Maxwell General Learning System Academic Statistics Complex
麦克斯韦通用高等习得系统综合统计学

Leo_Maxwell

March 2023

目录

统计学摘要

人们为什么要发展这门科学？人们试图达成什么目的?

1. 描述、总结通过统计获得的数据，发现其中的特征和规律。
2. 通过从总体中的一小部分（**样本**）中获得的**有限的**数据去**推知总体的数据特征**。
3. 对比**总体中的不同样本之间的数据特征**有没有**显著差异**。
4. 对比**不同作用方式在相同对象上产生的作用**有没有**显著差异**。

这些是统计分析的主要作用。后文我们所做的所有讨论，都是在这几个主要框架下进行的。文中大部分的统计量，都是为了满足这些目的而生。

本书需要分析学的基本知识，例如 Lebesgue 测度、Taylor 展开等。

第一章　随机变量及其分布 (Random Variable, RV) and its (Distribution)

一个随机取值的数字，它不代表任何一个具体、固定的数字，而是一个表示随机量的对象。然而这不同的随机量，其取什么值的概率，也有所不同，可能落在什么区间之间的概率大小，更是千奇百怪。这便是量化统计学的最基本、最基础的定义和对象。这个值在数学中一般用大写字母表示。正是因为它的随机性不同，所以才有许多描述这个值的属性。这个值有可能只能取某些固定的数字，有可能取遍某个区间内的所有实数，后文中不少量便是继承这个对象的属性而来的。

1.1　连续 (Continuous) 或离散 (Discrete)

投掷一枚标准骰子，它给出的值便只有六个整数，从一取到六。假若用 X 代表投掷一枚骰子所得点数，那么这 X 就是一个随机变量，并且是离散型的。但如果用 X 表示某一地的年降雨量，那这降雨量若是测量准确，则必定要取遍一个合理区间内的全体实数了，此时称 X 是连续型随机变量。在下文中的各种随机变量的属性，如无特别说明，属性对离散型或连续型随机变量都适用。

1.2　分布 Distribution

分布就是随机变量等于某个值（离散型）或落在某个区间内（连续型）的概率大小的函数。假设这个函数是 $f(x)$，则这个随机变量等于某个值 a 的概率就是 $f(a)$（离散型），落在区间 $[a,b]$ 上的概率就是 $\int_a^b f(x)$（连续型）。

容易理解的是，任何一个连续随机变量落在一个特定值上的概率都是零，与公式符合。所以要探究一个连续型随机变量取值的概率，一般都考虑对应在哪一个区间。

1.2.1　概率质量函数 (Probability Mass Function, PMF) 和概率密度函数 (Probability Density Function, PDF)

PMF 是针对离散型随机变量而言的，它详细列举了离散型随机变量取每一个值对应的概率的大小。而 PDF 是针对连续型随机变量而言的，对它的某个区间进行积分就能获得该随机变量落在该区间内的概率大小。

1.2.2　累计分布函数 (Cumulative Distribution Function, CDF)

实际的数学计算中还是累计分布函数用得更多一些。它所描述的是某个随机变量落在小于某个具体数字的概率。对于一个随机变量 X 而言，$\mathrm{CDF}(c) = \mathrm{P}(X \le c)$，故它一定是单调增、不小于零的。利用这个定义，要计算一个随机变量落在某个区间内的概率，只需要用两个 CDF 函数相减即可，免去了积分的麻烦。

分布函数一定右连续。

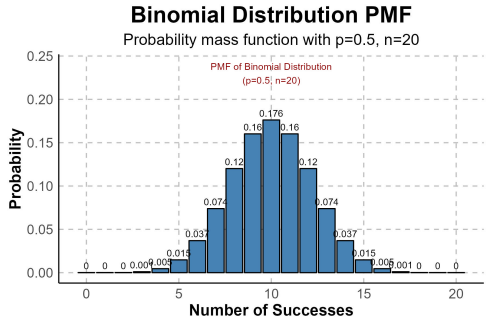

图 1.1: Standard Normal Distribution PDF

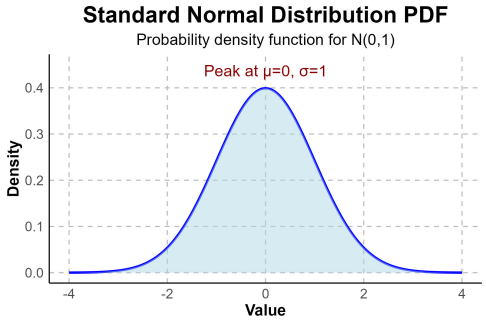

图 1.2: Standard Normal Distribution PDF

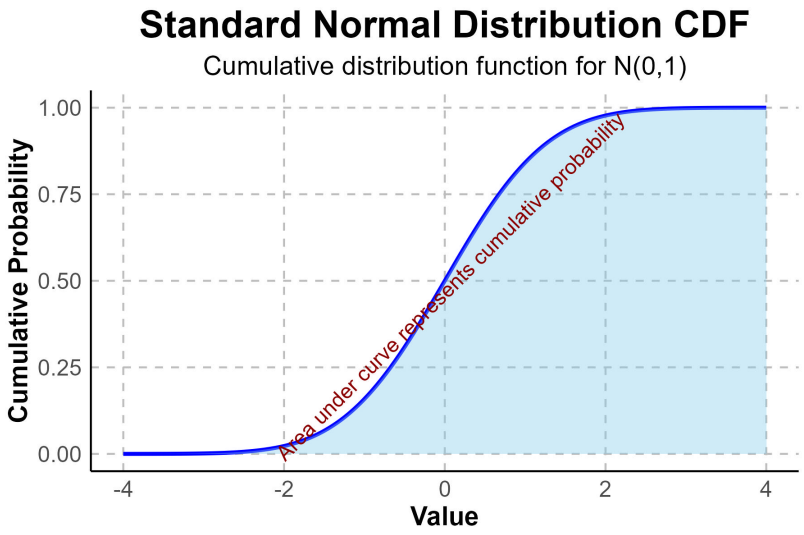

图 1.3: Standard Normal Distribution CDF

1.2.3 分位数 (Quantile)

分位数是 CDF 的反函数。CDF 接收一个随机变量可能取的具体数字，给出随机变量取值小于这个数字的概率。但若已经知道一个概率值，想要知道随机变量的取值小于哪一个数字的概率等于这个概率值，这个数字便称作分位数。分位数在后文的假设检验中常常用到。

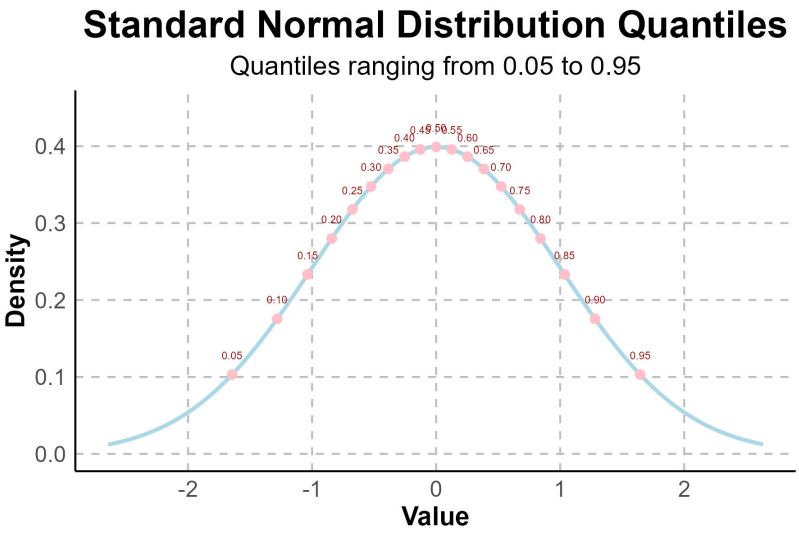

图 1.4: Standard Normal Distribution Quantiles

1.2.4 特征函数 (Characteristic Function)

特征函数是用来描述一个随机变量的函数。虽然前文介绍的分布相关的函数对一个随机变量已经作了很好的描述，特征函数在作部分概率相关的证明时非常有用。对于一个随机变量 X，它的特征函数是

$$\varphi_X(t) = \mathrm{E}[e^{itX}]$$

其中，i 是虚数单位。它是一个和 t 有关的函数，这个自变量某种程度上类似于 CDF 的自变量。它完全确定了随机变量 X 的行为和分布。

1.3 期望 (Mean)

期望，也称均值，描述的是重复足够多次取一个随机变量的值，这些值的均值是多少。对随机变量 X 而言，它的期望被表示为 $\mathrm{E}[X]$. 它的数学定义是：

离散型： $\mathrm{E}[X] = \sum_{i=1}^{n} x_i f(x_i)$

连续型： $\mathrm{E}[X] = \int_{\mathbb{R}} x_i f(x_i) \mathrm{d}x$

随机变量的期望有以下性质（证明见5.2.1）：

常数之期望为其本身： 对任意常数 c，都有 $\mathrm{E}[c] = c$.

期望是线性函数： $\forall a, b \in \mathbb{R}$，都有 $\mathrm{E}[aX + bY] = a\mathrm{E}[X] + b\mathrm{E}[Y]$

两个随机变量的积的期望： 对两个随机变量 X 和 Y 而言，有 $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y] + \mathrm{Cov}(X, Y)$，其中 $\mathrm{Cov}(X, Y)$ 是它们的协方差。协方差的定义在后文给出。

关于随机变量的期望，有一个重要的定律：（证明请见5.2.2）

定理 1.3.1 (无意识统计学家法则 (Law of the Unconscious Statistician, LUTOS)) 对随机变量 X 而言，其函数的期望 $\mathrm{E}[g(X)]$ 满足：

$$\mathrm{E}[g(X)] = \int_{\mathbb{R}} g(x)f(x)\mathrm{d}x \qquad \text{连续型（要求 } g \text{ 可逆且可微）}$$

$$\mathrm{E}[g(X)] = \sum_{i=1}^{n} g(x)f(x) \qquad \text{离散型}$$

1.4 方差 (Variance) 与标准差 (Standard Deviation)

方差和标准差都是用于衡量随机变量离散程度的指标，即若某个随机变量能在很大的一个区间上处处取值，且每处概率还不低的话，它的方差和标准差就会很大，反之则小。方差的计算方式是将随机变量可能取到的所有值减去期望的差取平方，然后再相加。若用 X 和 Y 表示两个随机变量，则方差表示为 $\mathrm{Var}(X)$，也常常被记为 σ^2. 方差的数学定义是：

$$\mathrm{Var}(X) = \mathrm{E}[(X - (\mathrm{E}[X])^2]$$

要计算方差的大小，除了定义的方法，还常常用这个公式，它的推导是平凡的：

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$$

标准差是方差的算术平方根，常常记为 σ.
以下是方差计算的性质：（证明见5.2.3）

常数的离散程度为零： $\mathrm{Var}(c) = 0$

随机变量整体的平移不影响其离散程度： $\mathrm{Var}(X + a) = \mathrm{Var}(X)$

伸缩后的随机变量离散程度增加： $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$

多个随机变量的方差： $\mathrm{Var}\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i,j}^{n} a_ia_j\mathrm{Cov}(X_i, X_j)$

1.5 协方差 (Covariance)

对于两个随机变量 X 和 Y 而言，若观测到其中一个的值偏大通常意味着另一个值的观测值偏大（或偏小），则认为这两个随机变量之间存在关联。协方差是用于描述这种关联性的指标之一。它的数学定义是：

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$$

化简后得到：（过程可见5.2.1）

$$\mathrm{Cov}(X, Y) = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]$$

它的运算性质：（证明见5.2.4）

和任意常数的协方差都为零： $\mathrm{Cov}(X, c) = 0$

自身和自身的协方差就是方差： $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$

协方差没有顺序性： $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$

随机变量的伸缩对协方差是线性的： $\mathrm{Cov}(aX, bY) = ab\mathrm{Cov}(X, Y)$

随机变量的平移不影响协方差： $\mathrm{Cov}(X + a, Y + b) = \mathrm{Cov}(X, Y)$

多个随机变量的协方差： $\mathrm{Cov}(aX+bY, cW+dV) = ac\mathrm{Cov}(X, W)+ad\mathrm{Cov}(X, V)+bc\mathrm{Cov}(Y, W)+bd\mathrm{Cov}(Y, V)$

1.6 Markov 不等式 (Markov's Inequality)

对于非负随机变量 X 和 $a \in (0, +\infty)$，则：（证明见5.2.5）

$$P(X > a) \leq \frac{\mathrm{E}[X]}{a}$$

1.7 Chebyshev 不等式 (Chebyshev's Inequality)

对于存在有限期望值 μ 和有界非零方差 σ^2 的随机变量 X 以及任意 $k \in (0, +\infty)$，有：（证明见5.2.6）

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

1.8 依概率收敛 (Convergence in Probability)

取一个随机变量构成的序列 $X_1, X_2, X_3, \ldots, X_n$，再取一个随机变量 X。若 $\forall \varepsilon > 0$，都有：

$$\lim_{n \to +\infty} P(|X - X_n| \geq \varepsilon) = 0$$

则称这列随机变量依概率收敛到 X，记作 $X_n \xrightarrow[n \to +\infty]{P} X$.

