Foundations of Data Science

Introduction to Statistics

尹一通、刘明谋 Nanjing University, 2024 Fall

统计的起源

- **8-13世纪（伊斯兰黄金时代）**：文本加密基于逐字替换。阿拉伯数学家和密码学家们通过统计字母出现的频率来解码。现在被称为频数分析。
  - 《密码信息之书》《密码信息解密手稿》
- 春秋战国、秦汉、古希腊、古罗马时代**威权机关**对人口、物资等资源的**普查**。
  - statistics与state同源，“国家科学”
  - 最早中文：传教士马礼逊《华英字典》“统纪”。传至日本，再反传回“统计”。
- 对赌术**的研究** (games of chance)、对误差**的研究** (theory of errors)，等等。

生活中的统计问题

- 保险：我应该买保险吗？
  - 中国卫生部公布的数据显示，人的一生罹患重大疾病的机会高达72.18%
  - 人寿险？重疾险？意外险？养老险？
  - 保险公司只有两种情况不保：这也不保、那也不保。
- 医疗方案：
  - 方案一：“100%有效”，应用3例治愈3例。
  - 方案二：“95%有效”，应用20例治愈19例。
  - 方案三：“90%有效”，应用100例治愈90例。
    - 哪种更有效？我该选哪种？
- 生男生女的概率？你觉得是 50% 吗?
  - 2023年全国男性人口72032万人，女性人口68935万人
  - 2021年全国出生人口性别比108.3，2019年全国出生人口性别比110.14
  - John Arbuthnot (1710): 1629 to 1710, in every year, the number of males born in London exceeded the number of females.

$$1/2^{82} = 1/4835703278458516698824704$$

战争中的统计问题

- 伦敦轰炸：德军已经掌握了伦敦的部署了吗?
  - R. D. Clarke (保险公司精算师)

    An Application of the Poisson Distribution
  - 弹坑分布服从泊松分布

TABLE 1 Comparing the expected results if V-1s followed the Poisson distribution to the actual results over a region of south London. The expected number of squares is given by $P(\lambda, k) \times 576$ with $\lambda = 537/576$ and $k = 0, 1, 2, \ldots$. Clarke's reported p-value of 0.88 is for a chi-squared test with 4 degrees of freedom (DOF).[1] There are 6 classes in the table and we have estimated one parameter (λ), so DOF = (6 – 1) – 1 = 4.

Number of V-1s in square (k)	Expected number of squares (Poisson)	Observed number of squares
0	226.74	229
1	211.39	211
2	98.54	93
3	30.62	35
4	7.14	7
5 and over	1.57	1
	576	576

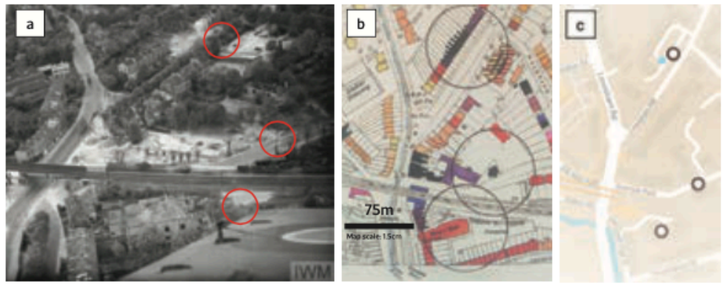


FIGURE 1 A cluster of three V-1s near a railway line in Lewisham, south London. Patterns like these motivated Clarke to test whether their distribution was random. (a) A contemporary aerial photograph of the V-1 sites, showing the large radius of destruction of houses around the bomb sites (© IWM, catalogue no. CH 15109). (b) The same region from the LCC bomb damage maps[4] (p. 162), showing the three V-1 hits (black circles). Damage to houses is coded by colour, with black indicating total destruction. (c) The same region in the Google Maps layer (hit locations V1.509–V1.511).

