 $z < c_L$



临界值和 *P* 值

Critical Value and the *P* Value

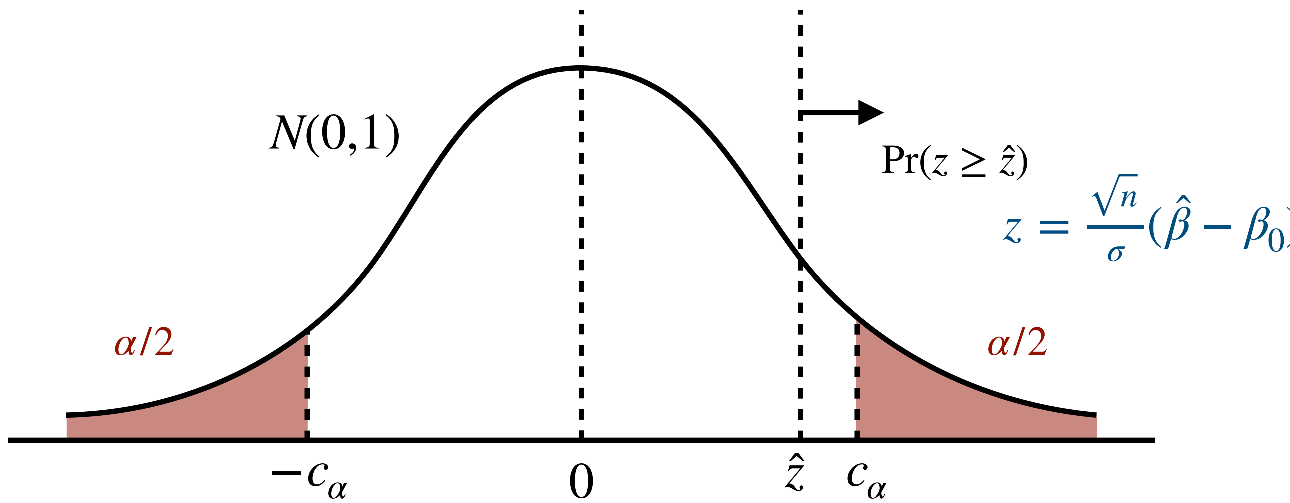
当给出检验的显著性水平 α 时，我们可以通过确定临界值的方式给出该检验的拒绝域。

z 统计量服从标准正态分布，则双尾检验的临界值 c_α 可以通过下面的隐函数定义

$$\Phi(c_\alpha) = 1-\frac{\alpha}{2}$$

$$\Leftrightarrow \quad c_\alpha = \Phi^{-1}(1-\frac{\alpha}{2})$$

Φ 是标准正态分布的累积分布函数



当我们观测到估计值 $\hat{z}$ 时，我们将 $2\mathrm{Pr}(z > |\hat{z}|)$ 定义为该检验的 *P* 值（marginal/observed significance level）。*P* 值是临界值为 $\hat{z}$ 的检验所对应的显著性水平。

通过比较 *P* 值和 α 的大小，可以直观地判断检验结果（拒绝或接受零假设）。

P 值受样本量的影响。样本量 n 越大，$\hat{z}$ 的取值就越大，*P* 值就越小。因此，在大样本下，我们应当审视 5% 或 1% 常用显著性水平的合理性（应当选择更小的 α），并关注估计值的经济学含义（如何解释回归系数）。

置信区间

Confidence Interval

真实参数值 θ 和估计量 $\hat{\theta}$ 之间的关系是

$$\theta = \hat{\theta} + \text{抽样误差}$$

$\hat{\theta}$ 是 θ 的点估计（point estimation），而 $\hat{\theta}$ 加上抽样分布可以给出区间估计（interval estimation）。