1.9 几乎必然收敛 (Almost Sure Convergence)

取一个随机变量构成的序列 $X_1, X_2, X_3, \ldots, X_n$，再取一个随机变量 X。若：

$$P(\lim_{n \to +\infty} X_n = X) = 1$$

则称这列随机变量几乎必然收敛到 X，记作 $X_n \xrightarrow[n \to +\infty]{\text{a.s.}} X$. 它也常常被称为“以概率 1 收敛”或者“强收敛”。

1.10 常见分布 Common Distributions

名称 (Name)	PMF/PDF	期望 (Mean)	方差 (Variance)
伯努利分布 (Bernoulli Dist.)	$p(x)=\begin{cases}p & x=1\\ 1-p & x=0\end{cases}$	p	$p(1-p)$
二项分布 (Binomial Dist.)	$p(x)=C_n^x p^x(1-p)^{n-x}$	np	$np(1-p)$
几何分布 (Geometric Dist.)	$p(x)=p(1-p)^{x-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
泊松分布 (Poission Dist.)	$p(x)=\frac{e^{-\lambda}\lambda^x}{x!}$	λ	λ
均匀分布 (Uniform Dist.)	$\begin{cases}\frac{1}{b-a} & \text{if } a<x<b\\ 0 & \text{otherwise}\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
指数分布 (Exponential Dist.)	$f(x)=\lambda e^{-\lambda x}$, for $x>0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
标准正态分布 (Std. Normal Dist.)	$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$	0	1
正态分布 (Normal Dist.)	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ

1.11 人造的分布 Artificial Distributions

有一些分布是在统计学发展的过程中，人们为了方便假设检验等等统计过程，用数学方法造出来的。下面介绍常用的几个人造的分布。

1.11.1 卡方分布 (Chi-Square Distribution)

定义：若 Z 符合标准正态分布，令 $X=\sum_{i=1}^{n}Z_i^2$，有 $X\sim\chi_n^2$，称 X 满足自由度为 n 的 χ^2（卡方）分布。X 的期望是 n，方差是 $2n$.

卡方分布具有可加性。对于 n 个分别满足自由度为 m_i 的卡方分布的随机变量 X_i，有 $\sum_{i=1}^{n}X_i\sim\chi^2(\sum_{i=1}^{n}m_i)$.

1.11.2 学生 T 分布 (Student T Distribution)

定义：若 Z 符合标准正态分布，$X\sim\chi_n^2$，令 $T=\frac{Z}{\sqrt{\frac{X}{n}}}$，则称 T 符合自由度为 n 的 T 分布。自由度若是很大或趋于无穷，则学生 T 分布趋于标准正态分布。

1.11.3 F 分布 (F Distribution)

定义：若 W_1 和 W_2 是两个相互独立的随机变量，分别符合自由度为 v_1 和 v_2 的卡方分布，令 $F=\frac{W_1/v_1}{W_2/v_2}$，则称 F 符合分子自由度为 v_1，分母自由度为 v_2 的 F 分布，记为 $F(v_1,v_2)$。F 分布是非对称的，是右偏分布。

由定义立得若 $F \sim F(n_1, n_2)$，则 $\dfrac{1}{F} \sim F(n_2, n_1)$. 若 $X \sim T(n)$，则 $X^2 \sim F(1, n)$.

1.12 分布族 Distribution Family

分布族一般被分为两类：**参数分布族 (Parametric Distribution Family)** 和**非参数分布族 (Nonparametric Distribution Family)**，其中，参数分布族的定义是：

分布族含有有限个未知的实参数，表示为

$$\mathcal{F} = \{f(\vec{\boldsymbol{x}}; \vec{\boldsymbol{\theta}}) : \theta \in \Theta\}$$

其中，f 是 PMF 或 PDF，$\vec{\boldsymbol{\theta}}$ 是有限多个未知参数，Θ 是所有可能的参数构成的集合，称为参数空间 (Parameter Space)。

1.12.1 指数型分布族 Exponential Family

对参数族 $\mathcal{F} = \{F(\vec{\boldsymbol{x}}; \vec{\boldsymbol{\theta}}) : \theta \in \Theta\}$，如果分布族中的所有分布都能表示为如下形式：

$$f(x) = c(\theta)\exp\left\{\sum_{j=1}^{k} c_j(\theta)T_j(x)\right\}h(x)$$

其中，$k \in \mathbb{N}$，$c(\theta) > 0$ 且它和 $c_i(\theta)$ 是定义在参数空间 Θ 上的函数，$h(x) > 0$ 且它和 $T_i(x)$ 都是 x 的函数，并且 $T_i(x)$ 之间线性无关，则称该分布族是指数型分布族，简称为指数族。

指数族的支撑集（Support），即 x 的定义域一定和参数 θ 无关。反之，支撑集和参数 θ 有关的分布族一定不是指数族。常见的大多数分布，如正态分布、二项分布、Gamma 分布都是指数族，但均匀分布不是指数族。

来自指数族分布的样本，其联合分布仍是指数族。

1.12.2 Pitman-Koopman 定理

第二章　样本 (Sample) 与总体 (Population)

总体，是真实的、客观的一个集合，包括了所有欲观测的变量，例如一国所有国民之身高，则称为总体。但是在实际统计场景下，绝大多数时候都不可能获得获知总体，而只能获知总体中的一个子集，称作是样本。

2.1　总体和样本的数学结构 Mathematical Structure of Population and Sample

总体从数学上来说是一**大堆数字**，但是通常来说人们感兴趣的总体（即人们感兴趣的一堆数字）符合一定的规律。所以我们会用大写拉丁字母 X 来代表这一堆数字。在不给出任何其他条件的情况下，X 不是一个固定的数字，而是前文中提到过的随机变量。

样本则略有不同。样本虽然也是一些数字，但是其中每一个数字都是从总体中**随机抽取**，所以

- 在抽取前，样本是许多个随机变量 $X_1, X_2, X_3, \ldots, X_n$
- 在抽取后，样本是许多个固定数字 $x_1, x_2, x_3, \ldots, x_n$

2.2　得到有效的样本 Obtaining Valid Sample

没有抽样就没有统计学。所以我们说，抽样是统计学的重要一环。在上述样本的定义下，如果 $X_1, X_2, X_3, \ldots, X_n$ 满足：

- 相互独立
- 和总体 X 服从同一分布

则称该样本为从总体中得到的容量为 n 的**简单随机样本**，抽取简单随机样本的抽样方法称为**简单随机抽样 (Simple Random Sampling, SRS)**。为了能抽取出简单随机样本，这种抽取方法必须满足：

- 代表性：总体中的每个个体有同等机会被选中（即必要时需要放回抽取）
- 独立性：每次抽样的结果互不影响

若无特殊说明，后文中的所有样本都是简单随机样本。

2.3　统计量和估计量 Statistics and Estimators

总体的各类属性如均值、方差等等被称为是**参数 (Parameter)**，所有参数的可能取值构成**参数空间 (Parameter Space)**。

而由样本算出的量（即样本的函数）被称为**统计量 (Statistics)**，用来估计参数的统计量被称为**估计量 (Estimator)**。由于样本具有随机性，所以统计量也具有随机性，其分布是用统计量进行统计推断的依据。

对一列来自总体 X 的样本 $Y_1, Y_2, Y_3, \ldots, Y_n$，该样本的统计量有如下几种。

2.3.1 次序统计量 (Order Statistics)

将其从小到大排列得到一列有序的样本：

$$Y_{(1)} \leq Y_{(2)} \leq Y_{(3)} \leq \cdots \leq Y_{(n)}$$

则称 $Y_{(1)}, Y_{(2)}, Y_{(3)}, \ldots, Y_{(n)}$ 为样本的次序统计量，其中 $Y_{(k)}$ 为第 k 个次序统计量。

2.3.2 样本极值 (Sample Extremum)

极小值 $Y_{(1)}$，极大值 $Y_{(n)}$，都称为是样本极值。

2.3.3 样本极差 (Sample Range)

定义样本极差 R_n 为：

$$R_n = Y_{(n)} - Y_{(1)}$$

2.3.4 样本中位数 (Sample Median)

定义样本中位数 $m_{0.5}$ 为：

$$m_{0.5} = \begin{cases} x_{(\frac{n+1}{n})} & n\text{为奇数} \\ \frac{1}{2}\left[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right] & n\text{为偶数} \end{cases}$$

2.3.5 样本 p 分位数 (Sample pth Quantile)

定义样本 p 分位数 m_p 为：（此处 $p \in (0, 1]$）

$$m_p = \begin{cases} x_{(np+1)} & np\text{不是整数} \\ \frac{1}{2}\left[x_{(np)} + x_{(np+1)}\right] & np\text{是整数} \end{cases}$$

2.4 样本分布 Sample Distribution

2.4.1 联合分布 (Joint Distribution)

联合分布指的是样本 $Y_1, Y_2, Y_3, \ldots, Y_n$ 取得特定一组值 $y_1, y_2, y_3, \ldots, y_n$ 时的分布。

对于来自总体 X 的样本 $Y_1, Y_2, Y_3, \ldots, Y_n$，如果总体有累积分布函数 $F(x)$，则样本的联合累积分布函数为

$$F^*(y_1, y_2, y_3, \cdots, y_n) = \prod_{k=1}^{n} F(y_k)$$

类似地，若总体 X 有概率密度函数 $f(x)$，则样本的联合概率密度函数是

$$f^*(y_1, y_2, y_3, \cdots, y_n) = \prod_{k=1}^{n} f(y_k)$$

2.4.2 经验分布 (Empirical Distribution)

经验分布指的是样本 $Y_1, Y_2, Y_3, \ldots, Y_n$ 在每个样本都等可能取值时，对总体分布的估计。

对于来自总体 X 的样本 $Y_1, Y_2, Y_3, \ldots, Y_n$，在不知道总体分布的情况下，我们构造一个分布 F_n 来模拟总体分布，该分布的基本假设是每一个样本 Y_i 都是等可能的，所以我们定义经验分布函数 $F_n(x)$ 为：

$$F_n(x) = \begin{cases} 0 & x < y_{(1)} \\ \dfrac{k}{n} & y_{(k)} \le x \le y_{(k+1)} \\ 1 & x \ge y_{(n)} \end{cases}$$

2.5 抽样分布 (Sampling Distribution)

我们称统计量的概率分布为**抽样分布 (Sampling Distribution)**。评价点估计的优劣、构造置信区间、执行后文的假设检验（四）都需要用到抽样分布。令样本量大小为 n，研究的统计量为 T，则有这些不同的抽样分布。

2.5.1 精确分布 (Exact Distribution)

对任意给定的样本量大小 n，统计量 T 的真实分布。

2.5.2 渐进分布 (Asymptotic Distribution)

当 n 趋于无穷时，统计量所趋于的分布。

2.5.3 近似分布 (Approximate Distribution)

没有对 T 进行直接观察，而是通过模拟等手段得到一份近似的统计量 T 的样本的分布。

2.6 充分统计量 (Sufficient Statistics)

所有的统计量，本质上都是在对样本进行加工。我们现在感兴趣的是，在对样本加工的过程中，我们有没有损失样本包含的所有原始信息?

严格地说，令样本是 $\vec{x} = (x_1, x_2, x_3, \ldots, x_n)$，我们感兴趣的未知参数是 θ，统计量（即对样本的一种加工）$T(\vec{x})$ 有没有丢失任何与 θ 有关的信息是我们所感兴趣的问题。统计学上我们把这种性质，即“样本加工不损失信息”称为“充分性”。

那么如何从数学上判断样本加工是否损失信息呢？我们认为有这样的关系：

样本 $\vec{x}$ 中关于 θ 的信息 = 统计量 $T = T(\vec{x})$ 中有关 θ 的信息 + 在 $T(\vec{x})$ 取值为 t 以后样本 $\vec{x}$ 中关于 θ 的信息

于是我们说，如果上式右侧最后一项为零，则 $T(\vec{x})$ 中包含了所有和 θ 有关的信息。使用数学语言描述，即 $F(\vec{x}|T = t)$ 和参数 θ 无关。

2.6.1 充分统计量的性质

充分统计量的一一映射仍是充分统计量。

次序统计量是充分统计量。

2.7 因子分解定理 Factorization Theorem

对参数分布族

$$\mathcal{F} = \{f(\vec{x}) : \theta \in \Theta\}$$

其中，$f(\vec{x})$ 表示 PMF 或 PDF，则对任意定义在样本空间内的统计量 $T(\vec{x})$ 是充分统计量，当且仅当存在定义在统计量 $T(\vec{x})$ 的取值空间上的函数 $g[T(\vec{x})]$，和定义在样本空间内的函数 $h(\vec{x})$，使得

- $h(\vec{x}) \geq 0$
- $f(\vec{x}, \theta) = g[T(\vec{x}), \theta]h(\vec{x})$

即样本分布一定能分解为两个因子的乘积，其中一个是与充分统计量 $T(\vec{x})$ 和参数 θ 有关的函数，另一个是只和样本 $\vec{x}$ 有关的函数。

第三章　点估计 Pointwise Estimate

点估计就是要构造一个合适的统计量，使用这个统计量估计未知参数 θ.