- 德国坦克问题：每辆德军坦克的不同部件都有序列号，且是顺序编号的。
  - 某个情报人员已经发现了k = 4辆坦克，其序列号分别为2、6、7、14，观测到的最大的序列号为m = 14。坦克总数可能是？每月产量可能是？
- Enigma密码机：$150738274937250 \approx 1.51 \times 10^{14}$ 种组合
  - P表示接线板的连线，U表示反射器，L、M、R表示左、中、右转子。$E = PRMLUL^{-1}M^{-1}R^{-1}P^{-1}E_0$
  - 波兰三杰：统计字母重复规律、估计转子循环圈、发现缺陷与规律：密钥总数为105456个

统计的起源

- **8-13世纪（伊斯兰黄金时代）**：文本加密基于逐字替换。阿拉伯数学家和密码学家们通过统计字母出现的频率来解码。现在被称为频数分析。
  - 《密码信息之书》《密码信息解密手稿》
- 春秋战国、秦汉、古希腊、古罗马时代**威权机关**对人口、物资等资源的**普查**。
  - statistics与state同源，“国家科学”
  - 最早中文：传教士马礼逊《华英字典》“统纪”。传至日本，再反传回“统计”。
- 对赌术**的研究** (games of chance)、对误差**的研究** (theory of errors)，等等。
- 本质：观察、收集、去芜存菁、总结、抽象、建模

  - 科学

  - “All models are wrong, but some are useful.” —— George E. P. Box

从这些中可以看到什么？统计是在干什么？

- 问题/目标
  - 量化：值得吗？有效吗？可能是多少？
- 调查/数据收集
  - 有的放矢
- 数据分析
  - 去芜存菁、抽象、总结
- 解释、推断、决策
  - 为什么？可能会怎么样？该怎么样？多大把握？

概率与统计

- 概率论：在良定义/理想化的模型中、基于严格的理论、考虑一个事件的可能性
  - 丢硬币：伯努利试验、二项分布、几何分布、负二项分布、泊松分布
  - Balls into Bins：birthday、coupon collector、max load
  - Erdős–Rényi random graph model
  - Hashing, QuickSort, data streaming (count distinct elements)
- 统计学：观察、收集、去芜存菁、总结、抽象、假设、建模、预测、推断
  - 现实中的随机性、数据的随机性、 噪声与干扰
  - 以解释、预测和推断为目的

All models are wrong,
but some are useful.

概率与统计

- 概率论：已知参数，具有确定参数的模型，计算概率
  - 已知某治疗方案有效率80%。可以预计如果对100位病患采用该方案，平均可治愈80位，且至少有 99.99% 的概率治愈至少65位。

- 统计学：已知数据，需要猜测模型，估计概率
  - 对100位病患采用某治疗方案，治愈了78位。可以推测：如果再对100位病患采用该方案，以95%的置信度，至少治愈69位至多治愈87位。

Point Estimation

一个简单的例子

- 一枚硬币，我们丢一次，出现正面的概率是多少？
- 假设：每一次丢硬币都是相互独立的伯努利试验

  - 为什么是随机试验?

    - 我们对世界的理解太少，于是我们**相信**世界是随机的
    - 量子力学：世界**本身**有内在的**随机**性

  - 为什么是相互独立的伯努利试验?

    - 对世界的理解太少，理想化的抽象：我们**相信**试验之间的关联足够少

All models are wrong,
but some are useful.

一个简单的例子

- 一枚硬币，我们丢一次，出现正面的概率是多少？
- 假设：每一次丢硬币都是相互独立的伯努利试验
- 此前丢了100次，我们观察到49次正面。怎么估计比较好？
- 49/100?
- 大数定律：$\lim_{n\to\infty}$ 样本均值=期望=正面概率

All models are wrong,
but some are useful.