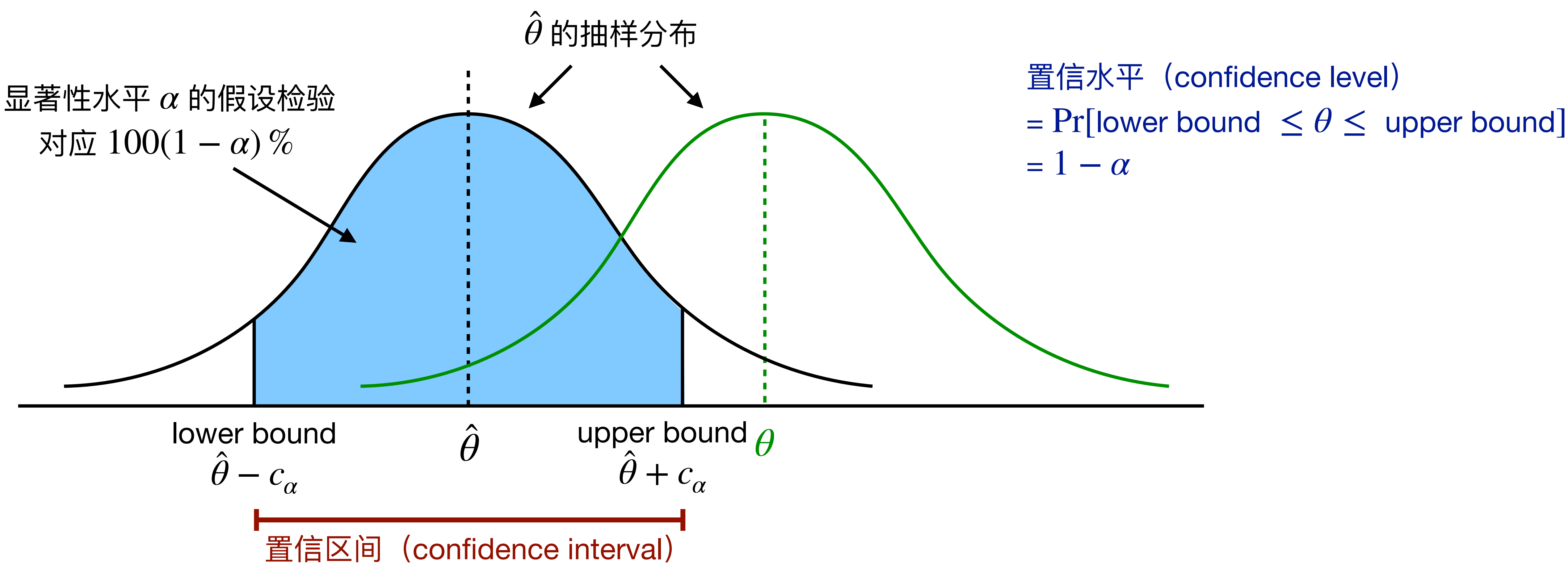
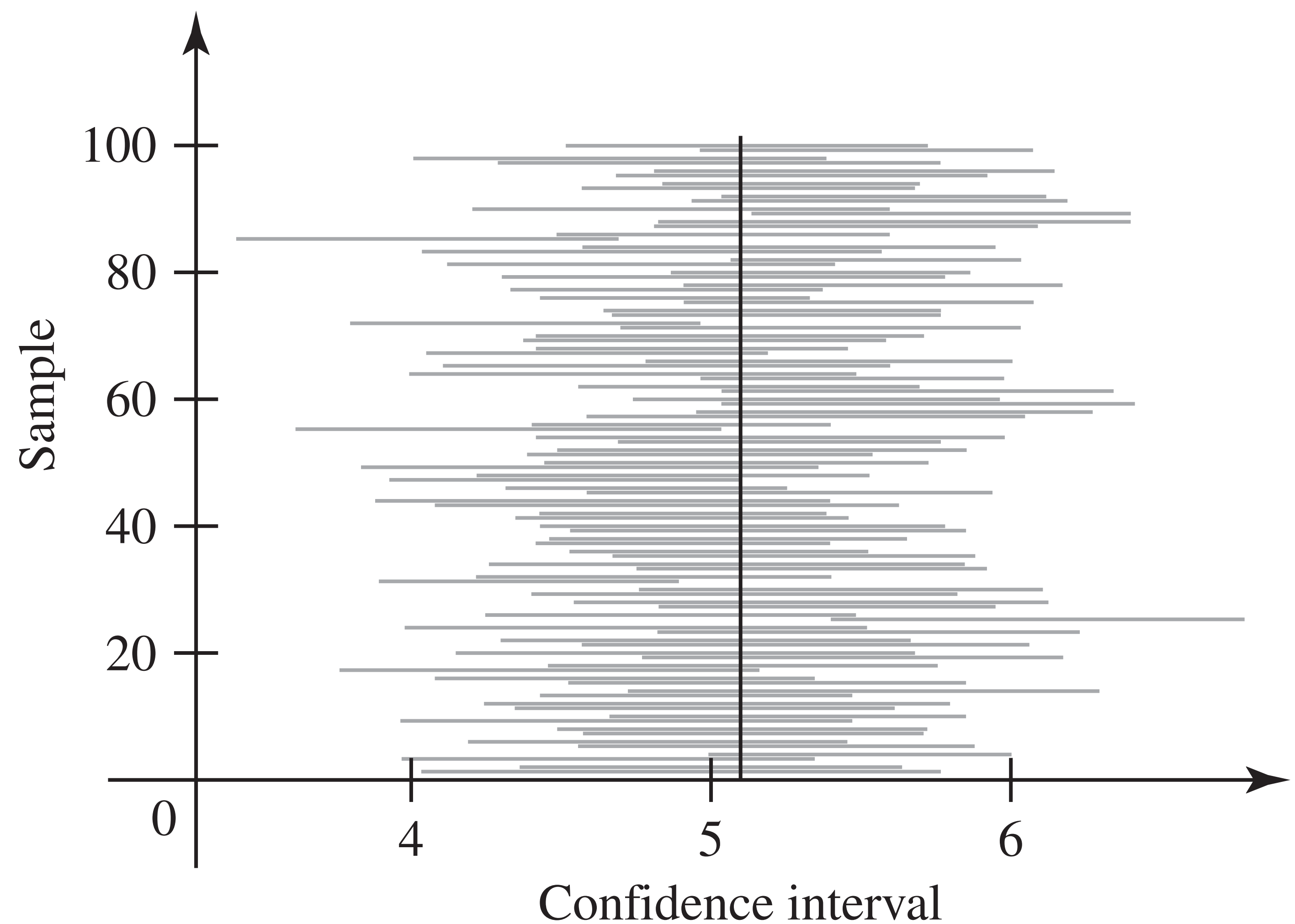