3.1　估计方法 Estimation Methods

3.1.1　矩估计 Method of Matching Moments

矩估计的基本思想是“替代”，即用样本的矩估计总体的矩，或用样本的矩函数估计总体的矩函数。矩估计是一种简单的估计方式，它不要求事先知道总体的分布类型。但是在实际使用中有很多缺点。例如，矩估计一般只能使用样本的一阶矩和二阶矩，更高阶的矩通常不稳定。它没有充分利用总体分布函数所提供的信息，因此难以保证所的估计量有优良的性质。

3.1.2　最大似然估计 Maximum Likelihood Estimation, MLE

最大似然估计的根本思想是找到令当前观测出现的概率最大的未知参数。对于观测到的一列样本 $\vec{x}$ 和未知参数 θ，我们使用一个 $\hat{\theta}$ 估计总体参数 θ，使得似然函数 $L(\vec{x},\theta)$ 取得最大值。

似然函数 $L(\vec{x},\theta)=P(X_1=x_1,X_2=x_2,X_3=x_3,\ldots,X_n=x_n)$，等于：

$$\prod_{i=1}^{n} f(x_i,\theta)$$

其中 $f(x_i,\theta)$ 是总体的 PDF 或 PMF，θ 是和总体分布有关的参数。将 $L(\vec{x},\theta)$ 取自然对数变为 $l(\vec{x},\theta)$，然后再对 θ 求导得到使导数为零的点 $\hat{\theta}$，即为所求的最大似然估计。

最大似然估计一定是充分统计量的函数。

3.1.2.1　最大似然估计给出的其他概念 Concepts in MLE

Score:

定义：我们称

$$s(\theta)=\frac{\partial l(\vec{x},\theta)}{\partial\theta}$$

为 θ 的 Score. Score 的均值为零，而其方差是：

Fisher 信息量：

定义：Score 的方差是 Fisher 信息量，记为 $I(\theta)$。由于 Score 的方差通常计算复杂，一般用以下公式计算：

$$I(\theta)=\mathrm{E}\left[\frac{\partial^2 l(\vec{x},\theta)}{\partial\theta^2}\right]$$

3.1.3　估计量的不变性 Invariance of Statistics

若 $\hat{\theta}$ 是 θ 的矩估计，若函数 g 连续，则 $g(\hat{\theta})$ 也是 $g(\theta)$ 的矩估计。

若 $\hat{\theta}$ 是 θ 的最大似然估计，则对任意函数 g，$g(\hat{\theta})$ 是 $g(\theta)$ 的最大似然估计。

3.2 点估计的评价标准 Criteria about Pointwise Estimate

点估计是一个统计量，统计量是随机变量，所以我们不可能要求它一定等于参数的真值，所以如何评价估计量的好坏是我们关心的问题。

3.2.1 大样本评价标准 Criteria for Large Samples

3.2.1.1 相合性 Consistency

根据 Glivenko-Cantelli 定理（3.5.3），随着样本容量的不断增大，经验分布函数逼近真实分布函数，所以可以要求估计量同样随着样本容量的不断增加而逐渐逼近参数真值。这种行为被称为**相合性（Consistency）**。将这种行为严谨地表达为：

定义：设 $\theta \in \Theta$ 是未知参数，$\hat{\theta}_n = \hat{\theta}_n(x_1, x_2, x_3, \ldots, x_n)$ 是从容量为 n 的样本中获得的估计量，若 $\forall \varepsilon > 0$，都有

$$\lim_{n \to +\infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0$$

则称 $\hat{\theta}_n$ 为 θ 的相合估计。

常见的矩估计都具有相合性，并且矩估计的连续函数也是对应参数的相合估计。

3.2.1.2 渐进正态性 Asymptotic Normality

相合性是对估计量的一种较低的要求，它只要求随着 n 的增大而收敛，没有对收敛的速度进行讨论。渐进正态性对这一点进行了补充，它是在相合性的基础上进行讨论的。

定义：$\hat{\theta}_n = \hat{\theta}_n(x_1, x_2, x_3, \ldots, x_n)$ 是相合估计序列，若存在趋于零的正数列 $\sigma_n(\theta)$，使得：

$$\frac{\hat{\theta}_n - \theta}{\sigma_n(\theta)} \xrightarrow{\mathrm{L}} N(0, 1)$$

则称 $\hat{\theta}_n$ 具有渐进正态性，记为 $\hat{\theta}_n \sim AN(\theta, \sigma_n^2(\theta))$. 其中，$\sigma_n^2(\theta)$ 称为 $\hat{\theta}_n$ 的渐进方差，$\dfrac{\hat{\theta}_n - \theta}{\sigma_n(\theta)}$ 称为规范变量。

上述定义中的数列 $\sigma_n^2(\theta)$ 表示 $\hat{\theta}_n$ 依概率收敛于 θ 的速度。

满足条件的 $\sigma_n(\theta)$ 不唯一。

在有多个相合估计的场合时，它们的渐进正态分布的方差大小常常被用来比较它们的好坏。

在合适的正则条件下，参数 θ 的最大似然估计 $\hat{\theta}_n$ 具有相合性和渐进正态性，并且还有：

$$\hat{\theta}_n \sim AN\left(\theta, \frac{1}{nI(\theta)}\right)$$

指数族中的分布都满足该“合适的正则条件”。

3.2.2 小样本评价标准 Criteria for Small Samples

3.2.2.1 偏差 (Bias)

定义：设用于估计参数 θ 的估计量是 T，则称偏差 (Bias) 为 $b_T(\theta) = \mathrm{E}[T] - \theta$，如果对任意 θ 而言，偏差都为零，则称该估计量 T 是一个**无偏估计 (Unbiased Estimator)**。

无偏性不具有不变性。

3.2.2.2 有效性 (Efficiency)

对于两个无偏估计量 $\hat{\theta}_1$ 和 $\hat{\theta}_2$，如果 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 的方差更小，则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 更有效。

3.2.2.3 均方差 (Mean Squared Error, MSE)

定义：设用于估计参数 θ 的估计量是 T，则称均方差 (MSE) 为 $MSE_T(\theta) = \mathrm{E}[(T-\theta)^2]$.

有性质：$MSE_T(\theta) = \mathrm{Var}(T) + (b_T(\theta))^2$，其证明请见附录5.2.10。

3.2.2.4 总结 Conclusion

一般而言，均方差是评价点估计的一个重要标准，但因为均方差由偏差和方差共同组成，所以想要达到最小的均方差通常意味着在准度和精度之间作取舍，略微增大偏差可能导致方差的大大减少，从而导致均方差的下降，但此时的估计量便不再是无偏的了。

由于这种取舍较为困难和复杂，所以人们转而不去寻找使均方差最小的估计量，而转而在所有的无偏估计中寻找方差最小的估计量了。这么做实际上是用精度换准度，因为我们限制偏差为零，在这个基础上寻找使均方差最小的估计量，即后文的 MVUE.

3.3 最小方差无偏估计 (Minimum Variance Unbiased Estimator, MVUE)

定义：若对于参数 θ，存在它的某个无偏估计 $\hat{\theta}$，使得对于参数 θ 的所有无偏估计 $\hat{T}$，满足：

$$\mathrm{Var}(\hat{\theta}) \leq \mathrm{Var}(\hat{T})$$

则称 $\hat{\theta}$ 是 θ 的最小方差无偏估计（MVUE）。

想要找到 MVUE 并不简单。

3.3.1 Cramér-Rao 不等式 (Cramér-Rao Inequality)

指数族分布全部满足 Cramér-Rao 正则条件。

如果 $\hat{\theta}$ 是 θ 的无偏估计，在 Cramér-Rao 正则条件下，有：

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

其中，我们称 $\dfrac{1}{I(\theta)}$ 为 **Cramér-Rao 下界 (Cramér-Rao Lower Bound, CRLB)**. 当等号成立时，称 $\hat{\theta}$ 是**有效估计 (Efficient Estimator)**. 更一般地，对于参数的函数 $g(\theta)$ 的无偏估计 $\hat{g}(\theta)$，有类似的结果：

设 $\hat{g}(\theta)$ 是 $g(\theta)$ 的无偏估计，且满足 Cramér-Rao 正则条件，则有：

$$\mathrm{Var}[\hat{g}(\theta)] \geq \frac{[g^{'}(\theta)]^2}{I(\theta)}$$

3.3.1.1 Cramér-Rao 正则条件下估计的效率 Efficiency under Cramér-Rao Regularity Conditions

设 $\hat{g}(\theta)$ 是 $g(\theta)$ 的无偏估计，且满足 Cramér-Rao 正则条件，则称：

$$e_n = \frac{[g^{'}(\theta)]^2 / I(\theta)}{\mathrm{Var}(\hat{g}(\theta))}$$

为无偏估计 $\hat{g}(\theta)$ 的**效率 (Efficiency)**.

若 $e_n = 1$，则按照先前的定义，称 $\hat{g}(\theta)$ 是 $g(\theta)$ 的**有效估计 (Efficient Estimator)**. 若 $\lim\limits_{n\to+\infty} e_n = 1$，则称 $\hat{g}(\theta)$ 是 $g(\theta)$ 的**渐进有效估计 (Asymptotically Efficient Estimator)**.

3.3.2 零的无偏估计法 Unbiased Estimator of Zero

这种方法的主要想法很简单，我们想知道一个无偏估计什么时候达到最优（即方差不能够继续减小了）?

我们取一个 θ 的无偏估计 $\hat{\theta}$，现在我们另取一个估计量 U，要求该估计量满足 $\mathrm{E}[U]=0$，称为零的无偏估计。那么任取常数 a，都有 $\mathrm{E}[\hat{\theta}+aU]=\theta$，新的估计量 $\tilde{\theta}=\hat{\theta}+aU$ 仍然是无偏的。

接下来我们想知道，新的估计量 $\tilde{\theta}$ 是否比 $\hat{\theta}$ 更优，即方差是否更小? 我们计算新估计量的方差:

$$\begin{aligned}\mathrm{Var}(\tilde{\theta}) &= \mathrm{Var}(\hat{\theta}+aU)\\ &= \mathrm{Var}(\hat{\theta})+2a\mathrm{Cov}(\hat{\theta},U)+a^2\mathrm{Var}(U)\end{aligned}$$

当 $\mathrm{Cov}(\hat{\theta},U)<0$ 时，$\forall a\in\left(0,-\dfrac{2\mathrm{Cov}(\hat{\theta},U)}{\mathrm{Var}(U)}\right)$，都有 $\mathrm{Var}\tilde{\theta}<\mathrm{Var}(\hat{\theta})$，即新估计量 $\tilde{\theta}$ 更优。

反之，当 $\mathrm{Cov}(\hat{\theta},U)>0$ 时，新估计量不会更优。

3.3.2.1 MVUE 与零偏估计 MVUE and Unbiased Estimator of Zero

参数 θ 的无偏估计 $\hat{\theta}$ 是 MVUE 的充要条件是：$\hat{\theta}$ 与零的所有无偏估计不相关，即

$$\forall U,\mathrm{E}[U]=0\Longrightarrow\mathrm{Cov}(\hat{\theta},U)=0$$

该定理给出了 MVUE 的特征，即它与零的所有无偏估计不相关。这是因为零的无偏估计实质上是随机噪声，即不含有信息。该定理的意义大部分局限于给出 MVUE 的特征，而不是告诉我们如何找到它，因为要验证一个估计量与零的所有无偏估计都不相关很难。

3.4 一般随机样本估计量 Estimators of General Random Sample

设总体 X 的均值是 μ，方差是 σ^2，分布函数（CDF）是 $F(x)$，密度函数（PDF）是 $f(x)$，总体的某一个参数用 θ 表示，对一列来自总体 X 的独立同分布样本 $Y_1,Y_2,Y_3,\ldots,Y_n$，该样本的估计量有以下属性。

3.4.0.1 次序统计量 (Order Statistics)

次序统计量之间既不独立也不同分布。

任意两个次序统计量的联合分布不同。

对于第 k 个次序统计量 $X_{(k)}$，其密度函数 $f_k(x)$ 为:

$$f_k(x)=\frac{n!}{(k-1)!(n-k)!}[F(x)]^{k-1}[1-F(x)]^{n-k}f(x)$$

对于两个次序统计量 X_i 和 $X_j(i<j)$，它们的联合分布密度函数 $f_{ij}(x_i,x_j)$ 为:

$$f_{ij}(x_i,x_j)=\frac{n!}{(i-1)!(j-i-1)!(n-j)!}[F(x_i)]^{i-1}[F(x_j)-F(x_i)]^{j-i-1}[1-F(x_j)]^{n-j}f(x_i)f(x_j)$$

对于 n 个次序统计量的联合分布密度函数 $f(x_1,x_2,x_3,\ldots,x_n)$，有:

$$f(x_1,x_2,x_3,\ldots,x_n)=\begin{cases}n!\prod_{i=1}^{n}f(x_i) & (x_1\le x_2\le x_3\le\cdots\le x_n)\\ 0 & \text{else}\end{cases}$$

3.4.0.2 分位数 (Quantile)

Y_p 是样本的 p 分位数，则样本的 p 分位数 m_p 的渐进分布是 $N(Y_p,\dfrac{p(1-p)}{n[f(Y_p)]^2})$.

3.4.0.3 样本均值 (Mean)

均值 $\bar{Y}$ 是

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

无论总体遵循什么分布，随机样本的均值的期望都和总体的均值相同，并且随机样本的均值的方差等于总体方差的 $\frac{1}{n}$：（此处证明见5.2.17）

$$\mathrm{E}[\bar{Y}] = \mu, \mathrm{Var}(\bar{Y}) = \frac{\sigma^2}{n}$$

无论总体遵循什么分布，随机样本的渐进分布都是正态分布 $N\left(\mu, \frac{\sigma^2}{n}\right)$，记作 $\bar{x} \sim AN\left(\mu, \frac{\sigma^2}{n}\right)$，证明见5.2.18。

特别地，若总体 X 服从两点分布 $b(1,p)$，则对应的随机样本均值 $\bar{Y}$ 的渐进分布是 $N(p, \frac{p(1-p)}{n})$.

