

Probability Theory and Mathematical Statistics

概率论与数理统计

Wentao Zhu
Queens' College, University of Cambridge

June 2025

目录

1 Probability 概率	7
1.1 Diverse notions of ‘probability’	7
1.2 Classical probability	8
1.2.1 Classical probability	8
1.2.2 Sample space and events	8
1.2.3 Combinatorial analysis: permutations and combinations	9
1.3 Mathematical probability	9
1.3.1 Measure	9
1.3.2 Probability measure	10
1.3.3 Axioms of probability	11
1.3.4 Boole’s inequality	11
1.3.5 Inclusion-exclusion formula	11
1.3.6 Bonferroni’s inequalities	12
1.3.7 Probability limit of a sequence of sets	12
1.4 Independence	12
1.4.1 Independence of two events	12
1.4.2 Independence of multiple events	13
1.5 Conditional probability	13
1.5.1 Conditional probability	13
1.5.2 Properties of conditional probability	13
1.5.3 Law of total probability	14
1.5.4 Bayes’ formula	14
1.6 Further exercises	15
1.7 Appendix: Proofs	19
2 Random variables and univariate distributions 随机变量和单变量分布	20
2.1 Random variables	20
2.1.1 Mapping outcomes to real numbers	20
2.2 Discrete random variables	20

2.2.1	Discrete random variables	20
2.2.2	Probability mass function and cumulative distribution function	20
2.3	Continuous random variables	21
2.3.1	Continuous random variables	21
2.3.2	Probability density function and cumulative distribution function	21
2.4	Mixed random variables	22
2.4.1	Mixed random variables	22
2.4.2	Functions that uniquely define a probability distribution	23
2.5	Expectation, variance, and higher moments	24
2.5.1	Expectation	24
2.5.2	Variance	24
2.5.3	Inequalities involving expectation	24
2.5.4	Moments	25
2.6	Common distributions of random variables	26
2.6.1	Common discrete distributions	26
2.6.2	Common continuous distributions	28
2.6.3	Parameters and families of distributions	29
2.6.4	Univariate distribution relationships	29
2.7	Generating functions	29
2.7.1	Probability generating function	29
2.7.2	Moment generating function	30
2.7.3	Cumulant generating functions and cumulants	31
2.8	Functions of random variables	32
2.8.1	Distribution and mass/density for $g(X)$	32
2.8.2	Monotone functions of random variables	32
2.9	Sequences of random variables and convergence	33
2.10	Further exercises	36
2.11	Appendix: Proofs	43
3	Multivariate distributions 多变量分布	44
3.1	Joint and marginal distributions	44
3.2	Joint mass and joint density	45
3.2.1	Mass for discrete distributions	45
3.2.2	Density for continuous distributions	46
3.3	Expectation and joint moments	48
3.3.1	Expectation of a function of several variables	48
3.3.2	Covariance and correlation	48
3.3.3	Joint moments	49
3.3.4	Joint moment generating function	49
3.4	Independent random variables	50
3.4.1	Independent for pairs of random variables	50
3.4.2	Mutual independence	51
3.4.3	Identical distributions	51

3.5	Random vectors and random matrices	52
3.6	Transformations of continuous random variables	52
3.6.1	Bivariate transformations	52
3.6.2	Multivariate transformations	54
3.7	Sums of random variables	55
3.7.1	Sum of two random variables	55
3.7.2	Sum of n independent random variables	56
3.8	Multivariate normal distribution	57
3.8.1	Bivariate case	57
3.8.2	n -dimensional multivariate case	58
3.9	Further exercises	59
3.10	Appendix: Proofs	60
4	Conditional distributions 条件分布	61
4.1	Discrete conditional distributions	61
4.2	Continuous conditional distributions	61
4.3	Relationship between joint, marginal, and conditional	61
4.4	Conditional expectation and conditional moments	62
4.4.1	Conditional expectation	62
4.4.2	Conditional moments	62
4.4.3	Conditional moment generating functions	63
4.5	Hierarchies and mixtures	63
4.6	Random sums	63
4.7	Conditioning for random vectors	63
4.8	Further exercises	65
4.9	Appendix: Proofs	68
Appendix A:	Cheatsheet for Expectation, Variance, and Covariance	69
Appendix B:	Cheatsheet for Common Univariate Distributions	70
5	Sample moments and quantiles 样本矩和分位数	71
5.1	Core Mathematical Models Revisit: Random Variables and Distributions	71
5.2	Population, sampling, sample, and observed sample	72
5.2.1	What is ‘Population’: Population and Data Generating Process (DGP)	72
5.2.2	Sampling, sample, and observed sample	74
5.3	Independent and identically distributed (IID) sequences	74
5.3.1	Structural and distributional assumptions	74
5.3.2	Random sample	74
5.4	Functions of a sample	75
5.4.1	Statistics	75
5.4.2	Sampling distribution	75
5.4.3	Pivotal functions	75
5.5	Common sampling distributions	76

5.5.1	Chi-squared distribution	76
5.5.2	<i>t</i> -distribution	76
5.5.3	<i>F</i> distribution	76
5.6	Sample mean	76
5.6.1	Mean and variance of the sample mean	77
5.6.2	Central limit theorem (CLT)	77
5.6.3	A Complete Proof of CLT	78
5.7	Higher-order sample moments	79
5.7.1	Sample variance	79
5.7.2	Joint sample moments	80
5.8	Sample mean and variance for a normal population	81
5.9	Sample quantiles and order statistics	82
5.9.1	Sample minimum and sample maximum	82
5.9.2	Distribution of i^{th} order statistic	83
5.10	Further exercises	84
5.11	Appendix: Proofs	85
6	Estimation, testing, and prediction 估计, 检验和预测	86
6.1	Point estimation	86
6.1.1	Bias, variance, and mean squared error	86
6.1.2	Consistency	87
6.1.3	The method of moments (MM)	88
6.1.4	Ordinary least squares (OLS)	89
6.2	Interval estimation	90
6.2.1	Coverage probability and length	90
6.2.2	Constructing interval estimators using pivotal functions	92
6.2.3	Constructing interval estimators using order statistics	92
6.2.4	Confidence sets	93
6.3	Hypothesis testing	94
6.3.1	Statistical hypotheses	94
6.3.2	Decision rules	95
6.3.3	Types of error and the power function	96
6.3.4	Basic ideas in constructing tests	97
6.3.5	Conclusions and <i>p</i> -values from tests	97
6.3.6	Hypothesis test for μ and σ^2	97
6.4	Prediction	98
6.5	Further exercises	101
6.6	Appendix: Proofs	102
7	Statistical models 统计模型	103
7.1	Linear regression	103
7.1.1	Simple linear regression	103
7.1.2	Multiple linear regression	103