Figure 8.5 A sample of one hundred observed 95% confidence intervals based on samples of size 26 from the normal distribution with mean $\mu = 5.1$ and standard deviation $\sigma = 1.6$. In this figure, 94% of the intervals contain the value of μ.



DeGroot & Schervish (2012), *Probability and Statistics*, 4th Edition, Pearson. (p.478)

根据正态分布 $N(\mu = 5.1, \sigma^2 = 1.6^2)$ 随机生成 $n = 26$ 的样本，然后生成置信区间。图中包含了100个这样的置信区间，其中94个包含真实的分布均值 $\mu = 5.1$。

常用的概率分布

正态分布
The Normal Distribution

正态分布也称高斯分布（Gaussian distribution），由期望值 μ 和方差 σ^2 两个参数决定。

$N(\mu, \sigma^2)$ 的密度函数是

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

如果 $x \sim N(\mu, \sigma^2)$，则 $z = \frac{x-\mu}{\sigma} \sim N(0,1)$。

- 服从正态分布的独立变量的线性结合也服从正态分布。

多变量正态分布

The Multivariate Normal Distribution

随机向量 $\boldsymbol{x} = (x_1, \ldots, x_m)$ 服从期望值为 $\boldsymbol{\mu}$，协方差矩阵为 $\boldsymbol{\Omega}$ 的多变量正态分布可以表达为 $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega})$。其密度函数为

$$f(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Omega}) = \frac{1}{(2\pi)^{m/2} \, |\, \boldsymbol{\Omega} \,|^{1/2}} \exp\Big(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^{\top} \boldsymbol{\Omega}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\Big)$$

因为 $\boldsymbol{\Omega}$ 是协方差矩阵，因此是半正定。
又因 $\boldsymbol{\Omega}$ 可逆，因此是正定矩阵。

- 如果 $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega})$，则 $\boldsymbol{a}^{\top}\boldsymbol{x} \sim N(\boldsymbol{a}^{\top}\boldsymbol{\mu}, \boldsymbol{a}^{\top}\boldsymbol{\Omega}\boldsymbol{a})$。
- 如果 $\boldsymbol{x}$ 服从多变量正态分布，且协方差都为零（要素间相互不相关），则 $x_1, \ldots, x_m$ 相互独立。

相互独立（mutual independence）：
$f(x_1, x_2, \ldots, x_m) = f(x_1)f(x_2)\cdots f(x_m)$