3.4.0.4 样本方差 (Variance)

方差 $\mathrm{Var}(Y)$ 是

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

无论总体遵循什么分布，随机样本的方差的期望都等于总体方差。换句话说，随机样本的方差是总体方差的一个无偏估计：（此处证明见5.2.19）

$$\mathrm{E}[S^2] = \sigma^2$$

3.4.0.5 样本标准差 (Standard Error)

定义：样本的标准差 $S(Y)$ 为 $\sqrt{\mathrm{Var}(Y)}$. 虽然 S^2 是总体方差 σ^2 的一个**无偏估计**（5.2.19），但 S 并不是 σ 的无偏估计，证明见附录5.2.20。

3.4.0.6 样本矩 (Sample Moments)

$k \in \mathbb{Z}_+$，我们称

$$A_k = \frac{1}{n}\sum_{i=1}^{n} Y_i^k$$

为样本的 k 阶原点距。称

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^k$$

为样本的 k 阶中心距。

3.4.0.7 样本偏度 (Skewness)

称

$$\hat{\beta}_s = \frac{B_3}{B_2^{\frac{3}{2}}}$$

为样本偏度。样本的偏度越大，其分布在其中心之外的概率密度就越大、越多。

3.4.0.8 样本峰度 (Kurtosis)

称

$$\hat{\beta_k} = \frac{B_4}{B_2^2} - 3$$

为样本峰度。样本的峰度越大，其概率密度图像就在其中心越尖。

3.4.0.9 线性估计量 (Linear Estimator)

定义：如果一个估计量 T 是这样得到的（其中，a_i 都是常数，Y_i 都是随机变量）

$$T = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n = \sum_{i=1}^{n} a_iY_i$$

则称 T 是一个线性估计量。

对于线性估计量而言，若给出条件 $\mathrm{E}[Y_i] = \mu_i$, $\mathrm{Var}(Y_i) = \sigma_i^2$ 且 Y_i 之间独立同分布，那么我们就有以下结论，证明请见附录5.2.11

$$\mathrm{E}[T] = \sum_{i=1}^{n} a_i\mu_i,\ \mathrm{Var}(T) = \sum_{i=1}^{n} a_i^2\sigma_i^2$$

3.4.0.10 最优线性无偏估计量 (Best Linear Unbiased Estimator)

定义：在所有对 θ 进行估计的线性无偏估计量 (Linear Unbiased Estimator) 中，方差最小的那个 T 被称为最优线性无偏估计量 (BLUE)。其满足如下几条性质：

1. T 是无偏的，即 $\mathrm{E}[T] = \theta$.
2. T 是线性的，即 $T = \sum_{i=1}^{n} a_iY_i$.
3. T 在所有线性无偏估计值中的方差最小。即对所有线性无偏估计量 T'，有 $\mathrm{Var}(T) \leq \mathrm{Var}(T')$.

如果 $\mathrm{E}[Y_i] = \mu$, $\mathrm{Var}[Y_i] = \sigma$ 且 Y_i 之间独立同分布，对一个估计 μ 的线性估计量 $T = \sum_{i=1}^{n} a_iY_i$ 而言，它是无偏的，当且仅当 $\sum_{i=1}^{n} a_i = 1$，证明请见附录5.2.12。由此引理，可以得到估计 μ 的最优线性无偏估计量的公式是：（证明请见附录5.2.13）

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

3.4.0.11 随机样本和 T 分布

该统计量满足自由度为 $n-1$ 的 T 分布（证明请见附录5.2.22），即：

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

3.5 正态随机样本估计量 Estimators of Normal Random Sample

若总体 X 符合 $N(\mu, \sigma^2)$ 的正态分布，则抽取的每个随机变量 $Y_1, Y_2, Y_3, \ldots, Y_n$ 都满足相同的 $N(\mu, \sigma^2)$ 正态分布，且相互独立。这些样本除了满足上述一般随机样本的所有性质之外，还有以下性质。

3.5.0.1 样本均值

均值 $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ 的精确分布是正态分布 $N(\mu, \frac{\sigma^2}{n})$，证明请见附录5.2.15。

根据3.4.0.10，该均值是估计参数 μ 的最佳线性无偏估计量。令 $Z = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$，则有 $Z \sim N(0,1)$，称这里的 Z 是标准化后的正态随机变量。证明请见附录5.2.15。

在未知总体方差的情况下，令 $Z = \frac{\bar{Y}-\mu}{S/\sqrt{n}}$，则有 $Z \sim T(n-1)$.

对 μ 而言，其中一个置信水平为 $100(1-\alpha)\%$ 的置信区间为:

$$\left(\bar{y} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

其中，$z_{\frac{\alpha}{2}}$ 的取值满足 $P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$，即分位数。由于正态分布是对称的，所以该区间也关于 $\bar{y}$ 对称。此结论证明请见附录5.2.16。

3.5.0.2 样本方差

一般地，我们使用 S^2，也就是**样本方差**来估计总体方差。

由样本方差构造的该估计量符合卡方分布:（此处证明见附录5.2.21）

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

有了分布，该估计量在给定置信水平下的置信区间就是可求的。

3.5.0.3 样本均值和方差的关系

样本均值 $\bar{Y}$ 和方差 S^2 是相互独立的。

3.5.0.4 两个不同正态样本的关系

对于来自两个不同正态分布 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$，容量分别是 n_1, n_2 的样本，它们的平均值分别是 $\bar{X}_1, \bar{X}_2$，方差分别是 S_1^2, S_2^2，则有

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

并且

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

特别地，当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 时，有

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

3.5.0.5 线性估计量

对线性估计量 T:

$$T = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n = \sum_{i=1}^{n} a_iY_i$$

有:（证明详见5.2.14）

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

3.5.1 大数定律 (Law of Large Numbers, LLN)

大数定律的中心思想是，样本数量越多，则其算术平均值就有越高的概率接近期望值。

3.5.1.1 弱大数定律 (Weak Law of Large Numbers, WLLN)

弱大数定律的表达是: 对一列独立同分布且期望等于 μ, 方差等于 σ^2 的样本 $X_1, X_2, X_3, \ldots, X_n$, 成立:（证明请见5.2.7）

$$\bar{X}_n \xrightarrow[n\to+\infty]{P} \mu$$

即 $\forall \varepsilon > 0$，都有:

$$\lim_{n\to+\infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0$$

3.5.1.2 强大数定律 (Strong Law of Large Numbers, SLLN)

强大数定律的表达是: 对一列独立同分布且期望等于 μ, 方差等于 σ^2 的样本 $X_1, X_2, X_3, \ldots, X_n$, 成立:（证明请见5.2.8）

$$\bar{X}_n \xrightarrow[n\to+\infty]{\text{a.s.}} \mu$$

即

$$P(\lim_{n\to+\infty} \bar{X}_n = \mu) = 1$$

3.5.2 中心极限定理 (Central Limit Theorem, CLT)

不管是从什么类型的分布中取的样本，只要样本足够大，则样本的平均值一定服从正态分布 $N(\mu, \frac{\sigma^2}{n})$，其中 μ 是总体的平均值，n 是样本大小,σ 是总体标准差（当总体标准差未知时，用样本标准差代替）。所以在许多情况下，我们对样本的平均值进行统计分析时，都直接认为其符合正态分布。

中心极限定理仅仅在样本数量很大的时候才生效。中心极限定理的证明见5.2.9

3.5.3 Glivenko-Cantelli 定理 (Glivenko-Cantelli Theorem)

该定理表明，当 n 充分大时，经验分布函数的任意一个观察值 $F_n(x)$ 几乎必然收敛到总体分布函数 $F(x)$，即（证明见5.2.23）

$$P\left(\lim_{n\to+\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0\right) = 1$$

第四章　假设检验 (Hypothesis Test)

假设检验是用来判断样本与样本、样本与总体的差异是由抽样误差引起还是由本质差别造成。数学上说，这便是检验一个已知分布的随机变量，它的某个属性（期望、方差等）落在人们所期望或不期望的区间内的大小。

假设的定义是一个统计学上的**陈述**，或者是一种**声称**，比如，声称该参数等于零，这就是一种假设。

零假设 (Zero Hypothesis) 是默认的假设，一般在一开始认为是正确的。备择假设 (Alternative Hypothesis) 一般是零假设的补集，比如若零假设是某参数等于零，则备择假设是某参数不等于零。

零假设和备择假设都是人造的，其主要目的是为了支持我们想要证明的结论。零假设是我们想要拒绝的假设，也就是和我们想要的结论所相反的假设，而备择假设则是我们真正想要的。通过一组给定的数据，我们可以得出判断，是否拒绝零假设，从而获得有意义的统计学结果。

在执行假设检验的过程中，我们一般都会构造一个**检验统计量 (Test Statistic)**，以及一个**拒绝域 (Critical Region)**，如果检验统计量落在拒绝域内，则我们拒绝原假设，反之不能拒绝原假设。

总结起来，假设检验至少涉及这几个要素：

- 零假设
- 备择假设
- 检验统计量
- 拒绝域

4.0.1　显著性水平 (Significance Level)

在检验统计量满足一个已知分布的情况下，如果这个量落在该分布的某个区间内的概率已经小于一个取定的、较小的值（一般记为 α），称作**显著性水平**，我们就认为这个量在实际生活中不可能落在这个区间内。所以说，显著性水平和选定的区间有关。显然地，当某个区间上的显著性水平越低，则该变量越不可能落在这个区间上，这个变量的值所对应的某个具体事件从统计学上来说就越不可能发生，越可以说明问题，所以这个变量叫做“显著性水平”。

4.0.2　弃真错误 (Type I Error) 与取伪错误 (Type II Error)

在执行假设检验的过程中，无论是拒绝还是不拒绝零假设，都不一定能保证这个决定绝对正确，都是有可能发生错误的。可能发生的错误有两种，分别称为弃真错误和取伪错误。这两个错误不尽相同，详见表格：

	原假设正确	原假设错误
拒绝原假设	弃真错误 (Type I Error)	Success
无法拒绝原假设	Success	取伪错误 (Type II Error)

在执行假设检验的过程中，导致弃真错误和取伪错误的概率同时存在，并且我们会希望这两个概率都尽量小，但是这两个概率是互斥的。也就是说，减小发生弃真错误概率的同时将会增大发生取伪错误的概率，反之

亦然。一般情况下来说，为了解决这个问题，我们会将发生弃真错误的概率固定在一个较小的值，这是与前文中构造零假设和备择假设呼应的。然后我们在检验过程中进行控制，使得发生取伪错误的概率尽量小。

此处给出几个概念：α, β 和 $1-\beta$。第一个就是显著性水平，它是 $P(\text{reject } H_0|H_0 \text{ is true})$，即发生弃真错误的概率。而 $\beta = P(\text{retain } H_0|H_0 \text{ is false})$，即发生取伪错误的概率。$1-\beta$ 被称为**功效 (Power)**，它是在给定显著性水平和样本大小的情况下，假设检验能够正确拒绝错误假设的概率。也就是说，Power 表示了检验能够发现实际差异并拒绝虚假假设的能力大小，这可以通过公式 $\beta = P(\text{reject } H_0|H_0 \text{ is false})$ 看出。Power 越高，说明检验的能力越强，更容易发现实际存在的差异。通常，我们希望检验的 Power 越高越好，一般认为 Power 大于 0.8 时可以接受。

4.0.3 样本大小 (Sample Size)

样本大小会显著影响 α 和 β 的取值，也就是说样本大小越大，我们就能获得越高的功效。在样本量不变的情况下，α 和 $1-\beta$ 是正相关的，但是我们希望 α 尽量小，$1-\beta$ 尽量大。为了解决这个问题，我们可以通过增加样本大小来改变二者的关联关系，尽管它们仍然是正相关的，但是可以将二者都控制在一个合理的范围。4.1 解释了前者，4.2 解释了后者。在 Figure 2. 中，下图较上图的标准误差更小，因为两个分布都更加集中。标准误差 SE= $\frac{\sigma}{\sqrt{n}}$，所以一个简单的减小标准误差的方法就是增大样本量。

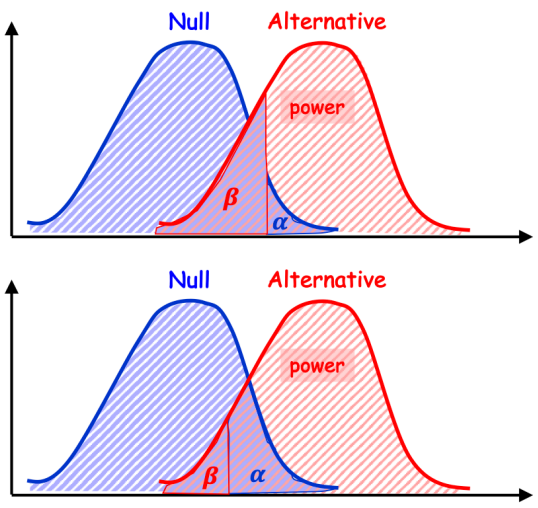

图 4.1: α，β，与 Power 的关系

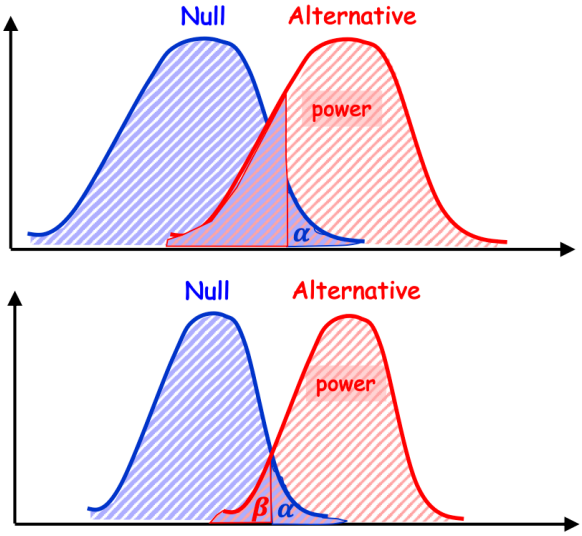

图 4.2: 标准差大小不同对它们关系产生的影响

我们通过给定需要的 α 和 β，可以计算出所需的样本大小。对于假设检验 $H_0 : \mu = \mu_0, H_a : \mu > \mu_0$ 和给定的 α 和 β，该单边检验所需的样本容量为:

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_a - \mu_0)^2}$$

而对于假设检验 $H_0 : \mu_1 = \mu_2, H_a : \mu_1 \neq \mu_2$ 而言，给定 α 和 β，以及两均值相差大小 δ，所需的样本量为

$$n = \frac{2\sigma^2 \left(z_{\alpha/2} + z_\beta\right)^2}{\delta^2}$$

对于假设检验 $H_0 : p = p_0, H_a; p \neq p_0$，给定 α 和 β，如想要至少检验出 p 与 p_0 有 δ 大小的差异，需要的样本量为:

$$n = \frac{z^2_{\alpha/2} p(1-p)}{\delta^2}$$

而对于检验两个不同比例的差异 $H_0 : p_1 = p_2, H_a : p_1 \neq p_2$，若要想证明二者之间的确有差异，需要的样本量为:

$$n = \frac{\left(z_{\alpha/2} + z_\beta\right)^2 \left[p_1\left(1-p_1\right) + p_2\left(1-p_2\right)\right]}{\left(p_1 - p_2\right)^2}$$

以上两个比较比例的假设检验中，都可以把对应的 $z_{\frac{\alpha}{2}}$ 换成单边检验的数值，然后进行单边的假设检验的样本量计算。

4.0.4 p-value

在假设检验中，p-value 指的是在原假设成立的情况下，无限次重复地从总体中抽取相同样本量大小的样本，观测到和当前样本同样或更极端的结果出现的概率。如果该值的大小是 $\frac{1}{1000000}$，则说明如果原假设为真，观察到如此极端或者更加极端的数据的概率是百万分之一，所以我们很显然会认为原假设不成立。

这种方法也可以用来进行假设检验，若 p-value 的值小于 α，则拒绝原假设。

p-value 的计算方式很简单，是 $\int_{-\infty}^{-|Z|} f(x) + \int_{|Z|}^{+\infty} f(x)$，其中 z 是检验统计量，$f(x)$ 是检验统计量符合的分布的 PDF。上式主要用于计算双边 p-value，在计算单边 p-value 时，计算方式和置信区间类似。

4.0.5 检验统计量 (Test Statistic, Z)

前文中提到，假设检验的本质是检验一个已知分布的随机变量，它的某个属性是否落在人们期望的区间内。统计学上为了方便，会希望将种类繁多、分布各异的各种随机变量，转化成符合一类人们已知分布的随机变量，这样就方便进行假设检验了。而这个随机变量被称为检验统计量，它符合一类已经被计算出的分布，例如学生 T 分布、正态分布等等。

一般来说，检验统计量的值都是这么计算的:（虽然计算方式相同，但是符合的分布却并不一定全部相同）

$$Z = \frac{\text{Estimate - Parameter}}{\text{Standard Error}}$$

4.0.6 标准误差 (Standard Error, SE)

标准误差就理解成标准化之后的标准差，这通过一张表可以计算获得（后附）。标准误差的计算较复杂，需要仔细参考表格，选择合适的模型进行计算。

4.0.7 置信水平 (Confidence Level) 和置信区间 (Confidence Interval,CI)

$(1-\alpha) \times 100\%$ 称为**置信水平**。容易理解的是，置信水平也和区间有关。置信水平越高则该变量可能落在的范围越大，越可能包含真值。而**置信区间**就是由置信水平计算出来的一个区间。

事实上，所谓的置信区间就是一个**误差范围**，可以把它视作是**以区间形式存在的随机变量对象**。因为我们执行统计活动时，总是只能对其中一个或多个样本进行调查，**不可能**获得总体的所有数据。根据样本的均值去推测真值时，不能简单认为样本均值就是真值，而是通过样本的均值和标准差，推测出真值落在哪个范围上。而取样本这个过程就是随机的，所以我们说从样本中得出的估计值当然也是随机变量，置信区间这个对象也不例外。当置信水平是 95% 时，我们说真值有 95% 的概率落在这个置信区间上（即该区间有 95% 的概率覆盖真值，即随机取样调查 100 次，得出的 100 个置信区间至少有 95 个覆盖了真值）。我们当然希望置信水平充分大，以获得对于真值比较有把握的估计，但是又不希望它过分大，因为这样会使得置信区间过大使得统计工作失去意义。比如说，置信水平为 100% 的置信区间是全体实数！但这种统计结果有什么意义呢？我们当然知道真值一定是一个实数了。

要计算置信区间的值，有以下公式:

$$\left(\text{Estimate} - z_{\frac{\alpha}{2}} \times SE, \text{Estimate} + z_{\frac{\alpha}{2}} \times SE\right)$$

其中，α 是选定的 p-value，$z_{\frac{\alpha}{2}}$ 是对应分布的分位数，也称关键值 (critical value)，它与标准误差 (SE) 的乘积称为误差限 (margin of error)，所以我们说在选定的置信水平下能够得到关键值，再通过关键值求得误差限，数据的均值加减这个误差限的区间就是置信水平为 $(1-\alpha)\times 100\%$ 的置信区间。

4.0.8 理解显著性水平和置信水平的区别

它们之间**虽然被公式连接在一起，但是在统计分析中的角色几乎完全不同**。

显著性水平主要是用来说明一件事是多么的不可能发生。比如，在对比差异的时候，我们会说它们之间没有差异的概率只有 5%（有的时候会更小，参见后面的 p-value 说明），是非常小的，所以一定是有差异的。

而置信水平主要是用来计算置信区间的，这一般被用来估计某一些数据的真值落在的范围。(真值一般认为是真正的均值)

致谢 Acknowledgement

感谢我的父母在本人学习、生活中提供的慷慨资助和情感支持，你们是我的英雄。

感谢 LaTeX 社区的所有开源开发者，是你们的工作使得本书美丽、优雅的排版称为可能。

感谢 CTeX 社区的所有开源开发者，是你们使得 LaTeX 的中文本地化得以实现。此外，本书使用了 ctexbook 模板，它是一份质量很高的模板！

本书撰写的过程中大量参考了中国海洋大学李芙蓉教授制作的幻灯片，李教授在幻灯片中的语言和措辞大大增添了本书的专业性和严肃性。感谢李教授在本书撰写过程中提供的支持。

第五章　附录 Appendix

5.1　补充定义

5.1.1　极限集

对一个无限集合列 $\{A_n\}$，定义：上限集即为**在无限个集合中出现过的元素所成集，而不考察是否在连续的无限个集合中出现**，而下限集的定义是**在第 n 个集合后每一个集合中都出现的元素所成集**。数学上，上限集的定义是 $\bigcap\limits_{N=1}^{+\infty}\bigcup\limits_{n=N}^{+\infty}A_n=\overline{\lim\limits_{n\to\infty}}\,A_n$，下限集的定义是 $\bigcup\limits_{N=1}^{+\infty}\bigcap\limits_{n=N}^{+\infty}A_n=\underline{\lim\limits_{n\to\infty}}\,A_n$。需要注意的是，定义中的连续交并符号，应当先计算后者，再计算前者。

5.1.2　泰勒公式

5.1.3　Borel-Cantelli 引理

对一个无限集合列（或看作是无限个事件）$\{A_n\}$，若 $\sum\limits_{i=1}^{+\infty}P(A_n)<+\infty$，则 $P\left(\overline{\lim\limits_{n\to\infty}}\,A_n\right)=0$. 该定理的直观解释是，对无穷个概率事件而言，若它们发生的概率之和是有限的，那么其中的无限多个事件一同发生的概率是零。它的一个常用的等价表示是

$$\sum_{i=1}^{+\infty}P(A_n)<+\infty\Longrightarrow P(A_n\quad i.o.)=0$$

其中，$i.o.$ 是 indefinitely often 的缩写，意味着该事件发生无数多次。

证明

因为 $\sum\limits_{i=1}^{+\infty}P(A_n)<+\infty$，这等价于说正项无穷级数 $(P(A_n))_{n\geq1}$ 收敛。根据无穷级数的性质，级数的余项 $\sum\limits_{n=N}^{+\infty}P(A_n)$ 的下界是零，即

$$\inf_{N\geq1}\sum_{n=N}^{+\infty}P(A_n)=0$$

所以有

$$P\left(\overline{\lim_{n\to\infty}}\,A_n\right)=P\left(\bigcap_{N=1}^{+\infty}\bigcup_{n=N}^{+\infty}A_n\right)\leq\inf_{N\geq1}P\left(\bigcup_{n=N}^{+\infty}A_n\right)\leq\inf_{N\geq1}\sum_{n=N}^{+\infty}P(A_n)=0$$

因为概率一定不小于零，所以我们证明了 $P\left(\overline{\lim\limits_{n\to\infty}}\,A_n\right)=0$，证毕。

5.2 正文证明

5.2.1 期望的性质

常数 c 可以看成是取 c 的概率为 100% 的随机变量。根据期望的定义，$\mathrm{E}[c] = c$.

对分别满足分布密度 f_X 和 f_Y 的任意随机变量 X, Y 而言，设它们的联合分布密度函数是 $f_{X,Y}$，记它们的取值分别是 x, y，令 x, y 都取全体实数（实际取不到的数的概率设为零），这样就一般化了这两个随机变量。然后有证明：

$$\begin{aligned}
\mathrm{E}[aX + bY] &= \iint_{\mathbb{R}^2} (ax + by)\mathrm{d}P(X = x, Y = y) \\
&= \iint_{\mathbb{R}^2} (ax + by) f_{X,Y}(x, y)\mathrm{d}xy \\
&= a \iint_{\mathbb{R}^2} x f_{X,Y}(x, y)\mathrm{d}xy + b \iint_{\mathbb{R}^2} y f_{X,Y}(x, y)\mathrm{d}xy \\
&= a \iint_{\mathbb{R}} x f_X(x)\mathrm{d}xy + b \iint_{\mathbb{R}} y f_Y(y)\mathrm{d}xy \\
&= a\mathrm{E}[X] + b\mathrm{E}[Y]
\end{aligned}$$

根据协方差的定义（1.5），有以下证明：

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] \\
&= \mathrm{E}[XY - X\mathrm{E}[Y] - \mathrm{E}[X]Y + \mathrm{E}[X]\mathrm{E}[Y]] \\
&= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] - \mathrm{E}[X]\mathrm{E}[Y] + \mathrm{E}[X]\mathrm{E}[Y] \\
&= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]
\end{aligned}$$

所以有 $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y] + \mathrm{Cov}(X, Y)$，所有性质证毕。

5.2.2 无意识统计学家法则

给出对Theorem 1.3.1的证明：

5.2.2.1 离散情形

X 是随机变量，可以取 n 个值，即 $x_1, x_2, x_3, \ldots, x_n,$，并且其对应的 PMF 是 f_X. 有一函数 g 将 X 映成 m 个值。记 $g(X) = Y$，则有 $y_1, y_2, y_3, \ldots, y_m$，容易知道 $m \leq n$. 令 f_Y 是 Y 的 PMF. 然后有：

$$\mathrm{E}[g(X)] = \sum_{i=1}^{n} x_i f_X(x_i)$$

$$\mathrm{E}[Y] = \sum_{i=1}^{m} y_i f_Y(y_i)$$

现在定义 $g^{-1}(Y)$，由于 g 可能不是一一映射，所以假设 y_i 对应 $G(i)$ 个 x，将 $g^{-1}(y_i)$ 看作是多值函数，那么 $f_Y(y_i) = \sum_{j=1}^{G(i)} f_X(x_j)$. 将上式继续写成：

$$\begin{aligned}
\mathrm{E}[Y] &= \sum_{i=1}^{m} y_i f_Y(y_i) \\
&= \sum_{i=1}^{m} y_i \sum_{j=1}^{G(i)} f_X(x_j)
\end{aligned}$$

又因为对 y_i 对应的 $G(i)$ 个 $x_1, x_2, x_3, \ldots, x_{G(i)}$ 而言，有 $g(x_1) = g(x_2) = g(x_3) = \cdots = g(x_{G(i)}) = y_i$，所以上式进一步变形成为：

$$\begin{aligned}
\mathrm{E}[Y] &= \sum_{i=1}^{m} y_i \sum_{j=1}^{G(i)} f_X(x_j) \\
&= y_1 \sum_{j=1}^{G(1)} f_X(x_j) + y_2 \sum_{j=1}^{G(2)} f_X(x_j) + y_3 \sum_{j=1}^{G(3)} f_X(x_j) + \cdots + y_m \sum_{j=1}^{G(m)} f_X(x_j) \\
&= \sum_{j=1}^{G(1)} g(x_j) f_X(x_j) + \sum_{j=1}^{G(2)} g(x_j) f_X(x_j) + \sum_{j=1}^{G(3)} g(x_j) f_X(x_j) + \cdots + \sum_{j=1}^{G(m)} g(x_j) f_X(x_j) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{G(i)} g(x_j) f_X(x_j)
\end{aligned}$$