7.2	Time series models	105
7.2.1	Autoregressive models	105
7.2.2	Moving-average models	105
7.2.3	Autocovariance, autocorrelation, and stationarity	105
7.3	Poisson processes	105
7.3.1	Stochastic processes and counting processes	105
7.3.2	Definitions of the Poisson process	105
7.3.3	Thinning and superposition	105
7.3.4	Arrival and interarrival times	105
7.3.5	Compound Poisson process	105
7.3.6	Non-homogeneous Poisson process	105
7.4	Markov chains	105
7.4.1	Classification of states and chains	105
7.4.2	Absorption	105
7.4.3	Periodicity	105
7.4.4	Limiting distribution	105
7.4.5	Recurrence and transience	105
7.4.6	Continuous-time Markov chains	105
7.5	Further exercises	106
7.6	Appendix: Proofs	106
8	Likelihood-based inference 基于似然的估计	107
8.1	Likelihood function and log-likelihood function	107
8.2	Score and information	108
8.3	Maximum-likelihood estimation	111
8.3.1	Properties of maximum-likelihood estimates	114
8.3.2	Numerical maximization of likelihood	115
8.3.3	EM algorithm	116
8.4	Likelihood-ratio test (LRT)	117
8.4.1	Testing in the presence of nuisance parameters	118
8.4.2	Properties of the likelihood ratio	119
8.4.3	Approximate tests	119
8.5	Further exercises	121
8.6	Appendix: Proofs	127
9	Inferential theory 推断理论	128
9.1	Sufficiency	128
9.1.1	Sufficient statistics and the sufficiency principle	128
9.1.2	Factorisation theorem	131
9.1.3	Minimal sufficiency	134
9.1.4	Application of sufficiency in point estimation	137
9.2	Variance of unbiased estimators	139
9.3	Most powerful tests	143

9.4	Further exercises	144
9.5	Appendix: Proofs	145
10	Bayesian inference 贝叶斯推断	146
10.1	Prior and posterior distributions	146
10.2	Choosing a prior	147
10.2.1	Constructing reference priors	148
10.2.2	Conjugate priors	149
10.3	Bayesian estimation	149
10.3.1	Point estimators	149
10.3.2	Quadratic loss	150
10.3.3	Absolute loss	150
10.3.4	0-1 loss	150
10.3.5	Interval estimates	151
10.4	Hierarchical models and empirical Bayes	151
10.4.1	Hierarchical models	152
10.4.2	Empirical Bayes (EB)	152
10.4.3	Predictive inference	152
10.5	Further exercises	153
10.6	Appendix: Proofs	154
11	Simulation methods 模拟方法	155
11.1	Simulating independent values from a distribution	155
11.1.1	Table lookup	155
11.1.2	Probability integral	156
11.1.3	Box-Muller method	156
11.1.4	Accept/reject method	156
11.1.5	Composition	156
11.1.6	Simulating model structure and the bootstrap	156
11.2	Monte Carlo integration	156
11.2.1	Averaging over simulated instances	156
11.2.2	Univariate vs multivariate integrals	156
11.2.3	Importance sampling	156
11.2.4	Antithetic variates	156
11.3	Markov chain Monte Carlo (MCMC)	156
11.3.1	Discrete Metropolis	156
11.3.2	Continuous Metropolis	156
11.3.3	Metropolis-Hastings algorithm	156
11.3.4	Gibbs sampler	156
11.4	Further exercises	157
11.5	Appendix: Proofs	158

1 Probability 概率

1.1 Diverse notions of ‘probability’

考虑“概率”一词的一些用法。

1. 一枚均匀硬币正面朝上的概率是 $1/2$ 。
2. 从 49 个数字中选 6 个，赢得国家彩票头奖的概率是 $1/\binom{49}{6}$ 。
3. 一枚图钉落地时“针尖朝上”的概率是 0.62。
4. 未来 30 年内圣安德烈亚斯断层发生大地震的概率约为 21%。
5. 到 2100 年人类灭绝的概率约为 50%。

“某事件的概率”这一概念，如同“点”和“时间”一样，是我们无法精确定义但却依然有用的观点。关于概率通常有三种视角：上述 1, 2 对应**古典 (classical) 概率**或者叫**理论 (theoretical) 概率**；上述 3 对应**频率主义 (frequentist) 概率**或者叫**经验 (empirical) 概率**；上述 4, 5 对应**贝叶斯 (Bayesian) 概率**或者叫**主观 (subjective) 概率**。

表 1: 频率学派与贝叶斯学派对比

频率学派 (上帝视角)	贝叶斯学派 (观察者视角)
1. “概率”的解读	
概率定义基于上帝视角。如果宇宙存在随机性，即“上帝掷骰子”，那么事件发生存在在一个只有上帝知道的“真理”概率。定义为不同平行宇宙中事件发生次数比例的极限值： $\lim_{n \rightarrow \infty} \frac{\text{事件发生次数}}{n}$ 概率是客观的、外在的。	概率定义基于观察者视角。贝叶斯概率是：个体基于所有信息（知识、经验等）对命题为真的可信程度的度量。不依赖于“平行宇宙”频率，可应用于生物、心理等领域。概率不再是寻求客观真理，而是反映认知的动态主观状态。贴近人类的有限理性——信息永不完美，理解只是近似。概率不是世界“铁律”，而是脑中不断修正的世界地图。
2. “概率”解释力的边界	
概率解读依赖于物理真理。如果真理性不存在，频率学派将毫无意义。	概率解读仅依赖于人，只要人存在，概率就有意义。
3. “概率”的获得	
由于平行宇宙不可观测，对于可重复实验事件，只能通过多次重复实验构造近似“平行宇宙”。对于一次性事件，重复实验的方法失效，概率只停留在思想层面。事实上，如果不依赖平行宇宙概念，频率学派对于概率的解读是严格局限于可重复实验事件的。在平行宇宙框架下，对一次性事件的估计可解读为“在所有平行宇宙中事件发生的比例”。	贝叶斯概率仅基于人的主观想法，因此自然对使用场景无任何要求。无论是否一次性事件，对概率获取都无影响。定义是连续的、常规的。不同于频率学派需要重复实验才能得到有限估计，贝叶斯学派在任一时刻都能给出概率。概率意义由观察者自身赋予，而非上帝赋予。意义大小取决于估测与“真理性”的距离。

你会发现，贝叶斯对于概率的定义是更加宽松的，需要的假设也是更加少的，并且贝叶斯对于概率的理解是完全包括频率学派的。贝叶斯概率是你根据你目前所有的信息（包括知识和经验）得出的关于概率的主观信念。对于频率学派而言，他们的那套理论本身就属于他们获取的知识，是他