Maximum Likelihood Estimation (MLE)

最有可能是哪个参数：哪个参数最有可能让我们观察到现有的数据？

- 一枚硬币，我们丢一次，出现正面的概率是未知参数 p
  - 概率质量函数：$f(x;p) = p$ if $x = 1; f(x;p) = 1 - p$ if $x = 0.$
- 观察100次投硬币的结果的序列，其中49次正面
- 似然函数 $L(p) = L(p; x_1, \ldots, x_n) = \Pi_i f(x_i; p) = p^{49}(1-p)^{51}$
- $L'(p) = (1-p)^{50}p^{48}(49 - 100p) = 0 \implies p = 0.49$
- 估计量 $\hat{p}$ 的最大似然估计值 $\hat{p} = 0.49$
  - “ 哪个参数<u>最有可能</u>使得当前数据出现 ”: $\text{argmax}_{\hat{p}} \Pr(X = x \mid p = \hat{p})$

Maximum Likelihood Estimation (MLE)

哪个参数最有可能让我们观察到现有的数据？

- 100次投硬币中观察到了49次正面：最大似然估计值 $\hat{p} = 0.49$
- 最大似然估计期望并不总是等于样本均值
  - YES: 伯努利试验、二项分布、泊松分布、正态分布
  - NO: 几何分布、指数分布
- 求最大似然估计 = 最优化 $L(p)$
  - 不总是易解。迭代逼近：梯度下降、牛顿-拉弗森法、拟牛顿法、expectation–maximization algorithm

Method of moments (MOM)

Moment problem + Law of large numbers

- Moment problem: 两个概率分布如果所有的矩都相同，那么它们是同一个分布
  - 根据各阶矩可以确定一个概率分布，即给定参数之后的那个未知分布
- 大数定律：$\overline{X}_n^k = (X_1^k + \ldots + X_n^k)/n \to \mathbb{E}\left[X^k\right]$ as $n \to \infty$
  - k-阶样本矩依概率/几乎处处收敛到k-阶矩

根据样本矩可以估计一个概率分布

Method of moments (MOM)

Moment problem + Law of large numbers

- 各阶矩确定一个概率分布 + $\overline{X}_n^k = (X_1^k + \ldots + X_n^k)/n \to \mathbb{E}\left[X^k\right]$ as $n \to \infty$
  - 根据**样本矩**可以**估计**一个概率分布
- 假设样本 $X_1, \ldots, X_n$ 服从某正态分布 $N(\mu, \sigma^2)$：
  - $\mathbb{E}[X] = \mu, \mathbb{E}[X^2] = \mu^2 + \sigma^2$
  - 样本矩 $m_1 = (X_1 + \ldots X_n)/n, m_2 = (X_1^2 + \ldots X_n^2)/n$
  - 令 $m_1 = \mu, m_2 = \mu^2 + \sigma^2$，解得

$$\begin{cases} \hat{\mu} = (X_1 + \ldots X_n)/n = \overline{X}_n; \\ \hat{\sigma^2} = ((X_1 - \overline{X}_n)^2 + \ldots (X_n - \overline{X}_n)^2)/n = S_n^2 \end{cases}$$

Method of moments (MOM)

Moment problem + Law of large numbers

- 各阶矩确定一个概率分布 + $\overline{X}_n^k = (X_1^k + \ldots + X_n^k)/n \to \mathbb{E}\left[X^k\right]$ as $n \to \infty$
  - 根据**样本矩**可以**估计**一个概率分布
- 假设样本 $X_1, \ldots, X_n$ 服从某泊松分布 $\mathrm{Poi}(\lambda)$：
  - $\mathbb{E}[X] = \lambda, \mathbb{E}[X^2] = \lambda^2 + \lambda$
  - 样本矩 $m_1 = (X_1 + \ldots X_n)/n, m_2 = (X_1^2 + \ldots X_n^2)/n$
  - 令 $m_1 = \lambda, m_2 = \lambda + \lambda^2$，解得 $\hat{\lambda}_1 = \overline{X}_n, \hat{\lambda}_2 = \sqrt{S_n^2 - \overline{X}_n} - 1/4$