注意到上述公式中 x 下标的含义。这些下标不是一列 $x_1, x_2, x_3, \ldots, x_n$ 含义下的下标，而是每一个 y_i 对应的第多少个 x 的含义。上式第三行以及之后的所有 x 的下标，都是这个含义。由于 g 对每个 x 而言都对应单个确定的 y，所以多个不同的 y 不可能对应同一个 x。即：

$$g^{-1}(y_i) \cap g^{-1}(y_j) = \varnothing \quad i \neq j$$

但每一个 x 都必定有一个或多个与之对应的 y，所以我们得到：

$$\sum_{i=1}^{m} G(i) = n$$

并且每一个可能的 x 在计算 $\sum_{i=1}^{m} \sum_{j=1}^{G(i)} g(x_j) f_X(x_j)$ 的过程中都被取且仅取一次。所以有：

$$\mathrm{E}[Y] = \sum_{i=1}^{m} \sum_{j=1}^{G(i)} g(x_j) f_X(x_j) = \sum_{i=1}^{n} g(x_i) f_X(x_i) = \mathrm{E}[g(X)]$$

于是离散情形下的定理证毕。

5.2.2.2 连续情形

对符合分布密度函数 f_X 和累积分布函数 F_X 的随机变量 X 以及一个可逆且可微的函数 g，令随机变量 $Y = g(X)$，并记它的分布密度函数为 f_Y，累积分布函数为 F_Y。扩充 X 和 Y 的取值到全体实数（原先不能取到的值概率为零），那么有：

$$\frac{\mathrm{d}(g^{-1}(y))}{\mathrm{d}y} = \frac{1}{g'(g^{-1}(y))}$$

将 $x = g^{-1}(y)$ 代入上式，得到：

$$\mathrm{d}x = \frac{1}{g'(g^{-1}(y))} \mathrm{d}y$$

接下来求 Y 的累积分布函数：

$$\begin{aligned}
F_Y(y) &= \mathrm{P}(Y \leq y) \\
&= \mathrm{P}(g(X) \leq y) \\
&= \mathrm{P}(X \leq g^{-1}(y)) \\
&= F_X(g^{-1}(y))
\end{aligned}$$

求导，得到 Y 的概率密度函数：

$$\begin{aligned} f_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) \\ &= \frac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y)) \\ &= f_X(g^{-1}(y)) \frac{\mathrm{d}}{\mathrm{d}y}(g^{-1}(y)) \\ &= f_X(g^{-1}(y)) \frac{1}{g'(g^{-1}(y))} \end{aligned}$$

所以有：

$$\begin{aligned} \int_{\mathbb{R}} g(x) f_X(x) \mathrm{d}x &= \int_{\mathbb{R}} y f_X(g^{-1}(y)) \frac{1}{g'(g^{-1}(y))} \mathrm{d}y \\ &= \int_{\mathbb{R}} y f_Y(y) \mathrm{d}y \\ &= \mathrm{E}[Y] \\ &= \mathrm{E}[g(X)] \end{aligned}$$

于是连续情形下的定理证毕。

5.2.3 方差的性质

对常数 c 而言，由方差的计算公式得：

$$\mathrm{Var}(c) = \mathrm{E}[c^2] - (\mathrm{E}[c])^2 = c^2 - c^2 = 0$$

对平移后的随机变量 $X + a$，有：

$$\begin{aligned} \mathrm{Var}(X + a) &= \mathrm{E}[(X + a)^2] - (\mathrm{E}[X + a])^2 \\ &= \mathrm{E}[X^2 + 2aX + a^2] - (\mathrm{E}[X] + a)^2 \\ &= \mathrm{E}[X^2] + 2a\mathrm{E}[X] + a^2 - ((\mathrm{E}[X])^2 + 2a\mathrm{E}[X] + a^2) \\ &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 \\ &= \mathrm{Var}(X) \end{aligned}$$

对伸缩后的随机变量 aX，有：

$$\begin{aligned} \mathrm{Var}(aX) &= \mathrm{E}[a^2X^2] - (\mathrm{E}[aX])^2 \\ &= a^2\mathrm{E}[X^2] - (a\mathrm{E}[X])^2 \\ &= a^2\mathrm{E}[X^2] - a^2 (\mathrm{E}[X])^2 \\ &= a^2(\mathrm{E}[X^2] - (\mathrm{E}[X])^2) \\ &= a^2\mathrm{Var}(X) \end{aligned}$$

多个随机变量 $X_i, i = 1, 2, 3, \ldots$ 的线性组合的方差：

$$\begin{aligned}
\operatorname{Var}\left(\sum_{i=1}^{n} a_i X_i\right) &= \mathrm{E}\left[\left(\sum_{i=1}^{n} a_i X_i\right)^2\right] - \left(\mathrm{E}\left[\sum_{i=1}^{n} a_i X_i\right]\right)^2 \\
&= \mathrm{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j X_i X_j\right] - \left(\mathrm{E}\left[\sum_{i=1}^{n} a_i X_i\right]\right)^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathrm{E}[X_i X_j] - \left(\sum_{i=1}^{n} a_i \mathrm{E}[X_i]\right)^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathrm{E}[X_i X_j] - \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{E}[X_i]\mathrm{E}[X_j] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j (\mathrm{E}[X_j X_j] - \mathrm{E}[X_i]\mathrm{E}[X_j]) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \operatorname{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{n} a_i^2 \operatorname{Var}(X_i) + 2\sum_{i=1}^{n}\sum_{j>i}^{n} a_i a_j \operatorname{Cov}(X_i, X_j)
\end{aligned}$$

所有性质证毕。

5.2.4 协方差的性质

对任意常数 c 和随机变量 X 而言，有：

$$\begin{aligned}
\operatorname{Var}(X, c) &= \mathrm{E}[Xc] - \mathrm{E}[X]\mathrm{E}[c] \\
&= c\mathrm{E}[X] - c\mathrm{E}[X] \\
&= 0
\end{aligned}$$

自身的协方差：

$$\begin{aligned}
\operatorname{Var}(X, X) &= \mathrm{E}[X \times X] - \mathrm{E}[X]\mathrm{E}[X] \\
&= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 \\
&= \operatorname{Var}(X)
\end{aligned}$$

容易证明协方差没有顺序性。接下来证明随机变量的伸缩对协方差是线性的：

$$\begin{aligned}
\operatorname{Cov}(aX, bY) &= \mathrm{E}[abXY] - \mathrm{E}[aX]\mathrm{E}[bY] \\
&= ab\mathrm{E}[XY] - ab\mathrm{E}[X]\mathrm{E}[Y] \\
&= ab\operatorname{Cov}(X, Y)
\end{aligned}$$

随机变量的平移不影响方差：

$$\begin{aligned}
\operatorname{Cov}(X + a, Y + b) &= \mathrm{E}[(X + a)(Y + b)] - \mathrm{E}[X + a]\mathrm{E}[Y + b] \\
&= \mathrm{E}[XY + aY + bX + ab] - (\mathrm{E}[X] + a)(\mathrm{E}[Y] + b) \\
&= \mathrm{E}[XY] + a\mathrm{E}[Y] + b\mathrm{E}[X] + ab - \mathrm{E}[X]\mathrm{E}[Y] - a\mathrm{E}[Y] - b\mathrm{E}[X] - ab \\
&= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] \\
&= \operatorname{Cov}(X, Y)
\end{aligned}$$

多个随机变量 X,Y,W,V 的协方差:

$$\begin{aligned}\mathrm{Cov}(aX+bY,cW+dV) &= \mathrm{E}[(aX+bY)(cW+dV)]-\mathrm{E}[aX+bY]\mathrm{E}[cW+dV]\\ &= \mathrm{E}[acXW+adXV+bcYW+bdYV]-(a\mathrm{E}[X]+b\mathrm{E}[Y])(c\mathrm{E}[W]+d\mathrm{E}[V])\\ &= ac\mathrm{E}[XW]+ad\mathrm{E}[XV]+bc\mathrm{E}[YW]+bd\mathrm{E}[YV]-\\ &ac\mathrm{E}[X]\mathrm{E}[W]-ad\mathrm{E}[X]\mathrm{E}[V]-bc\mathrm{E}[Y]\mathrm{E}[W]-bd\mathrm{E}[Y]\mathrm{E}[V]\\ &= ac\mathrm{Cov}(X,W)+ad\mathrm{Cov}(X,V)+bc\mathrm{Cov}(Y,W)+bd\mathrm{Cov}(Y,V)\end{aligned}$$

所有性质证毕。

5.2.5 Markov 不等式

根据定义，$\mathrm{E}[X]=\int_{-\infty}^{+\infty} xf(x)$，其中，$f(x)$ 是 X 的概率密度函数（1.2.1）。但因为 X 是非负的随机变量，所以有:

$$\mathrm{E}[X]=\int_{-\infty}^{+\infty} xf(x)=\int_{0}^{+\infty} xf(x)$$

于是，对于 $a\in(0,+\infty)$，有

$$\begin{aligned}\mathrm{E}[X] &= \int_0^{+\infty} xf(x)\\ &= \int_0^{a} xf(x)+\int_a^{+\infty} xf(x)\\ &\geq \int_a^{+\infty} xf(x)\\ &\geq \int_a^{+\infty} af(x)\\ &= a\int_a^{+\infty} f(x)\\ &= aP(X\geq a)\end{aligned}$$

两边同时除 a，有:

$$P(X\geq a)\leq\frac{\mathrm{E}[X]}{a}$$

对于离散型随机变量，将积分换成累加同理可证，证毕。

5.2.6 Chebyshev 不等式

$$\begin{aligned}P(|X-\mu|\geq k\sigma) &= P((X-\mu)^2\geq k^2\sigma^2)\\ &\leq \frac{\mathrm{E}[(X-\mu)^2]}{k^2\sigma^2} && \text{根据 Markov 不等式（1.6）}\\ &= \frac{\sigma^2}{k^2\sigma^2} && \text{根据方差定义}\\ &= \frac{1}{k^2}\end{aligned}$$

证毕。

5.2.7 弱大数定律

因为 $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + X_3 + \cdots + X_n)$，并且 $X_i, i = 1, 2, 3, \ldots, n$ 之间互相独立，所以有

$$\mathrm{Var}(\bar{X}_n) = \mathrm{Var}(\frac{1}{n}(X_1 + X_2 + X_3 + \cdots + X_n)) = \frac{1}{n^2}\mathrm{Var}(X_1 + X_2 + X_3 + \cdots + X_n) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

容易证明 $\mathrm{E}[\bar{X}_n] = \mu$. $\forall \varepsilon > 0$，利用 Chebyshev 不等式（1.7）得到：

$$P(|\bar{X}_n - \mu| \geq \varepsilon) = P(|\bar{X}_n - \mu| \geq \frac{\sqrt{n}\varepsilon}{\sigma}\frac{\sigma}{\sqrt{n}}) \leq \frac{\sigma^2}{n\varepsilon^2}$$

这意味着

$$\lim_{n\to+\infty} P(|\bar{X}_n - \mu| \geq \varepsilon) = \lim_{n\to+\infty} \frac{\sigma^2}{n\varepsilon^2} = 0$$

根据按概率收敛的定义（1.8），得到：

$$\bar{X}_n \xrightarrow[n\to+\infty]{P} \mu$$

证毕。

5.2.8 强大数定律

对一列独立同分布的样本 $X_1, X_2, X_3, \ldots, X_n$，其满足 $\mathrm{E}[X_i] = \mu < +\infty, \mathrm{Var}(X_i) = \sigma^2 < +\infty$. 不失一般性地，我们令 $\mathrm{E}[X_i^4] = \tau < +\infty$，并且通过样本的整体平移令 $\mu = 0$，然后给出证明。

欲证强大数定律给出的结论

$$P\left(\lim_{n\to+\infty} \bar{X}_n = 0\right) = 1$$

等价于证明对于任意一个在样本空间 Ω 内的样本 ω，有

$$P\left(\omega \in \Omega : \lim_{n\to+\infty} \frac{S_n(\omega)}{n} = 0\right) = 1$$

上述公式的直观解释是，在样本空间内任意选取样本，样本均值等于零的概率是 100%。这也等价于

$$P\left(\omega \in \Omega : \lim_{n\to+\infty} \frac{S_n(\omega)}{n} \neq 0\right) = 0$$

上述公式的直观解释是，在样本空间内任意选取样本，样本均值不等于零的概率是零。注意到

$$\omega \in \Omega : \lim_{n\to+\infty} \frac{S_n(\omega)}{n} \neq 0 \Longleftrightarrow \omega \in \Omega : \exists \varepsilon > 0, \left|\frac{S_n(\omega)}{n} > \varepsilon\right| \quad i.o.$$

即：对无数个样本空间 Ω 内的样本 ω，都存在一个 $\varepsilon > 0$，使得样本均值大于 ε. 所以，要证强大数定律给出的结论，即证：$\forall \varepsilon > 0$，都有

$$P\left(\omega \in \Omega : |S_n(\omega)| \geq n\varepsilon \quad i.o.\right) = 0$$

定义事件 $A_n = \{\omega \in \Omega : |S_n(\omega)| \geq n\varepsilon\}$，则我们希望证明的是 $P(A_n \quad i.o.) = 0$. 由 Borel-Cantelli 引理（5.1.3），若我们能够证明 $\sum_{i=1}^{+\infty} P(A_n) < +\infty$，则 $P(A_n \quad i.o.) = 0$ 是显然地。接下来证明这个命题。

由 Markov 不等式 (1.6) 得

$$P\left(|S_n| \geq n\varepsilon\right) = P\left(|S_n|^4 \geq n^4\varepsilon^4\right) \leq \frac{\mathrm{E}[|S_n|^4]}{n^4\varepsilon^4}$$