们信息集的一部分，因此他们得出的概率自然不会与贝叶斯的范式有任何冲突。可以理解为，对于频率学派，他们主观上相信存在客观概率真理值。但除此之外，他们还额外纠结于这个上帝才知道的真理值是否存在，以及如何通过重复实验的这个频率思想从上帝视角来解释概率究竟是什么这些问题。然而概率本身就是人们对于不确定性的一种含混的主观感受，如果非要用过于强的假设把它拉离这个感受的层面可能除了丰富自己的主观信念也不会产生额外客观的现实洞见。人们已经用比较弱的公理规范了概率的数学定义，在这个范式下，纠结 P 的内涵不会对数学世界中的研究有任何贡献，相反，同样有较弱假设的贝叶斯概率才是与公理化概率的数学定义更契合的那个内涵。公理化定义只要求概率满足最基本的一些条件，比如在 0-1 之间，关于它的内涵留出了足够开放的空间，每个人可以有自己的理解，但不管怎么样，概率都是来自于人的一种主观信仰。这样的缝隙反而大大拓展了概率的适用场域。这种从上帝和真理到观察者的视角转换，帮助概率挣脱了“真与假”的形而上学束缚，赋予了概率作为人的“理性思考工具”的自由生命。正映照了那句哲学名言“人是万物的尺度”，这是我们卑微的体现，但意识到要“从神回到人”同样也是人类想要发展进步必然的思想跃迁——不是通过僭越为神，而是通过诚实面对自身认知的局限性，并在此基础上构建我们的理性大厦。

表 2: 不同宇宙假设下的概率学派表现

	假设 1: 宇宙存在随机的部分 (存在平行宇宙)	假设 2: 宿命论 (不存在平行宇宙)
频率学派	逻辑上可以观测平行宇宙来估测概率，但现实中看不到平行宇宙。对于可重复实验的事件，我们可以通过重复实验人造平行宇宙进行估计。对于一次性事件，频率方法失效，概率的获取只能依靠主观猜测。	不存在概率的真理性，频率学派崩塌，扔硬币某一面朝上的概率要么是 0 要么是 1，通过重复实验人造平行宇宙的方法失效，得到的比值是无意义的。
贝叶斯学派	永远可以得出概率。对于频率方法可行的可重复实验事件，贝叶斯学派完全包含频率视角的解读。当一个贝叶斯主义者获得的数据趋于无限时，他的主观信念会收敛到频率学派的客观概率。贝叶斯框架是频率框架的超集。对于频率方法失效的一次性事件，贝叶斯学派依然可以得出概率，它将内在随机性和认知不确定性一同纳入“信念”的框架中进行处理。	优势尽显。它坦率地承认“终极答案”已定但我们不知道，从而将概率明确定义为我们对于这个确定答案的无知程度。这完美地解决了在确定性宇宙中如何理性决策的问题。

1.2 Classical probability

1.2.1 Classical probability

古典概率适用于只有有限个等可能结果的情形。

1.2.2 Sample space and events

考虑一个具有随机结果的试验 (experiment)。所有可能结果的集合称为样本空间 (sample space)。如果可能结果的数量是可数的，我们可以将它们列出，如 $\omega_1, \omega_2, \dots$ ，那么样本空间就是 $\Omega = \{\omega_1, \omega_2, \dots\}$ 。选择特定的点 $\omega \in \Omega$ 就得到一次观测 (observation)。

集合论基础知识回顾：

交换律： $A \cap B = B \cap A$ 和 $A \cup B = B \cup A$

结合律： $A \cap (B \cap C) = (A \cap B) \cap C$ 和 $A \cup (B \cup C) = (A \cup B) \cup C$

分配律: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ 和 ...

德摩根反演律: $(A \cap B)^c = A^c \cup B^c$ 和 $(A \cup B)^c = A^c \cap B^c$

集合的差运算: $A - B = A \setminus B = A \cap B^c$

某些集合论概念在概率的语境下使用时具有特殊的含义和术语。

1. Ω 的一个子集 A 称为一个事件 (event), 它是我们感兴趣的一组结果。

2. 对于任意事件 $A, B \in \Omega$,

- 补事件 (complement event) $A^c = \Omega \setminus A$ 是事件 A 不发生的事件, 即 “非 A ”。
- $A \cup B$ 是 “ A 或 B ”。
- $A \cap B$ 是 “ A 且 B ”。
- $A \subseteq B$: 事件 A 的发生意味着事件 B 的发生。
- $A \cap B = \emptyset$ 表示 “ A 和 B 是互斥 (mutually exclusive) 或不相交 (disjoint) 事件”。

如前所述, 在古典概率中, 样本空间由有限个等可能结果组成, $\Omega = \{\omega_1, \dots, \omega_N\}$ 。对于 $A \subseteq \Omega$,

$$P(A) = \frac{\text{事件 } A \text{ 中包含的结果数}}{\text{样本空间 } \Omega \text{ 中的结果总数}} = \frac{|A|}{N}$$

1.2.3 Combinatorial analysis: permutations and combinations

如前所示, 由于古典概率中的样本空间仅包含有限个等可能结果, 我们可以将其视为一个简单的特例, 其中概率值可以通过直接计数结果得出。由于古典概率涉及计数, 通常我们需要依赖一些组合计数方法 (例如排列和组合) 来计算我们想要的结果数。

无放回抽样且有顺序

$$\begin{aligned} P_n^m &= n \times (n-1) \times (n-2) \times \cdots \quad (\text{一共 } m \text{ 个因子}) \\ &= n(n-1)(n-2) \cdots (n-m+1) \\ &= \frac{n(n-1)(n-2) \cdots (n-m+1) \cdots 1}{(n-m) \cdots 1} = \frac{n!}{(n-m)!} \end{aligned}$$

无放回抽样且无顺序

$$C_n^m = \frac{P_n^m}{P_m^m} = \frac{n!}{(n-m)!m!}$$

1.3 Mathematical probability

1.3.1 Measure

为了建立更严格概率框架, 我们简要介绍一个称为测度论 (measure theory) 的数学领域。

考虑一个集合 Ψ 及其子集 $A \subseteq \Psi$ 。我们想要了解 A 的 “大小”。如果 A 是有限集, 一种明显的方法是计算 A 中元素的个数。测度 (measure) 是作用于子集的函数, 用于衡量其大小, 推广了计数元素的概念。由于测度作用于样本空间的子集, 测度的定义域将是一个子集构成的集合。为了确保测度能被合理地定义, 我们需要这个集合具有某些性质。

设 Ψ 是一个集合, \mathcal{G} 是 Ψ 的一些子集构成的集合。当以下条件满足时, 我们称 \mathcal{G} 是定义在 Ψ 上的一个 σ -代数 (σ -algebra):