不唯一

贝叶斯估计 (Bayes estimator)
如何利用先验知识？

- 先验知识：圆周率大约是3.14、地球重力加速度大约是9.8、男女比大约是1:1
  - 布丰投针重复执行若干次，MLE的结果是 3.1。应该宣布估计$\pi = 3.1$吗?
- 频率学派：直接让可选的$\pi \in [3.14,3.15]$，MLE修正为 $\pi = 3.14$
- 贝叶斯学派：假设π服从与3.14有关的**先验概率分布** $\Pr(\pi = \hat{\pi})$。
  - 通过**后验概率分布** $\Pr(\pi = \hat{\pi} \mid X = x)$ 选择最好的估计值 $\hat{\pi}$
  - 给定数据 x，已知 $\Pr(X = x \mid \pi = \hat{\pi})$ 和 $\Pr(X = x)$,
  - 贝叶斯定理: $\Pr(\pi = \hat{\pi} \mid X = x) = \dfrac{\Pr(X = x \mid \pi = \hat{\pi}) \Pr(\pi = \hat{\pi})}{\Pr(X = x)}$

(possibly portrait)
Thomas Bayes
（1701-1761）

贝叶斯估计 (Bayes estimator)
如何利用先验知识？

- 贝叶斯学派：**假设**π服从与3.14有关的**先验概率分布** $\Pr(\pi = \hat{\pi})$。
- 先验分布应该反应现实情况
  - 实践中的选择其实相当**任意**
  - 选择**适应面较广**的分布，通过调参具体确定
- 应该方便计算
  - 我们需要计算 $\Pr(\pi = \hat{\pi} \mid X = x) = \dfrac{\Pr(X = x \mid \pi = \hat{\pi}) \Pr(\pi = \hat{\pi})}{\Pr(X = x)}$
- 共轭（conjugacy）*
  - 共轭分布：$P(\pi)$ 与 $P(\pi \mid x)$ 同分布族，$P(\pi)$ 是 $P(x \mid \pi)$ 的共轭先验。
  - 注意到 $P(\pi \mid x) \propto P(x \mid \pi) \cdot P(\pi)$, 选择 $\hat{\pi}$ 只需要比较 $P(x \mid \pi)P(\pi)$
  - 估计是为了预测，直接算后验预测分布 $P(\tilde{x} \mid x) = \int_{\pi} P(\tilde{x} \mid \pi)P(\pi \mid x)\mathrm{d}\pi$？

贝叶斯估计 (Bayes estimator)

$$\mathrm{B}(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t$$

$$\sim\sqrt{2\pi}\frac{x^{x-1/2}y^{y-1/2}}{(x+y)^{x+y-1/2}}$$

共轭先验分布*

- 假设样本 $X_1,\ldots,X_n$ 服从某伯努利分布 $p=\theta$，令 $Y=\Sigma_i X_i\sim \mathrm{Bin}(n,p)$
- 伯努利分布和二项分布的共轭先验分布是贝塔分布
  - 假设先验分布 $P(\theta)=\theta^{\alpha-1}(1-\theta)^{\beta-1}/B(\alpha,\beta)$，$\alpha,\beta$ 是可灵活调节的参数
- $P(y)=\int_0^1 P(y\,|\,\theta)P(\theta)\,\mathrm{d}\theta=\binom{n}{y}B(\alpha+y,n+\beta-y)/B(\alpha,\beta)$
- $P(\theta\,|\,y)=\dfrac{P(y,\theta)}{P(y)}=\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}/B(\alpha+y,n+\beta-y)$
- $P(\theta\,|\,y)\propto P(y\,|\,\theta)\cdot P(\theta)=\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}/B(\alpha,\beta)$
- “给定当前数据，最有可能的那个 $\hat{\pi}$”: $\mathrm{argmax}_{\hat{\pi}}\Pr(\pi=\hat{\pi}\,|\,X=x)$
- Maximum A Posteriori (MAP) 最大后验: $\hat{\theta}=\dfrac{y+\alpha-1}{n+\alpha+\beta-2}$