卡方分布

The Chi-squared Distribution

当随机向量 $z \sim N(\mathbf{0}, \boldsymbol{I})$ 时，随机变量

$$y \equiv \|z\|^2 = z^\top z = \sum_{t=1}^{m} z_t^2$$

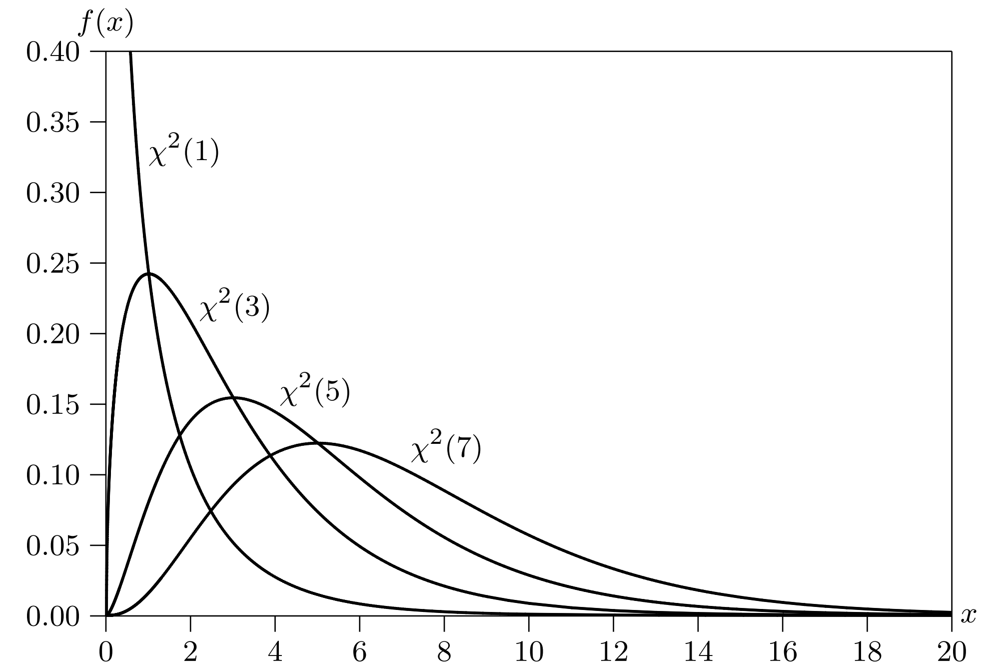


Figure 4.4 Various chi-squared PDFs

服从自由度为 m 的卡方分布（chi-squared distribution with m degrees of freedom），写成 $y \sim \chi^2(m)$。

- $E[y] = \sum_{t=1}^{m} E[z_t^2] = \sum_{t=1}^{m} 1 = m$
- $\mathrm{Var}[y] = \sum_{t=1}^{m} \mathrm{Var}[z_t^2] = mE[(z_t^2 - 1)^2]$
  $= mE[z_t^4 - 2z_t^2 + 1] = m(3 - 2 + 1) = 2m$
- $y_1 \sim \chi^2(m_1),\ y_2 \sim \chi^2(m_2) \quad \Rightarrow \quad (y_1 + y_2) \sim \chi^2(m_1 + m_2)$

服从卡方分布的随机变量

定理

1. 如果维度为 m 的随机向量 $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{\Omega})$，则 $\boldsymbol{x}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{x} \sim \chi^2(m)$；
2. 如果 $\boldsymbol{P}$ 是投影矩阵且 $\operatorname{rank}(\boldsymbol{P}) = r$，$\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{I})$ 是维度为 n 的随机向量，则 $\boldsymbol{z}^\top \boldsymbol{P} \boldsymbol{z} \sim \chi^2(r)$。

证明：

1. $\boldsymbol{\Omega}$ 是对称正定矩阵，因此存在非奇异矩阵 $\boldsymbol{A}$ 满足 $\boldsymbol{\Omega} = \boldsymbol{A}\boldsymbol{A}^\top$（Section 3.4）。此时考虑 $\boldsymbol{z} = \boldsymbol{A}^{-1}\boldsymbol{x}$：$E[\boldsymbol{z}] = E[\boldsymbol{A}^{-1}\boldsymbol{x}] = \boldsymbol{A}^{-1}E[\boldsymbol{x}] = \boldsymbol{0}$，且
$$\begin{aligned}\operatorname{Var}[\boldsymbol{z}] &= \operatorname{Var}[\boldsymbol{A}^{-1}\boldsymbol{x}] = E[\boldsymbol{A}^{-1}\boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{A}^\top)^{-1}] \\ &= \boldsymbol{A}^{-1}E[\boldsymbol{x}\boldsymbol{x}^\top](\boldsymbol{A}^\top)^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{A}^\top(\boldsymbol{A}^\top)^{-1} = \boldsymbol{I}_m\end{aligned}$$
因此 $\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{I})$。而 $\boldsymbol{x}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{x} = \boldsymbol{x}^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{A}^{-1}\boldsymbol{x} = \boldsymbol{z}^\top\boldsymbol{z}$，所以 $\boldsymbol{x}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{x} \sim \chi^2(m)$。
2. 假设 $\boldsymbol{P}$ 投影到 $n \times r$ 矩阵 $\boldsymbol{Z}$ 的列空间，因此 $\boldsymbol{P} = \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$，
$$\boldsymbol{z}^\top\boldsymbol{P}\boldsymbol{z} = \boldsymbol{z}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{z}$$
如果令 $\boldsymbol{x} = \boldsymbol{Z}^\top\boldsymbol{z}$，则 $E[\boldsymbol{x}] = \boldsymbol{0}$, $\operatorname{Var}[\boldsymbol{x}] = E[\boldsymbol{Z}^\top\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{Z}] = \boldsymbol{Z}^\top\boldsymbol{Z}$，因此 $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{Z}^\top\boldsymbol{Z})$。可见 $\boldsymbol{z}^\top\boldsymbol{P}\boldsymbol{z} = \boldsymbol{x}^\top(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{x} \sim \chi^2(r)$。