- i. $\emptyset \in \mathcal{G}$,

- ii. 如果 $A \in \mathcal{G}$, 则 $A^c \in \mathcal{G}$ (\mathcal{G} 对取补运算封闭),
- iii. 如果 $A_1, A_2, \dots \in \mathcal{G}$, 则 $\bigcup_{i=1}^{\infty} A_i \in \mathcal{G}$ (\mathcal{G} 对可数并运算封闭)。

由这些性质可知, σ -代数也对可数交运算封闭 (通过应用德摩根律)。

以下例子展示了如何为任何包含非平凡子集的集合构造两个 σ -代数。

考虑一个集合 Ψ 及其一个非平凡子集 $A \subset \Psi$ 。下面给出了定义在 Ψ 上的两个 σ -代数的例子。

1. 最小的非平凡 σ -代数包含 4 个元素: $\mathcal{G} = \{\emptyset, A, A^c, \Psi\}$, 其中 $A \subset \Psi$ 。

2. 包含成员最多的 σ -代数是通过包含 Ψ 的所有子集得到的, 即 $\mathcal{G} = \{A : A \subseteq \Psi\}$ 。这被称为 Ψ 的幂集 (power set), 有时记作 $\mathcal{P}(\Psi)$ 或 $\{0, 1\}^{\Psi}$ 。

由一个集合及其上定义的一个 σ -代数所构成的对 (Ψ, \mathcal{G}) 称为一个可测空间 (measurable space)。顾名思义, 我们在 (Ψ, \mathcal{G}) 上定义测度。

给定一个可测空间 (Ψ, \mathcal{G}) , 其上的一个测度 (measure) 是一个函数 $m : \mathcal{G} \rightarrow \mathbb{R}^+$, 满足:

- i. 对所有 $A \in \mathcal{G}$, 有 $m(A) \geq 0$,
- ii. $m(\emptyset) = 0$,
- iii. 如果 $A_1, A_2, \dots \in \mathcal{G}$ 是互不相交的, 则 $m(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} m(A_i)$ 。

三元组 (Ψ, \mathcal{G}, m) 称为一个测度空间 (measure space)。定义中的三条性质与我们对“大小”行为的直觉相符。前两条相当平凡: i. 说明“大小是非负的量”, ii. 说明“如果集合中没有任何元素, 则其大小为零”。第三条需要稍加思考, 但它本质上是以下思想的延伸: 如果两个集合没有公共元素, 那么将它们合并后, 新集合的大小是原来两个集合大小之和。如开头所述, 测度推广了计算集合元素个数的概念。

集合 $S \subseteq \mathbb{R}$ 的指示函数 (indicator function) 是 $1_S(x)$, 其中

$$1_S(x) = \begin{cases} 1 & \text{当 } x \in S \\ 0 & \text{否则} \end{cases}$$

集合 S 通常取区间的形式, 例如 $[0, 1]$ 或 $[0, \infty)$ 。

假设 (Ψ, \mathcal{G}) 是一个可测空间。

- 1. 如果 Ψ 是有限的, 并且我们通过 $m(A) = |A|$ (对所有 $A \in \mathcal{G}$) 来定义 m , 那么 m 是一个测度。
- 2. 如果 1_A 是集合 A 的指示函数, 那么由 $m(A) = 1_A(x)$ (对于 $x \in \Psi$) 定义的函数是一个测度。

1.3.2 Probability measure

测度给出了集合大小的概念。概率则告诉我们一个事件发生的可能性。我们将把这两个概念结合起来, 将概率定义为一个测度。

为了定义一个测度, 我们需要一个可测空间, 即一个集合及其上定义的一个 σ -代数。我们之前对概率的直观描述引入了样本空间 Ω 的概念, 即我们试验的所有可能结果的集合。我们还将事件定义为 Ω 的子集, 其中包含我们感兴趣的结果。通过这个设置, 我们可以生成一个可测空间 (Ω, \mathcal{F}) , 其中 \mathcal{F} 是定义在 Ω 上的一个 σ -代数。这里 \mathcal{F} 是 Ω 的子集的一个集合, 我们将 \mathcal{F} 中的元素解释为事件。

给定一个可测空间 (Ω, \mathcal{F}) , 其上的一个概率测度 (probability measure) 是一个测度 $P : \mathcal{F} \rightarrow [0, 1]$, 且满足性质 $P(\Omega) = 1$ 。

1.3.3 Axioms of probability

一个概率空间 (probability space) 是一个三元组 (Ω, \mathcal{F}, P) , 其中 Ω 是样本空间, \mathcal{F} 是 Ω 的子集的一个集合 (事件域), 而 P 是一个概率测度 $P : \mathcal{F} \rightarrow [0, 1]$ 。

为了获得一致的理论, 我们必须对 \mathcal{F} 提出要求 (要求它是定义在 Ω 上的一个 σ -代数):

$$F_1 : \emptyset \in \mathcal{F} \quad (\text{且 } \Omega \in \mathcal{F}).$$

$$F_2 : A \in \mathcal{F} \implies A^c \in \mathcal{F}.$$

$$F_3 : A_1, A_2, \dots \in \mathcal{F} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$$

习题 1: 设 \mathcal{F} 为一事件域, 若 $A_n \in \mathcal{F}, n = 1, 2, \dots$, 试证明: (1) 有限并 $\bigcup_{i=1}^n A_i \in \mathcal{F}$ (2) 有限交 $\bigcap_{i=1}^n A_i \in \mathcal{F}$ (3) 可列交 $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$ (4) 差运算 $A_1 - A_2 \in \mathcal{F}$

我们也对 P 提出要求: 它是一个定义在 \mathcal{F} 上的实值函数, 满足三条公理 (称为科尔莫戈罗夫公理):

I. 对所有 $A \in \mathcal{F}$, 有 $0 \leq P(A) \leq 1$ 。

II. $P(\Omega) = 1$ 。

III. 对于任意可数的互不相容事件序列 A_1, A_2, \dots (即 $A_i \cap A_j = \emptyset, i \neq j$), 有 $P(\bigcup_i A_i) = \sum_i P(A_i)$ 。

$P(A)$ 称为事件 A 的概率。

定理 1.1 (P 的性质) 公理 I-III 蕴含以下进一步的性质:

(i) $P(\emptyset) = 0$ 。 (空集的概率)

* 注意 $A = \emptyset \implies P(A) = 0$ 但是已知 $P(A) = 0$, A 未必是 \emptyset 。

(ii) $P(A^c) = 1 - P(A)$ 。

(iii) 如果 $A \subseteq B$, 则 $P(A) \leq P(B)$ 。 (单调性)