当数据量小的时候，
先验知识能辅助估计

贝叶斯估计 (Bayes estimator)

$$\mathrm{B}(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t$$

$$\sim\sqrt{2\pi}\frac{x^{x-1/2}y^{y-1/2}}{(x+y)^{x+y-1/2}}$$

共轭先验分布*

- 假设样本 $X_1,\ldots,X_n$ 服从某伯努利分布 $p=\theta$，令 $Y=\Sigma_i X_i\sim \mathrm{Bin}(n,p)$
- 伯努利分布和二项分布的共轭先验分布是贝塔分布
  - 假设先验分布 $P(\theta)=\theta^{\alpha-1}(1-\theta)^{\beta-1}/B(\alpha,\beta)$，$\alpha,\beta$ 是可灵活调节的参数
- $P(y)=\int_0^1 P(y\,|\,\theta)P(\theta)\,\mathrm{d}\theta=\binom{n}{y}B(\alpha+y,n+\beta-y)/B(\alpha,\beta)$
- $P(\theta\,|\,y)=\dfrac{P(y,\theta)}{P(y)}=\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}/B(\alpha+y,n+\beta-y)$
- $P(\theta\,|\,y)\propto P(y\,|\,\theta)\cdot P(\theta)=\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}/B(\alpha,\beta)$
- 后验预测分布$P(\tilde{x}\,|\,y)=\int_0^1 P(\tilde{x}\,|\,\theta)P(\theta\,|\,y)\mathrm{d}\theta\quad=\dfrac{y+\alpha}{n+\alpha+\beta}$

如何系统化地比较、评价估计量

- 最大似然估计（MLE）
  - 哪个参数最有可能使得当前数据出现
  - 直观，适合大样本，有时不易计算
- 矩估计（MOM）
  - 大数定律：样本矩会收敛到总体矩
  - 容易计算，有时不太准确，甚至自相矛盾
- 最大后验估计（MAP）
  - 给定当前数据，最有可能的那个参数
  - 引入先验知识的辅助，适合小样本，经常难以计算

相合/一致（Consistency）

- 大数定律：$\bar{X}_n \to \mathbb{E}[X]$ as $n \to \infty$
- 估计值为什么不能也有这样的性质：数据越多、估计越准
  - (弱)相合：$\hat{\theta} \xrightarrow{P} \theta$
  - 强相合*：$\hat{\theta} \xrightarrow{a.s.} \theta$
  - r阶矩相合*：$|\hat{\theta} - \theta|^r \to 0$
- 中心极限定理：$\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1)$ as $n \to \infty$
  - 估计值也能收敛到正态分布：渐近正态性 (Asymptotic normality)

无偏（Unbiasedness）

Friedrich Bessel
(1784-1846)

- 假设样本 $X_1, \ldots, X_n$ 服从某分布，试估计方差
- 二阶样本中心距 $S_n^2 = \Sigma_i (X_i - \bar{X})^2 / n$，大数定律 $S_n^2 \to \mathbb{E}[(X-\mu)^2]$
- $\mathbb{E}[S_n^2] = \Sigma_i (\mathbb{E}[X_i^2] - 2\mathbb{E}[\bar{X}_n X_i] + \mathbb{E}[\bar{X}_n^2])/n = \mathbb{E}[X^2] - \mathbb{E}[\bar{X}_n^2]$

$$= \mathbb{E}[X^2] - \left(\mathbb{E}[\Sigma_i X_i^2] + \Sigma_{i \neq j} \mathbb{E}[X_i X_j] \right) / n^2 = \frac{n-1}{n} (\mathbb{E}[X^2] - \mathbb{E}[X]^2) = \frac{n-1}{n} \sigma^2$$