于是解 $\mathrm{E}[S_n^4]$，有

$$\begin{aligned}\mathrm{E}[S_n^4] &= \mathrm{E}\left[\left(\sum_{i=1}^{n} X_i\right)\right] \\ &= \mathrm{E}\left[\sum_{1\leq i,j,k,l\leq n} X_i X_j X_k X_l\right]\end{aligned}$$

展开上式，我们知道因为样本之间互相独立，所以除了 $\mathrm{E}[X_i^2X_j^2]=(\mathrm{E}[X_i^2])^2$ 和 $\mathrm{E}[X_i^4]$ 以外的所有项都为零（包括一次因子项的所有期望都能拆出一次因子的期望，而我们已经假定所有样本的期望都为零了）。于是我们得到 n 个 $\mathrm{E}[X_i^4]$ 项和 $3n(n-1)$ 个 $(\mathrm{E}[X_i^2])^2$ 项。所以有

$$\begin{aligned}\mathrm{E}[S_n^4]&=n\mathrm{E}[X_i^4]+3n(n-1)(\mathrm{E}[X_i^2])^2\\&=\mathrm{E}[X_i^4]+3n(n-1)\sigma^4\\&=3n^2\sigma^4+n(\mathrm{E}[Y_i^4]-3\sigma^4)\\&\le Cn^2\quad\text{对于足够大的 }n\text{ 和 }C\ge3\sigma^4+1\end{aligned}$$

所以现在有

$$P\left(|S_n|\ge n\varepsilon\right)\le\frac{\mathrm{E}[|S_n|^4]}{n^4\varepsilon^4}\le\frac{Cn^2}{n^4\varepsilon^4}=\frac{C}{n^2\varepsilon^4}$$

所以

$$\sum_{n\ge n_0}P(A_n)\le\sum_{n\ge n_0}\frac{C}{n^2\varepsilon^4}\le+\infty$$

于是根据 Borel-Cantelli 引理（5.1.3），得到 $P(A_n\quad i.o.)=0$，证毕。

5.2.9 中心极限定理

对独立同分布的一组随机变量（可以进化成样本）$X_1,X_2,X_3,\ldots,X_n$，每个随机变量的期望是 μ，方差是 σ^2. 则我们说 $X_1+X_2+X_3+\cdots+X_n$ 的期望是 $n\mu$，方差是 $n\sigma^2$. 于是我们考虑随机变量

$$Z_n=\frac{X_1+X_2+X_3+\cdots+X_n}{\sqrt{n\sigma^2}}=\sum_{i=1}^{n}\frac{X_i-\mu}{\sqrt{n\sigma^2}}=\sum_{i=1}^{n}\frac{1}{\sqrt{n}}Y_i$$

上式中我们定义 $Y_i=\dfrac{X_i-\mu}{\sigma}$. 容易看出每一个 Y_i 的期望都是 0，方差都是 1. 于是 Z_n 的特征函数（1.2.4）是

$$\varphi_{Z_n}(t)=\varphi_{\sum_{i=1}^{n}\frac{1}{\sqrt{n}}Y_i}(t)=\prod_{i=1}^{n}\varphi_{Y_i}\left(\frac{t}{\sqrt{n}}\right)=\left[\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n\quad\text{（所有的 }Y_i\text{ 独立同分布）}$$

根据泰勒公式（5.1.2）

5.2.10 均方差和方差、偏差的关系

方法一：

$$\begin{aligned}MSE_T(\theta)&=\mathrm{E}[(T-\theta)^2]\\&=\mathrm{E}[(T-\mathrm{E}[T]+\mathrm{E}[T]-\theta)^2]\\&=\mathrm{E}[(T-\mathrm{E}[T])^2]+2\mathrm{E}[(T-\mathrm{E}[T])(\mathrm{E}[T]-\theta)]+\mathrm{E}[(\mathrm{E}[T]-\theta)^2]\\&=\mathrm{Var}[T]+2(\mathrm{E}[T]-\theta)E[T-E[T]]+(\mathrm{E}[T]-\theta)^2\\&=\mathrm{Var}[T]+(b_T(\theta))^2\end{aligned}$$

方法二：由 $\mathrm{Var}(T-\theta)=\mathrm{E}[(T-\theta)^2]-[\mathrm{E}[T-\theta]]^2$，得到：

$$\begin{aligned}\mathrm{E}[(T-\theta)^2]&=\mathrm{Var}(T-\theta)+[\mathrm{E}[T-\theta]]^2\\&=\mathrm{Var}(T)+[\mathrm{E}[T]-\theta]^2\\&=\mathrm{Var}(T)+(b_T(\theta))^2\end{aligned}$$

5.2.11 线性估计量的期望与方差

$$\begin{aligned}\mathrm{E}[T] &= \mathrm{E}[\sum_{i=1}^{n} a_i Y_i] \\ &= \mathrm{E}[a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n] \\ &= a_1 \mathrm{E}[Y_1] + a_2 \mathrm{E}[Y_2] + \cdots + a_n \mathrm{E}[Y_n] \\ &= \sum_{i=1}^{n} a_i \mu_i\end{aligned}$$

$$\begin{aligned}\mathrm{Var}[T] &= \mathrm{E}[T^2] - (\mathrm{E}[T])^2 \\ &= \mathrm{E}[(a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n)^2] - (a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n)^2 \\ &= \mathrm{E}[a_1^2 Y_1^2] + \mathrm{E}[2a_1 a_2 Y_1 Y_2] + \cdots + \mathrm{E}[a_n^2 Y_n^2] - (a_1^2\mu_1^2 + 2a_1 a_2 \mu_1\mu_2 + \cdots + a_n^2\mu_n^2) \\ &= \sum_{i=1}^{n} \mathrm{E}[a_i^2 Y_i^2] - \sum_{i=1}^{n} a_i^2\mu_i^2 + \sum_{\substack{i=1 \\ j=1 \\ i\neq j}}^{n} \mathrm{E}[2a_i a_j Y_i Y_j] - \sum_{\substack{i=1 \\ j=1 \\ i\neq j}}^{n} 2a_i a_j \mu_i \mu_j\end{aligned}$$

（因为 Y_i 相互独立，所以 $\mathrm{E}[Y_i Y_j] = \mathrm{E}[Y_i]\mathrm{E}[Y_j],\ i \neq j$）

$$\begin{aligned}&= \sum_{i=1}^{n} a_i^2 (\mathrm{E}[Y_i^2] - \mu_i^2) \\ &= \sum_{i=1}^{n} a_i^2 \sigma_i^2\end{aligned}$$

5.2.12 线性无偏估计量估计均值的性质

若要使某线性估计量 $\sum_{i=1}^{n} a_i Y_i$ 无偏，则必须要使 $\mathrm{E}[\sum_{i=1}^{n} a_i Y_i] = \mu$。因为有 $\mathrm{E}[Y_i] = \mu$，所以

$$\begin{aligned}\mathrm{E}[T] &= \mathrm{E}[\sum_{i=1}^{n} a + iY_i] \\ &= \sum_{i=1}^{n} a_i \mathrm{E}[Y_i] \\ &= \mu \sum_{i=1}^{n} a_i\end{aligned}$$

综上所述，要使 T 是无偏线性估计量，就要使 $\sum_i a_i = 1$，反之仍然成立，证毕。

5.2.13 最优线性无偏估计量

首先任取一个线性无偏估计量 T，则有 $\mathrm{E}[T]=\mu$，且 $T=\sum_{i=1}^{n}a_iY_i$. 根据之前已经得出的结论，有

$$\begin{aligned}\mathrm{Var}[T]&=\sum_{i=1}^{n}a_i^2\sigma_i^2\\&=\sum_{i=1}^{n}a_i^2\sigma^2\\&=\sigma^2\sum_{i=1}^{n}a_i^2\end{aligned}$$

已知 $\sum_{i=1}^{n}a_i=1$，要令 $\sum_{i=1}^{n}a_i^2$ 取得最小值，则 a_i 必须全部相等，等于 $\frac{1}{n}$，即这个方差最小的线性无偏估计为 $\sum_{i=1}^{n}\frac{1}{n}Y_i=\frac{1}{n}\sum_{i=1}^{n}Y_i$，证毕。

5.2.14 线性估计量的正态性

首先我们知道，对于正态分布 $N(\mu,\sigma^2)$，其**特征函数**为:

$$\varphi(t)=e^{it\mu-\frac{t^2\sigma^2}{2}}$$

对于 n 个互相独立的，满足正态分布 $N(\mu_i,\sigma_i^2)$ 的随机变量 $Y_1,Y_2,\ldots,Y_n$，现在 T 是它们的线性组合 $T=a_1Y_1+a_2Y_2+\cdots+a_nY_n=\sum_{i=1}^{n}a_iY_i$，计算它的特征函数:

$$\begin{aligned}\varphi_T(t)&=\mathrm{E}[e^{it(a_1Y_1+a_2Y_2+\cdots+a_nY_n)}]\\&=\mathrm{E}[e^{it(a_1Y_1)}\times e^{it(a_2Y_2)}\times\cdots\times e^{it(a_nY_n)}]\\&=\mathrm{E}[e^{it(a_1Y_1)}]\times\mathrm{E}[e^{it(a_2Y_2)}]\times\cdots\times\mathrm{E}[e^{it(a_nY_n)}]\end{aligned}$$

而因为 Y_n 相互独立，所以上式中的乘积的期望等于期望的乘积。将计算继续进行下去，我们可以得到

$$\begin{aligned}&=\varphi_{Y_1}(a_1t)+\times\varphi_{Y_2}(a_2t)\times\cdots\times\varphi_{Y_n}(a_nt)\\&=e^{ia_1t\mu_1-\frac{a_1^2t^2\sigma_1^2}{2}}\times e^{ia_2t\mu_2-\frac{a_2^2t^2\sigma_2^2}{2}}\times\cdots\times e^{ia_nt\mu_n-\frac{a_n^2t^2\sigma_n^2}{2}}\\&=e^{it(a_1\mu_1+a_2\mu_2+\cdots+a_n\mu_n)-\frac{(a_1^2\sigma_1^2+a_2^2\sigma_2^2+\cdots+a_n^2\sigma_n^2)t^2}{2})}\\&=e^{it\sum_{i=1}^{n}a_i\mu_i-\frac{t^2\sum_{i=1}^{n}a_i^2\sigma_i^2}{2}}\end{aligned}$$

从形式来看,满足正态分布 $N(\sum_{i=1}^{n}a_i\mu_i,\sum_{i=1}^{n}a_i^2\sigma_i^2)$ 的特征函数,所以可以证明 T 符合正态分布 $N(\sum_{i=1}^{n}a_i\mu_i,\sum_{i=1}^{n}a_i^2\sigma_i^2)$，证毕。

5.2.15 标准化正态分布随机变量

由5.2.13和5.2.14可以得知, BLUE 满足的分布是 $N(\mu,\frac{\sigma^2}{n})$, 此处使用 $\bar{Y}$ 来表示 BLUE, 使用变换方法可以将 $\bar{Y}$ 转换为标准正态分布。已知 $\bar{Y}$ 的概率密度函数是 $f_{\bar{Y}}(y)=\frac{\sqrt{n}}{\sigma\sqrt{2\pi}}e^{-\frac{n}{2\sigma^2}(y-\mu)^2}$，而变换 $Z=h(\bar{Y})=\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$

是 $\bar{Y}$ 的一对一函数。其反函数是 $\bar{Y} = h^{-1}(Z) = \frac{\sigma}{\sqrt{n}}Z + \mu$，观察到 $\left|\frac{\mathrm{d}h^{-1}(Z)}{\mathrm{d}Z}\right| = \frac{\sigma}{\sqrt{n}}$，所以

$$\begin{aligned} f_Z(z) &= f_Y(h^{-1}(Z))\left|\frac{dh^{-1}}{dZ}\right| \\ &= \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} e^{-\frac{n}{2\sigma^2}(\frac{\sigma}{\sqrt{n}}Z+\mu-\mu)^2}\frac{\sigma}{\sqrt{n}} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \end{aligned}$$

符合正态分布的概率密度函数，证毕。注意到不仅仅是 BLUE，任何符合正态分布的随机变量都可以按照这种方式进行标准化。

5.2.16 置信区间证明 1

要证该区间是一个对 μ 而言置信水平为 $100(1-\alpha)\%$ 的置信区间，按照定义，就要证 $P(\bar{y} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{y} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$，接下来计算：

$$\begin{aligned} P(\bar{y} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{y} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) &= P(-z_{\frac{\alpha}{2}} < \frac{\mu-\bar{y}}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}) \\ &= P(-z_{\frac{\alpha}{2}} < -z < z_{\frac{\alpha}{2}}) \\ &= P(-z_{\frac{\alpha}{2}} < z < z_{\frac{\alpha}{2}}) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} \\ &= 1 - \alpha \end{aligned}$$

证毕。

5.2.17 样本方差和期望证明

设总体的均值、方差都已知，分别是 μ, σ^2，而随机变量 $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$，则我们知道 $\bar{Y}$ 的期望是

$$\begin{aligned} \mathrm{E}[\bar{Y}] &= \mathrm{E}[\frac{1}{n}\sum_{i=1}^{n} Y_i] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[Y_i] \\ &= \mu \end{aligned}$$