(iv) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ 。

(v) $P(A \cap B^c) = P(A) - P(A \cap B)$ 。

(vi) 如果 $A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots$, 则 $P(\bigcup_{i=1}^{\infty} A_i) = \lim_{n \rightarrow \infty} P(A_n)$ 。 ($P(\cdot)$ 是一个连续函数。)

1.3.4 Boole's inequality

定理 1.2 (布尔不等式) 如果 (Ω, \mathcal{F}, P) 是一个概率空间且 $A_1, A_2, \dots \in \mathcal{F}$, 那么

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i).$$

特例是:

$$P\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n P(A_i).$$

1.3.5 Inclusion-exclusion formula

定理 1.3 (容斥原理) 考虑一个概率空间 (Ω, \mathcal{F}, P) 。如果 $A_1, \dots, A_n \in \mathcal{F}$, 那么

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i) - \sum_{i_1 < i_2} P(A_{i_1} \cap A_{i_2}) + \sum_{i_1 < i_2 < i_3} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \dots + (-1)^{n-1} P(A_1 \cap \dots \cap A_n).$$

我们有时记 $P(ABC)$ 表示与 $P(A \cap B \cap C)$ 相同的意思。

1.3.6 Bonferroni's inequalities

定理 1.4 (邦费罗尼不等式) 对于任意事件 A_1, \dots, A_n 和任意 r ($1 \leq r \leq n$)，

$$P\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{t=1}^r (-1)^{t-1} \sum_{1 \leq i_1 < \dots < i_t \leq n} P(A_{i_1} \cap \dots \cap A_{i_t}),$$

当 r 为奇数时取 \leq ，当 r 为偶数时取 \geq 。

1.3.7 Probability limit of a sequence of sets

定理 1.5 (概率的连续性) *i.* 如果 $\{A_i : i = 1, 2, \dots\}$ 是一个递增的集合序列 (即 $A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots$)，那么

$$\lim_{n \rightarrow \infty} P(A_n) = P\left(\bigcup_{i=1}^{\infty} A_i\right).$$

ii. 如果 $\{A_i : i = 1, 2, \dots\}$ 是一个递减的集合序列 (即 $A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots$)，那么

$$\lim_{n \rightarrow \infty} P(A_n) = P\left(\bigcap_{i=1}^{\infty} A_i\right).$$

1.4 Independence

1.4.1 Independence of two events

两个事件 A 和 B 被称为**独立 (independent)** 的，如果

$$P(A \cap B) = P(A)P(B)$$

否则它们被称为**相关 (dependent)** 的。

注意，如果 A 和 B 独立，那么 A 和 B^c 独立， A^c 和 B 独立， A^c 和 B^c 也独立。

定理 1.6 如果 $P(A) = 0$ 或 $P(A) = 1$ ，那么 A 和 B 独立。

定理 1.7 如果 $P(A) > 0$ 且 $P(B) > 0$ ，那么：若 A 和 B 独立，则它们不互斥；若 A 和 B 互斥，则它们不独立。

注意，这里一个很容易产生的误解是“相互独立就是互不影响”。实际上事件 A 与事件 B 相互独立未必代表事件 A 与事件 B 在各个层面都互不影响，互不影响的只是彼此发生的概率，但这个定义并不涉及事件本身的内容如何变化。从维恩图上理解，只需要满足两个事件的面积之积等于它们交集的面积即为独立事件，但两个面积之积是多少在图像上似乎难以直观看出，我们不妨换一个视角，把独立事件的定义写作：

$$\frac{P(A \cap B)}{P(B)} = P(A)$$

这在维恩图中表示当 A 在全集中的面积占比等于 A 在 B 中的占比时，两事件相互独立。我们通过两个例子来说明仅仅靠直觉来判断两个事件是否独立的风险。首先思考以下两个事件是否独立：事件 A 为我早饭吃鸡蛋，事件 B 为美国总统晚饭吃汉堡，显然认为这两个事件是独立的是比较合理的，我早饭是否吃鸡蛋不足以影响美国总统晚饭是否吃汉堡的概率。但现在思考下面这两个事件：扔一枚均匀的骰子，事件 A 为骰子向上点数小于 5，事件 B 为骰子向上点数是偶数。你第一直觉可能会觉得它们不是独立的，因为原本事件 A 的点数可以是 1, 2, 3, 4，但如果事件 B 发生，事件 A 可以取的点数就会发生改变，1 和 3 显然就不能取了，同理已知事件 A 发生也会影响事件

B 的可能结果。但我们要注意，“独立”是指发生的概率不受影响，当我们按照独立的定义计算概率时，我们发现，已知事件 B 发生事件 A 发生的概率是 $2/3$ ，事件 A 发生的概率是 $4/6 = 2/3$ ，因此两个事件应是相互独立的。如果概率没有被影响，那是什么被影响了？被影响的是可能结果的分布：虽然概率没变，但在给定 B 发生的情况下，原本的样本空间可能被限制了，从而使得可能出现的取值发生了变化。

总结而言，独立性是指概率的稳定性，而不是具体可能取值的不变性。在事件独立的情况下，已知一个事件发生，另一个事件的概率不变，但可能的样本空间会缩小或改变，导致某些具体取值不能再出现，造成“看起来”受影响的错觉。

1.4.2 Independence of multiple events

事件 A_1, A_2, \dots 被称为**独立的**（或者为了强调是“相互独立”的），如果对所有 $i_1 < \dots < i_r$ ，都有

$$P(A_{i_1} \cap \dots \cap A_{i_r}) = P(A_{i_1}) \dots P(A_{i_r})$$

事件可以**两两独立**（pairwise independent）而不（相互）独立。

* 多事件互相独立是比两两独立更强的概念， A, B, C 两两独立只可以得到

$$P(A \cap B) = P(A)P(B), \quad P(A \cap C) = P(A)P(C), \quad P(B \cap C) = P(B)P(C)$$

但未必得到 $P(A \cap B \cap C) = P(A)P(B)P(C)$ 。

1.5 Conditional probability

1.5.1 Conditional probability

假设 B 是一个满足 $P(B) > 0$ 的事件。对于任意事件 $A \subseteq \Omega$ ，给定 B 发生的条件下 A 的**条件概率**（conditional probability）定义为

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

注意 $P(A \cap B) = P(A|B)P(B) = P(B \cap A) = P(B|A)P(A)$ 。

如果 A 和 B 独立，则

$$P(A|B) = P(A)$$

且 $P(A|B^c) = P(A)$ 。因此知道 B 是否发生不影响 A 发生的概率。

1.5.2 Properties of conditional probability

定理 1.8 1. $P(A \cap B) = P(A|B)P(B)$ 。

2. $P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C)$ 。

3. $P(A|B \cap C) = \frac{P(A \cap B|C)}{P(B|C)}$ 。

4. 如果 A 和 B 互斥，则 $P(A|B) = 0$ 。

5. 如果 $B \subseteq A$ ，则 $P(A|B) = 1$ 。

1.5.3 Law of total probability

考虑一个概率空间 (Ω, \mathcal{F}, P) 。一组事件 $\{B_1, \dots, B_n\}$ 称为样本空间 Ω 的一个划分(**partition**)，如果它满足以下性质：