- **渐近无偏**估计量S_n^2；**无偏**估计量**样本方差**：$S^2 = \frac{n}{n-1} S_n^2 = \frac{\Sigma_i (X_i - \bar{X})^2}{n-1}$

无偏（Unbiasedness）

- 假设样本 $X_1, \ldots, X_n$ 服从某分布，试估计方差
- **渐近无偏**估计量：$\mathbb{E}[S_n^2] \to \sigma^2$；**无偏**估计量：$\mathbb{E}[S^2] = \sigma^2$
- bias 偏差: $\mathbb{E}[\hat{\theta}] - \theta$
- average absolute deviation 平均绝对误差: $\mathbb{E}[\,|\hat{\theta} - \theta|\,]$
- mean square error (MSE) 均方误差: $\mathbb{E}[(\hat{\theta} - \theta)^2]$

贝叶斯估计 (Bayes estimator)
如何利用先验知识？

- 贝叶斯学派：假设π服从与3.14有关的**先验概率分布** $\Pr(\pi = \hat{\pi})$。
- 通过**后验概率分布** $\Pr(\pi = \hat{\pi} \mid X = x)$ 选择估计值 $\hat{\pi}$
  - Maximum A Posteriori (MAP) 最大后验估计
    - “ 给定当前数据，<u>最有可能</u>的那个 $\hat{\pi}$ ”: $\operatorname{argmax}_{\hat{\pi}} \Pr(\pi = \hat{\pi} \mid X = x)$
  - Posterior median 后验中位数估计
    - 平均绝对误差最小: $\operatorname{argmin}_{\hat{\pi}} \mathbb{E}\left[|\pi - \hat{\pi}| \,\middle|\, x \right]$
  - Least Mean Squares (LMS) 最小均方估计
    - 最小均方误差(MMSE)估计: $\operatorname{argmin}_{\hat{\pi}} \mathbb{E}[(\pi - \hat{\pi})^2 \mid x]$

(possibly portrait)
Thomas Bayes
（1701-1761）

有效（Efficiency）

- 估计量是关于样本的函数 => 随机变量的函数 => 随机变量
  - 估计量有方差
  - 方差更小的无偏估计量更有效 (efficient)
- 最小方差无偏估计（**MVUE**，minimum-variance unbiased estimator）
  - 若**无论** θ 的取值，$\hat{\theta}(X;\theta)$ 都是 MVUE，则它是一致最小方差无偏估计（**UMVUE**，Uniformly Minimum-Variance Unbiased Estimator）
  - UMVUE 唯一

Cramér–Rao bound* & Fisher information*

- **Cramér–Rao bound**: 假设样本 $X_1, \ldots, X_n$ 服从某分布，pdf 为 $f(x;\theta)$。若 $\hat{\theta}(X_1, \ldots, X_n)$ 是 θ 的无偏估计量，则

$$\operatorname{Var}(\hat{\theta}) \geq \frac{1}{n \cdot I(\theta)},$$

  其中 $I(\theta)$ 是该概率分布关于未知参数 θ 的费希尔讯息数（Fisher Information）

- $I(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2 \,\middle|\, \theta\right] = \int_{-\infty}^{+\infty}\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2 f(x;\theta)\,dx$
- Cramér–Rao bound不一定可达

Cramér–Rao bound* & Fisher information*

- **Cramér–Rao bound**: 假设样本 $X_1, \ldots, X_n$ 服从某分布，pdf 为 $f(x;\theta)$。若 $\hat{\theta}(X_1, \ldots, X_n)$ 是 θ 的无偏估计量，则