Student's t 分布

The Student's t Distribution

如果 $z \sim N(0,1)$，$y \sim \chi^2(m)$，且 z 和 y 相互独立，则

$$t_m \equiv \frac{z}{\sqrt{\frac{y}{m}}}$$

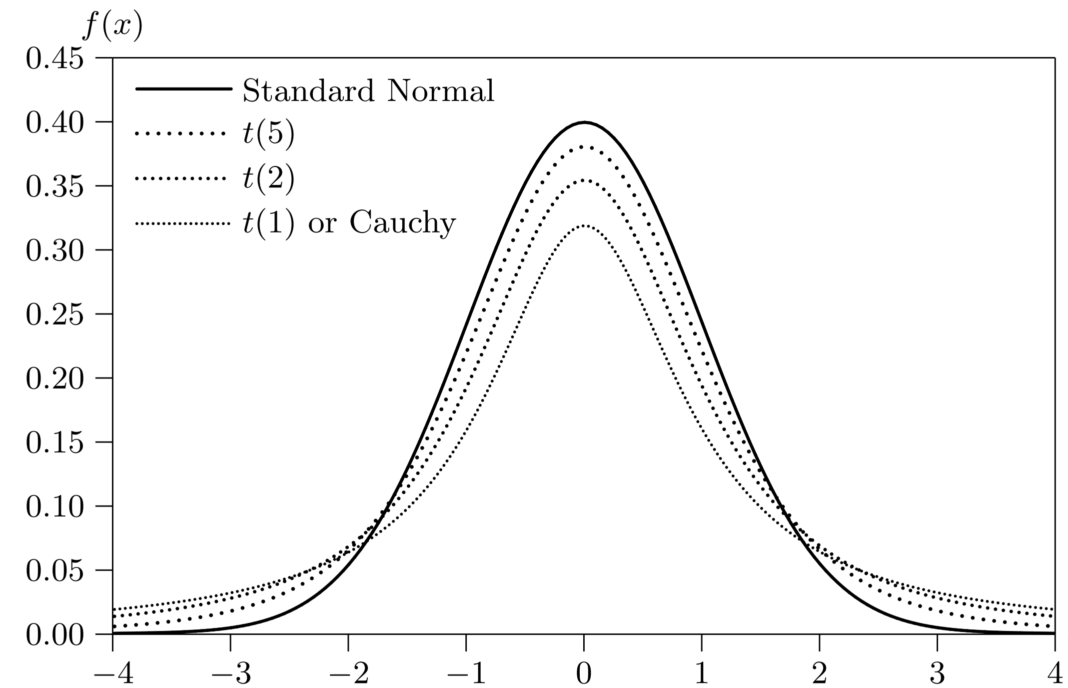


Figure 4.5 PDFs of the Student's t distribution

服从自由度为 m 的 Student's t 分布，写成 $t_m \sim t(m)$。

- 自由度为 m 的 t 分布存在前 $m-1$ 个矩。因此 $t(1)$ 没有矩，$t(2)$ 仅有期望值。
- 若存在，则 $E[t_m]=0$，$\text{Var}[t_m]=m\big/(m-2)$。
- 当 $m \to \infty$ 时，t 分布的方差收敛于 1，分布自身收敛于正态分布。

F 分布

The *F* Distribution

当 $y_1 \sim \chi^2(m_1)$，$y_2 \sim \chi^2(m_2)$，且 y_1 与 y_2 相互独立时，

$$F_{m_1,m_2} = \frac{y_1 \big/ m_1}{y_2 \big/ m_2}$$

服从自由度为 m_1 和 m_2 的 *F* 分布，写成 $F_{m_1,m_2} \sim F(m_1, m_2)$。

- $E[F_{m_1,m_2}] = (m_2 - 1)\big/(m_2 - 3)$。
- 当 $m_2 \to \infty$ 时，$m_1 F_{m_1,m_2} \to y_1 \sim \chi^2(m_1)$。
- 由 *t* 分布的定义可知，当 $x \sim t(m_2)$ 时，$x^2 \sim F(1, m_2)$。