- i. 完备性: $\bigcup_{i=1}^n B_i = \Omega$,
- ii. 互不相容: $B_i \cap B_j = \emptyset$ 对 $i \neq j$,
- iii. 概率非零: $P(B_i) > 0$ 对 $i = 1, \dots, n$ 。

定理 1.9 (全概率公式) 考虑概率空间 (Ω, \mathcal{F}, P) 及其划分 $\{B_1, \dots, B_n\}$ 。对所有 $A \in \mathcal{F}$,

$$P(A) = \sum_{i=1}^n P(A \cap B_i) = \sum_{i=1}^n P(A|B_i)P(B_i)$$

1.5.4 Bayes' formula

定理 1.10 (贝叶斯公式) 假设 $\{B_i\}_i$ 是样本空间的一个划分，且 A 是一个满足 $P(A) > 0$ 的事件。则对于划分中的任意事件 B_j ，有

$$P(B_j | A) = \frac{P(A | B_j)P(B_j)}{\sum_i P(A | B_i)P(B_i)}$$

1.6 Further exercises

1. 三箱零件与复杂检验

已知有三个箱子，箱中装有同种零件：

- 一号箱装 50 件，其中 10 件一等品。
- 二号箱装 30 件，其中 18 件一等品。
- 三号箱装 20 件，其中 12 件一等品。

先随机取一个箱子（每个箱子被选中的概率为 $1/3$ ），再从中随机取出一个零件。取出的零件是一等品。现在用一台不太可靠的仪器来检验这个零件的等级。如果零件实际是一等品，仪器有 0.9 的概率测为一等品，0.1 的概率测为次品。如果零件实际是次品，仪器有 0.8 的概率测为次品，0.2 的概率测为一等品。

求：在仪器测出该零件为一等品的条件下，它确实是来自一号箱的概率。

解：定义事件：

- B_i : 选到第 i 号箱 ($i = 1, 2, 3$), $P(B_i) = \frac{1}{3}$ 。
- A : 第一次取出的零件实际是一等品。
- E : 仪器测出该零件为一等品。

求 $P(B_1|E)$ 。

先计算先验概率：

$$P(A|B_1) = \frac{10}{50} = 0.2, \quad P(A|B_2) = \frac{18}{30} = 0.6, \quad P(A|B_3) = \frac{12}{20} = 0.6$$

$$P(A) = \sum P(B_i)P(A|B_i) = \frac{1}{3}(0.2 + 0.6 + 0.6) = \frac{1.4}{3}$$

计算 $P(E|B_i)$:

$$P(E|B_i) = P(E|A)P(A|B_i) + P(E|\bar{A})P(\bar{A}|B_i)$$

$$P(E|B_1) = 0.9 \times 0.2 + 0.2 \times 0.8 = 0.34$$

$$P(E|B_2) = 0.9 \times 0.6 + 0.2 \times 0.4 = 0.62$$

$$P(E|B_3) = 0.9 \times 0.6 + 0.2 \times 0.4 = 0.62$$

全概率公式求 $P(E)$:

$$P(E) = \sum P(B_i)P(E|B_i) = \frac{1}{3}(0.34 + 0.62 + 0.62) = \frac{1.58}{3}$$

贝叶斯公式：

$$P(B_1|E) = \frac{P(B_1)P(E|B_1)}{P(E)} = \frac{\frac{1}{3} \times 0.34}{\frac{1.58}{3}} = \frac{0.34}{1.58} = \frac{17}{79} \approx 0.215$$

答案： $\boxed{\frac{17}{79}}$

2. 双重贝叶斯问题

一家工厂有 A, B 两条生产线生产同一种产品。

- A 线的产量占总产量的 60%，次品率为 10%。
- B 线的产量占总产量的 40%，次品率为 20%。

质量检验员会进行两次检验：

- 第一次检验的准确率是 95%。
- 被第一次检验判为“次品”的产品，会送去进行第二次检验。第二次检验的准确率是 98%。

求：如果一个产品经过了两次检验，并且两次都被判定为“次品”，那么这个产品确实来自 A 生产线的概率是多少？

解：定义事件：

- A : 产品来自 A 生产线。 $P(A) = 0.6, P(\bar{A}) = 0.4$
- D : 产品是次品。 $P(D|A) = 0.1, P(D|\bar{A}) = 0.2$
- T_1^+ : 第一次检验判为次品。
- T_2^+ : 第二次检验判为次品。

检验准确性：

$$P(T_1^+|D) = 0.95, \quad P(T_1^+|\bar{D}) = 0.05$$

$$P(T_2^+|D) = 0.98, \quad P(T_2^+|\bar{D}) = 0.02$$

求 $P(A|T_1^+ \cap T_2^+)$ 。

计算似然：

$$\begin{aligned} P(T_1^+ \cap T_2^+|A) &= P(D|A)P(T_1^+|D)P(T_2^+|D) + P(\bar{D}|A)P(T_1^+|\bar{D})P(T_2^+|\bar{D}) \\ &= (0.1)(0.95)(0.98) + (0.9)(0.05)(0.02) = 0.0931 + 0.0009 = 0.094 \\ P(T_1^+ \cap T_2^+|\bar{A}) &= P(D|\bar{A})P(T_1^+|D)P(T_2^+|D) + P(\bar{D}|\bar{A})P(T_1^+|\bar{D})P(T_2^+|\bar{D}) \\ &= (0.2)(0.95)(0.98) + (0.8)(0.05)(0.02) = 0.1862 + 0.0008 = 0.187 \end{aligned}$$

全概率公式：

$$\begin{aligned} P(T_1^+ \cap T_2^+) &= P(A)P(T_1^+ \cap T_2^+|A) + P(\bar{A})P(T_1^+ \cap T_2^+|\bar{A}) \\ &= 0.6 \times 0.094 + 0.4 \times 0.187 = 0.0564 + 0.0748 = 0.1312 \end{aligned}$$

贝叶斯公式：

$$P(A|T_1^+ \cap T_2^+) = \frac{0.6 \times 0.094}{0.1312} = \frac{0.0564}{0.1312} \approx 0.4299$$