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{n \cdot I(\theta)},$$

  其中 $I(\theta)$ 是该概率分布关于未知参数 θ 的费希尔讯息数（Fisher Information）
- 若 $\hat{\theta}$ 是 θ 的无偏估计
  - 有效估计 (efficient estimator)：$\mathrm{Var}(\hat{\theta}) = 1/nI(\theta)$ 效率为 1
  - $\hat{\theta}$ 的效率(efficiency)：$e_n(\hat{\theta}) = 1/nI(\theta)\mathrm{Var}(\hat{\theta}) \in [0,1]$
  - 渐近有效（asymptotically efficient）估计：$e_n(\hat{\theta}) \to 1$ as $n \to \infty$

充分统计量（sufficient statistic）*

- 假设样本 $X_1, \ldots, X_n$ 服从某伯努利分布 $p = \theta$，令 $Y = \Sigma_i X_i \sim \mathrm{Bin}(n, \theta)$
- 已知Y，再获得别的信息，比如 X_1, X_2, X_n，能更好地帮助我们估计 θ 吗?
- 若给定统计量 $T(X_1, \ldots, X_n)$，样本 $(X_1, \ldots, X_n)$ 与 参数 θ 无关，则统计量 T 是参数 θ 的**充分统计量**。
  - “ 函数 T 包含了样本中关于 θ 的全部信息 ”
  - $\Pr[X = x \mid T] = \Pr[X = x \mid T, \theta]$
  - 马尔科夫链：参数 → 充分统计量 → 样本

充分统计量（sufficient statistic）*

因子分解定理

- 若给定统计量 $T(X_1, \ldots, X_n)$，样本 $(X_1, \ldots, X_n)$ 与 参数 θ 无关，则统计量 T 是参数 θ 的充分统计量。
- Fisher–Neyman **因子分解定理**：令 f 是似然函数/pdf/pmf，统计量 T 是参数 θ 的充分统计量当且仅当存在非负函数h, g 使得

$$f(x;\theta) = h(x) \cdot g(\theta, T(x))$$

  - 概率分布可以被分解为两个函数的乘积，其中一个与 θ 无关；另一个仅通过 T 与样本 x 产生关联。
  - “函数 T 包含了样本中关于 θ 的全部信息”

参数 → 充分统计量 → 样本

充分统计量（sufficient statistic）*

最小充分统计量

- 假设样本 $X_1, \ldots, X_n$ 服从某伯努利分布 $p = \theta$，令 $Y = \Sigma_i X_i \sim \mathrm{Bin}(n, \theta)$
- 若给定统计量 $T(X_1, \ldots, X_n)$，样本 $(X_1, \ldots, X_n)$ 与 参数 θ 无关，则统计量 T 是参数 θ 的充分统计量。
- 充分统计量不唯一：$Y = \Sigma_i X_i$、$\bar{X}_n = \Sigma_i X_i / n$、$(X_1, \ldots, X_n)$
- 如果统计量 $S(X)$ 是 θ 的充分统计量，且对于任意 θ 的充分统计量 $T(X)$ 都存在函数 g 使得 $S(X) = f(T(X))$，则 $S(X)$ 是**最小充分统计量**(Minimal sufficiency)
  - 最小充分统计量通常存在但不总是存在

充分统计量（sufficient statistic）*

- 若给定统计量 $T(X_1, \ldots, X_n)$，样本 $(X_1, \ldots, X_n)$ 与 参数 θ 无关，则统计量 T 是参数 θ 的充分统计量。
- 对于任意估计量 $\hat{\theta}_1(X)$，利用充分统计量的新估计量 $\hat{\theta}_2 = \mathbb{E}[\hat{\theta}_1(X) \mid T(X)]$
- Rao–Blackwell–Kolmogorov定理：

$$\mathbb{E}[(\hat{\theta}_2(X) - \theta)^2] \leq \mathrm{E}[(\hat{\theta}_1(X) - \theta)^2]$$

  - 利用充分统计量可以得到均方误差（MSE）更小的估计量。
- Lehmann–Scheffé 定理：若 $\hat{\theta}$ 是无偏估计量，T是完备（complete）充分统计量，则 $\mathbb{E}[\hat{\theta} \mid T]$ 是唯一的一致最小方差无偏估计（UMVUE）。
  - 不严格地，完备统计量只包含了样本中关于目标参数的信息，不含其他信息