而 $\bar{Y}$ 的方差是

$$\begin{aligned} \mathrm{Var}(\bar{Y}) &= \mathrm{E}(\bar{Y}^2) - (\mathrm{E}[\bar{Y}])^2 \\ &= \mathrm{E}[(\frac{Y_1}{n} + \frac{Y_2}{n} + \cdots + \frac{Y_n}{n})^2] - \mu^2 \\ &= \mathrm{E}[\frac{Y_1^2}{n^2}] + \mathrm{E}[\frac{Y_2^2}{n^2}] + \cdots + \mathrm{E}[\frac{Y_n^2}{n^2}] + \mathrm{E}[\frac{2Y_1Y_2}{n^2} + \frac{2Y_1Y_3}{n^2} + \cdots + \frac{Y_{n-1}Y_n}{n^2}] - \mu^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}[Y_i^2] + \frac{1}{n^2}\sum_{\substack{i=1\\j=1\\i\neq j}}\mathrm{E}[Y_iY_j] - \mu^2 \end{aligned}$$

因为每个 Y_i 都是同分布的，所以每个 $\mathrm{E}[Y_i^2]$ 也都相等，等于 $\mathrm{V}ar(Y_i)+(\mathrm{E}[Y_i])^2=\sigma^2+\mu^2$，并且这里的每个 Y_i 互相都是独立同分布的，所以对于任意 $i\neq j$，都有 $\mathrm{E}[Y_iY_j]=\mathrm{E}[Y_i]\mathrm{E}[Y_j]$，然后继续计算：

$$\begin{aligned}\mathrm{Var}(\bar{Y}) &= \frac{1}{n}(\sigma^2+\mu^2)+\frac{1}{n^2}(\mu^2(n^2-n))-\mu^2\\ &= \frac{1}{n}(\sigma^2+\mu^2)+\mu^2-\frac{\mu^2}{n}-\mu^2\\ &= \frac{\sigma^2}{n}\end{aligned}$$

注意到上述证明中没有对任何随机变量的分布作出任何假设，只使用了随机变量的期望和方差，证毕。

5.2.18 任意分布中的任意随机样本的均值的渐进分布都是正态分布

5.2.19 随机样本方差推测

设总体的均值、方差分别是 μ, σ^2，而每个随机变量 Y_i 的均值和方差都和总体相同，分布也和总体相同。其中，σ^2 的值我们并不清楚，且 $\bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$。接下来我们证明 $\mathrm{E}[S^2]=\sigma^2$：

$$\begin{aligned}S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\\ \mathrm{E}[S^2] &= \mathrm{E}[\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2]\\ &= \frac{1}{n-1}\mathrm{E}\left[\sum_{i=1}^{n}(Y_i^2-2Y_i\bar{Y}+\bar{Y}^2)\right]\\ &= \frac{1}{n-1}\mathrm{E}\left[\sum_{i=1}^{n}Y_i^2-2\bar{Y}\sum_{i=1}^{n}Y_i+n\bar{Y}^2\right]\\ &= \frac{1}{n-1}\mathrm{E}\left[\sum_{i=1}^{n}Y_i^2-2n\bar{Y}^2+n\bar{Y}^2\right]\\ &= \frac{1}{n-1}\mathrm{E}\left[\sum_{i=1}^{n}Y_i^2-n\bar{Y}^2\right]\\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathrm{E}[Y_i^2]-n\mathrm{E}[\bar{Y}^2]\right)\\ \mathrm{E}[Y_i^2] &= \mathrm{Var}(Y_i)+(\mathrm{E}[Y_i])^2\\ &= \sigma^2+\mu^2\\ \mathrm{E}[\bar{Y}^2] &= \mathrm{Var}(\bar{Y})+(\mathrm{E}[\bar{Y}])^2\\ &= \frac{\sigma^2}{n}+\mu^2 \qquad \text{（此处证明可见附录5.2.17）}\\ \mathrm{E}[S^2] &= \frac{1}{n-1}\left(n(\sigma^2+\mu^2)-n(\frac{\sigma^2}{n}+\mu^2)\right)\\ &= \sigma^2\end{aligned}$$

注意到上述证明中没有对任何 Y_i 的分布作出任何假设，我们只认为它们互相独立，并已知均值和方差。证毕。

5.2.20 随机样本的标准差不是总体标准差的无偏估计

我们有:

$$(\mathrm{E}[S])^2 = \mathrm{E}[S^2] - \mathrm{Var}(S)$$

因为

$$\mathrm{Var}(S) \geq 0$$

所以

$$(\mathrm{E}[S])^2 \leq \sigma^2$$

从而

$$\mathrm{E}[S] \leq \sigma$$

5.2.21 随机样本方差推测的分布

前提条件：随机样本 Y_i 分别独立地满足相同正态分布 $N(\mu,\sigma^2)$，以 $\bar{Y}$ 指代 Y_i 的平均数。已知此处的 $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$，我们先证明 S^2 和 $\bar{Y}$ 是相互独立的：

因为 $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$，有

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2 \\ &= \frac{1}{n-1}\left[(Y_n-\bar{Y})^2 + \sum_{i=1}^{n-1}(Y_i-\bar{Y})^2\right] \end{aligned}$$

因为 $\sum_{i=1}^{n}(Y_i-\bar{Y}) = \sum_{i=1}^{n}Y_i - n\bar{Y} = 0$，而 $\sum_{i=1}^{n}(Y_i-\bar{Y}) = Y_n - \bar{Y} + \sum_{i=1}^{n-1}(Y_i-\bar{Y}) = 0$，所以我们有 $Y_n - \bar{Y} = -\sum_{i=1}^{n-1}(Y_i-\bar{Y})$，代入上式可以得到

$$S^2 = \frac{1}{n-1}\left[\left(\sum_{i=1}^{n-1}(Y_i-\bar{Y})\right)^2 + \sum_{i=1}^{n-1}(Y_i-\bar{Y})^2\right]$$

所以我们能够获知，S^2 是 $Y_i-\bar{Y}$，$i=1,2,3,\ldots,n-1$ 的函数。接下来计算 $\bar{Y}$ 和 $Y_i-\bar{Y}$ 之间的协方差，以确定其独立性。

$$\begin{aligned} \mathrm{Cov}(\bar{Y}, Y_i-\bar{Y}) &= \mathrm{Cov}(\bar{Y}, Y_i) - \mathrm{Cov}(\bar{Y},\bar{Y}) \\ &= \mathrm{Cov}\left(\frac{1}{n}\sum_{i=j}^{n}Y_j, Y_i\right) - \mathrm{Var}(\bar{Y}) \\ &= \frac{1}{n}\mathrm{Cov}(Y_j, Y_i) - \frac{\sigma^2}{n} \qquad \text{（这一步证明见附录5.2.17）} \end{aligned}$$

由 Y_i 之间的独立性我们知道，对于不相等的 i,j 而言，$\mathrm{Cov}(Y_i,Y_j)=0$，而若 i,j 相等，则有 $\mathrm{Cov}(Y_i,Y_j) = \mathrm{Cov}(Y_i,Y_i) = \mathrm{Var}(Y_i) = \sigma^2$，所以上式对任意给定的 i，都有

$$\begin{aligned} \mathrm{Cov}(\bar{Y}, Y_i-\bar{Y}) &= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} \\ &= 0 \end{aligned}$$

所以 $Y_i - \bar{Y}$ 和 Y_i 相互独立，又因为 S^2 是 $Y_i - \bar{Y}$ 的函数，所以 S^2 和 $\bar{Y}$ 相互独立。

接下来证明 $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$:

$$\begin{aligned}
S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}[Y_i - \mu - (\bar{Y} - \mu)]^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}[(Y_i - \mu)^2 - 2(\bar{Y} - \mu)(Y_i - \mu) + (\bar{Y} - \mu)^2] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}(Y_i - \mu)^2 - 2(\bar{Y} - \mu)\sum_{i=1}^{n}(Y_i - \mu) + n(\bar{Y} - \mu)^2\right]
\end{aligned}$$

而我们知道 $\sum_{i=1}^{n}(Y_i - \mu) = n\bar{Y} - n\mu = n(\bar{Y} - \mu)$，带入原方程，有

$$\begin{aligned}
S^2 &= \frac{1}{n-1}\left[\sum_{i=1}^{n}(Y_i - \mu)^2 - 2n(\bar{Y} - \mu)^2 + n(\bar{Y} - \mu)^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}(Y_i - \mu)^2 - n(\bar{Y} - \mu)^2\right]
\end{aligned}$$

然后有:

$$\begin{aligned}
\frac{(n-1)S^2}{\sigma^2} &= \frac{n-1}{\sigma^2}\frac{1}{n-1}\left[\sum_{i=1}^{n}(Y_i - \mu)^2 - n(\bar{Y} - \mu)^2\right] \\
&= \sum_{i=1}^{n}\frac{(Y_i - \mu)^2}{\sigma^2} - n\frac{(\bar{Y} - \mu)^2}{\sigma^2} \\
&= \sum_{i=1}^{n}\frac{(Y_i - \mu)^2}{\sigma^2} - \left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}\right)^2
\end{aligned}$$

根据证明的前提条件以及附录5.2.15的证明和卡方分布的定义1.11.1，我们能知道 $\frac{Y_i - \mu}{\sigma}$ 以及 $\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$ 都满足标准正态分布，并且 $\sum_{i=1}^{n}\frac{(Y_i - \mu)^2}{\sigma^2}$ 满足自由度为 n 的卡方分布 χ^2_n，$\left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}\right)^2$ 满足自由度为 1 的卡方分布 χ^2_1。将上式进行整理，得到:

$$\frac{(n-1)S^2}{\sigma^2} + \left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}\right)^2 = \sum_{i=1}^{n}\frac{(Y_i - \mu)^2}{\sigma^2}$$

因为我们已经证明了 S^2 和 $\bar{Y}$ 的独立性，由独立随机变量和的矩生成函数 (Moment Generating Function, MGF) 等于每个随机变量矩生成函数的乘积，并且我们已经知道上式左侧第二项和右侧项满足的分布以及对应的矩生成函数，令 M_1, M_2, M_3 分别代表上式从左至右三个随机变量的矩生成函数，则有

$$M_1(t)M_2(t) = M_3(t)$$

带入卡方分布的矩生成函数后，得到

$$\begin{aligned}
M_1(t) \times (1-2t)^{-\frac{1}{2}} &= (1-2t)^{-\frac{n}{2}} \\
M_1(t) &= (1-2t)^{-\frac{n}{2}+\frac{1}{2}} \\
M_1(t) &= (1-2t)^{-\frac{n-1}{2}}
\end{aligned}$$

所以 $\frac{(n-1)S^2}{\sigma^2}$ 的矩生成函数等于自由度为 $n-1$ 的卡方分布 χ^2_{n-1} 的矩生成函数，由矩生成函数的性质知 $\frac{(n-1)S^2}{\sigma^2}$ 满足自由度为 $n-1$ 的卡方分布 χ^2_{n-1}，证毕。

5.2.22 满足 T 分布证明

$$\begin{aligned}
T &= \frac{\bar{Y}-\mu}{S/\sqrt{n}} \\
&= \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \times \frac{\sigma/\sqrt{n}}{S/\sqrt{n}} \\
&= \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \div \frac{S}{\sigma} \\
&= \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \div \sqrt{\frac{S^2(n-1)}{\sigma^2(n-1)}} \\
&= \frac{\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{S^2(n-1)}{\sigma^2(n-1)}}}
\end{aligned}$$

其中，根据附录5.2.15中的证明，我们知道分式上半部分满足标准正态分布；根据附录5.2.21，我们知道分式下半部分中 $\frac{S^2(n-1)}{\sigma^2}$ 满足自由度为 $n-1$ 的卡方分布 χ^2_{n-1}，根据 T 分布的定义1.11.2，我们知道 T 满足自由度为 $n-1$ 的 T 分布，证毕。

5.2.23 Glivenko-Cantelli 定理

我们先证明以下引理：

引理 1

对任意给定的定义在 $\mathbb{R}$ 上的分布函数 F，我们说，对任意给定的 $\varepsilon > 0$，都存在一个有限的分割 $-\infty < t_0 < t_1 < t_2 < \cdots < t_k = +\infty$，使得对任意 $j \in [0, k-1] \cup \mathbb{Z}$，有

$$F(t_{j+1}^-) - F(t_j) \leq \varepsilon$$

引理 1 证明

对一个给定的 ε，令 $t_0 = -\infty$，对 $j \geq 0 \in \mathbb{Z}$，定义

$$t_{j+1} = \sup\{z : F(z) \leq F(t_j) + \varepsilon\}$$

我们希望证明结论 $F(t_{j+1}) \geq F(t_j) + \varepsilon$. 假设 $F(t_{j+1}) < F(t_j) + \varepsilon$，因为分布函数 F 的右连续性质（1.2.2），所以 $\exists \delta > 0, F(t_{j+1} + \delta) < F_j + \varepsilon$，显然这和定义冲突。

所以我们说明了在 t_j 和 t_{j+1} 之间，函数 F 的值至少跳跃 ε. 由于我们限定了一个有限的 k，所以该过程最多能够发生有限次，于是我们得到一个符合定义形式的分割。

接下来证明 $F(t_{j+1}^-) - F(t_j) \leq \varepsilon$. 注意到根据 t_{j+1} 的定义，$\forall \sigma > 0, F(t_{j+1} - \sigma) < F(t_j) + \varepsilon$，该陈述符合 $F(t_{j+1}^-)$ 的定义，证毕。

主要证明

对一列次序统计量 $Y_{(1)}, Y_{(2)}, Y_{(3)}, \ldots, Y_{(n)}$ 以及其对应的经验分布函数 $F_n(x)$ 和真实的总体分布函数 $F(x)$，不妨设

Maxwell General Learning System Academic

麦克斯韦通用高等习得系统