答案： $\boxed{\frac{0.0564}{0.1312} \approx 0.43}$

3. 信号传输与干扰

一个通信系统通过信道发送信号 0 或 1。

- 发送信号 0 的概率为 $\frac{2}{3}$ ，发送信号 1 的概率为 $\frac{1}{3}$ 。
- 传输特性：

– 发送 0 时，接收为 0 的概率是 0.8，接收为 1 的概率是 0.2。

– 发送 1 时，接收为 1 的概率是 0.9，接收为 0 的概率是 0.1。

接收端根据后验概率进行猜测。

求：整个通信过程中，猜错原始信号的概率。

解：定义事件：

- S_0 : 发送 0, $P(S_0) = 2/3$
- S_1 : 发送 1, $P(S_1) = 1/3$
- R_0 : 收到 0
- R_1 : 收到 1

计算收到概率：

$$P(R_0) = P(S_0)P(R_0|S_0) + P(S_1)P(R_0|S_1) = \frac{2}{3} \times 0.8 + \frac{1}{3} \times 0.1 = \frac{17}{30}$$

$$P(R_1) = 1 - P(R_0) = \frac{13}{30}$$

计算后验概率：

$$P(S_0|R_0) = \frac{P(S_0)P(R_0|S_0)}{P(R_0)} = \frac{(2/3)(0.8)}{17/30} = \frac{16}{17}$$

$$P(S_1|R_0) = 1 - \frac{16}{17} = \frac{1}{17}$$

$$P(S_1|R_1) = \frac{P(S_1)P(R_1|S_1)}{P(R_1)} = \frac{(1/3)(0.9)}{13/30} = \frac{9}{13}$$

$$P(S_0|R_1) = 1 - \frac{9}{13} = \frac{4}{13}$$

计算猜错概率：

$$\begin{aligned} P(\text{猜错}) &= P(R_0) \times P(S_1|R_0) + P(R_1) \times P(S_0|R_1) \\ &= \frac{17}{30} \times \frac{1}{17} + \frac{13}{30} \times \frac{4}{13} = \frac{1}{30} + \frac{4}{30} = \frac{5}{30} = \frac{1}{6} \end{aligned}$$

答案： 1
6

4. 疾病检测与先验不确定性

某种疾病在人群中的患病率 $P(D)$ 本身是一个随机变量：

- $P(P(D) = 0.001) = 0.7$
- $P(P(D) = 0.01) = 0.3$

检测方法的准确率为 99%。

求：在检测结果为阳性的条件下，该人群的患病率 $P(D)$ 为 0.01 的概率。

解：定义：

- H_1 : 人群患病率 $\theta = 0.001$, $P(H_1) = 0.7$
- H_2 : 人群患病率 $\theta = 0.01$, $P(H_2) = 0.3$

- T^+ : 检测阳性

已知 $P(T^+|D) = 0.99, P(T^+|\bar{D}) = 0.01$ 。

计算 $P(T^+|H_i)$:

$$P(T^+|H_1) = 0.001 \times 0.99 + 0.999 \times 0.01 = 0.01098$$

$$P(T^+|H_2) = 0.01 \times 0.99 + 0.99 \times 0.01 = 0.0198$$

全概率公式:

$$P(T^+) = P(H_1)P(T^+|H_1) + P(H_2)P(T^+|H_2) = 0.7 \times 0.01098 + 0.3 \times 0.0198 = 0.013626$$

贝叶斯公式:

$$P(H_2|T^+) = \frac{P(H_2)P(T^+|H_2)}{P(T^+)} = \frac{0.3 \times 0.0198}{0.013626} \approx 0.436$$

答案: 0.436

5. 网络路径选择

数据包从节点 A 发送到节点 D:

- 选择路径 A-B-D 的概率是 p
- 选择路径 A-C-D 的概率是 $1 - p$

路径可靠性:

- 路径 A-B-D: 成功率 $\alpha\beta$
- 路径 A-C-D: 成功率 $\gamma\beta$

求: 若已知数据包成功从 A 发送到了 D, 求它选择了路径 A-B-D 的概率。

解: 定义:

- R : 选择路径 A-B-D, $P(R) = p$
- S : 成功从 A 到 D

$$P(S|R) = \alpha\beta, \quad P(S|\bar{R}) = \gamma\beta$$

全概率公式:

$$P(S) = P(R)P(S|R) + P(\bar{R})P(S|\bar{R}) = p\alpha\beta + (1-p)\gamma\beta = \beta(p\alpha + (1-p)\gamma)$$

贝叶斯公式:

$$P(R|S) = \frac{P(R)P(S|R)}{P(S)} = \frac{p\alpha\beta}{\beta(p\alpha + (1-p)\gamma)} = \frac{p\alpha}{p\alpha + (1-p)\gamma}$$

当 $\alpha = \gamma$ 时:

$$P(R|S) = \frac{p\alpha}{p\alpha + (1-p)\alpha} = \frac{p\alpha}{\alpha} = p$$

答案: $\frac{p\alpha}{p\alpha + (1-p)\gamma}$, 当 $\alpha = \gamma$ 时, $P(R|S) = p$

1.7 Appendix: Proofs

2 Random variables and univariate distributions 随机变量和单变量分布

2.1 Random variables

2.1.1 Mapping outcomes to real numbers

定义 2.1 (随机变量) 一个随机变量 (*random variable*) X 是一个函数: $\Omega \rightarrow \mathbb{R}$, 且满足性质: 若

$$A_x = \{\omega \in \Omega : X(\omega) \leq x\}$$

则对所有的 $x \in \mathbb{R}$, 有 $A_x \in \mathcal{F}$ 。因此, 对每个实数值 x , A_x 是一个事件。

随机变量 X 是一个函数, 如果 ω 是一个结果, 那么 $X(\omega)$ 是一个实数。这与我们对随机变量的直观定义——其值由试验结果决定的量——是一致的。

定义 2.2 (随机变量的函数) 如果 $g: \mathbb{R} \rightarrow \mathbb{R}$ 是一个性质良好的函数, $X: \Omega \rightarrow \mathbb{R}$ 是一个随机变量, 且 $Y = g(X)$, 那么 Y 是一个随机变量, $Y: \Omega \rightarrow \mathbb{R}$, 且对所有 $\omega \in \Omega$, 有 $Y(\omega) = g(X(\omega))$ 。

定义 2.3 (函数的支撑集) 一个正实值函数 f 的支撑集 (*support*) 是实直线中使得 f 取值严格大于零的子集, 即 $\{x \in \mathbb{R} : f(x) > 0\}$ 。

2.2 Discrete random variables

2.2.1 Discrete random variables

定义 2.4 (离散随机变量) 一个随机变量 X 被称为离散 (*discrete*) 的, 如果它只取值于 \mathbb{R} 的某个可数子集 $\{x_1, x_2, \dots\}$ 中的值。