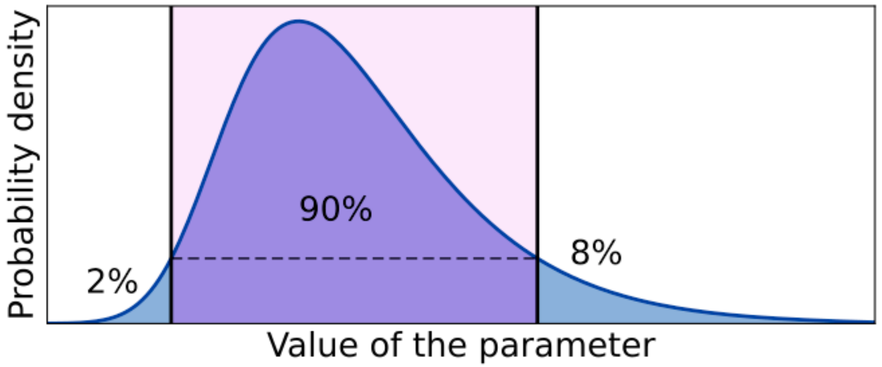
置信区间（confidence interval）

- 频率学派：估计量 $\hat{\theta}$ 是关于随机样本 $X_1, \ldots, X_n$ 的函数
- 贝叶斯学派：后验概率 $P(\theta \mid x)$

} 随机的估计量

- 估计量是随机变量、估计量具有方差、很可能任意估计量都不太会对
- 假设样本 $X_1, \ldots, X_n$ 服从某伯努利分布 $p = \theta$，令 $Y = \Sigma_i X_i \sim \mathrm{Bin}(n, \theta)$
  - Hoeffding不等式：$\Pr[\,|Y - np| \geq t] \leq 2 \cdot \exp(-2t^2/n)$
  - 令 $t = \sqrt{n \ln(1/10)/2}$，$\Pr[p = Y/n \pm t] = 0.95$
  - 我们可以宣布，以 95% 的概率，$p = Y/n \pm O(\sqrt{n})$！
  - $Y/n \pm O(\sqrt{n})$ 是一个随机的区间，取决于随机样本 $X_1, \ldots, X_n$

可信区间（credible interval）



- 频率学派：估计量 $\hat{\theta}$ 是关于随机样本 $X_1, \ldots, X_n$ 的函数
- **贝叶斯**学派：**后验概率** $P(\theta \mid x)$
- 估计量是随机变量、估计量具有方差、很可能任意估计量都不太会对
- 从后验分布中选择一段区间 $[a, b]$ 使得

$$\Pr[\theta \in [a, b] \mid X = x] = 1 - \alpha$$

- 多种选法*：
  - 选择最短的区间：highest density interval (HDI)
  - 基于分位数quantile-based interval (QBI)：如从中位数左右各取 $(1 - \alpha)/2$

区间估计（interval estimation）

- （原则上）给定置信度 $1-\alpha$，任意统计量 $a(X_1,\ldots), b(X_1,\ldots)$，使得

$$\Pr[a \le \theta \le b] \ge 1-\alpha$$

- 枢轴变量（pivot/pivotal quantity）法：假设样本 $X_1,\ldots,X_n$ 服从 $N(\mu,1)$,
  - 统计量 $\bar{X}_n \sim N(\mu,1/n)$，枢轴变量 $\bar{X}_n-\mu \sim N(0,1/n)$
  - 寻找分位数 $[c,d]$ 使得 $\Pr\left[\bar{X}_n-\mu \in [c,d]\right] \ge 1-\alpha$
  - 置信水平 $1-\alpha$ 的置信区间 $\left[\bar{X}_n-d, \bar{X}_n-c\right]$

68-95-99.7 (3σ) 经验法则