2.2.2 Probability mass function and cumulative distribution function

我们通常更关注与随机变量相关的概率, 而非从结果到实数的映射。与随机变量相关的概率可以通过几个函数来完全刻画。

定义 2.5 (概率质量函数) 离散随机变量 X 的概率质量函数 (*probability mass function, p.m.f.*) 是函数 $f_X: \mathbb{R} \rightarrow [0, 1]$, 定义为

$$f_X(x)(= p_x) = P(X = x)$$

它是在 $\{x_1, x_2, \dots\}$ 上的一个概率分布。

概率质量函数通常简称为质量函数。质量函数的性质很容易陈述如下。任何满足这些性质的函数都是一个有效的质量函数。

定理 2.1 (质量函数的性质) 如果 f_X 是一个质量函数, 那么

- i. 对所有 x , 有 $0 \leq f_X(x) \leq 1$ 。
- ii. 如果 $x \notin \{x_1, x_2, \dots\}$, 则 $f_X(x) = 0$ 。
- iii. $\sum_x f_X(x) = 1$ 。

显然, 离散变量质量函数的支撑集是该变量可以取值的可数集 $\{x_1, x_2, \dots\}$ 。

定义 2.6 (累积分布函数) 随机变量 X (离散、连续或其他) 的累积分布函数 (*cumulative distribution function, c.d.f.*) (或简称分布函数) 是函数 $F_X: \mathbb{R} \rightarrow [0, 1]$, 定义为

$$F_X(x) = P(X \leq x)$$

定理 2.2 (累积分布函数的性质) 如果 F_X 是一个累积分布函数, 那么

- i. F_X 是一个非递减函数。
- ii. $\lim_{x \rightarrow -\infty} F_X(x) = 0$ 且 $\lim_{x \rightarrow \infty} F_X(x) = 1$ 。
- iii. F_X 是右连续的: 对所有 $x \in \mathbb{R}$, 有 $F_X(x+) = F_X(x)$ 。

累积分布函数告诉我们低于某点的概率。显然还有其他感兴趣的概率。以下命题说明了如何使用累积分布函数计算这些概率。

定理 2.3 (从累积分布函数计算概率) 考虑实数 a 和 b , 且 $a < b$ 。那么:

- i. $P(X > a) = 1 - P(X \leq a) = 1 - F_X(a)$
- ii. $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a)$
- iii. $P(X < a) = F_X(a-)$
- iv. $P(X = a) = P(X \leq a) - P(X < a) = F_X(a) - F_X(a-) (= f_X(a))$

定理 2.4 (质量函数与分布函数的关系) 如果 X 是一个离散随机变量, f_X 是 X 的质量函数, F_X 是 X 的累积分布函数, 那么

- i. $f_X(x) = P(X = x) = F_X(x) - F_X(x-)$ 。
- ii. $F_X(x) = \sum_{u: u \leq x} f_X(u)$ 。

注意, 此性质的第一部分意味着 $F_X(x) = F_X(x-) + f_X(x)$ 。我们知道如果 $x \notin \{x_1, x_2, \dots\}$, 则 $f_X(x) = 0$ 。因此, 对于 $x \notin \{x_1, x_2, \dots\}$, 有 $F_X(x) = F_X(x-)$ 。这表明累积分布函数除了在点 $\{x_1, x_2, \dots\}$ 处有间断外, 其余部分是平坦的; 这类函数被称为阶梯函数。

2.3 Continuous random variables

2.3.1 Continuous random variables

到目前为止, 我们一直在考虑可能结果集 Ω 是有限或可数的试验。现在我们允许可能结果是连续的情况。连续随机变量可以定义为其累积分布函数是连续的变量。

2.3.2 Probability density function and cumulative distribution function

定义 2.7 (连续随机变量) 一个随机变量 X 是连续 (*continuous*) 的, 如果它的分布函数可以表示为

$$F_X(x) = \int_{-\infty}^x f_X(u) du \quad \text{对于 } x \in \mathbb{R}$$

其中 $f_X: \mathbb{R} \rightarrow [0, \infty)$ 是一个可积函数。函数 f_X 称为 X 的 (概率) 密度函数 (*probability density function, p.d.f.*)。

该定义展示了如何通过对密度函数积分来找到随机变量的分布函数。我们也可以反向操作: 通过对累积分布函数求导来找到密度函数。

定理 2.5 (从累积分布函数求密度函数) 对于具有累积分布函数 F_X 的连续随机变量 X , 其密度函数由下式给出:

$$f_X(x) = \frac{d}{du} F_X(u) \Big|_{u=x} = F'_X(x) \quad \text{对所有 } x \in \mathbb{R}$$

定理 2.6 (密度函数的性质) 如果 f_X 是一个密度函数, 那么

- i. 对所有 $x \in \mathbb{R}$, 有 $f_X(x) \geq 0$ 。
- ii. $\int_{-\infty}^{+\infty} f_X(x) dx = 1$ 。

我们对密度函数使用与质量函数相同的记号。这旨在强调密度对于连续变量所起的作用与质量对于离散变量所起的作用相同。然而, 有一个重要的区别: 密度函数的值可以大于 1 是合理的, 因为密度函数的值并不给出概率。

定理 2.7 (由密度函数求概率) 如果 X 是一个具有密度 f_X 的连续随机变量, 且 $a, b \in \mathbb{R}$ 满足 $a \leq b$, 那么

$$P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x) dx$$

关于概率密度函数及其用途的直觉可以从以下近似关系中获得:

$$P(X \in [x, x + \delta x]) = \int_x^{x+\delta x} f(z) dz \approx f(x)\delta x$$

然而, 请记住 $f(x)$ 不是一个概率。实际上, 对于 $x \in \mathbb{R}$, 有 $P(X = x) = 0$ 。因此根据公理 III, 如果 A 是 \mathbb{R} 的任何可数子集, 我们必须得出 $P(X \in A) = 0$ 的结论。

定理 2.8 (连续随机变量在单点处的概率) 如果 X 是一个连续随机变量, 那么

$$P(X = x) = 0 \quad \text{对所有 } x \in \mathbb{R}$$

2.4 Mixed random variables

2.4.1 Mixed random variables

混合随机变量 (mixed random variable) 是既非离散也非连续, 而是两者混合的随机变量。具体来说, 一个混合随机变量包含连续部分和离散部分。

例题

设 X 是一个具有以下概率密度函数的连续随机变量:

$$f_X(x) = \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & \text{其他} \end{cases}$$

同时令

$$Y = g(X) = \begin{cases} X & 0 \leq X \leq \frac{1}{2} \\ \frac{1}{2} & X > \frac{1}{2} \end{cases}$$

求 Y 的累积分布函数。

解:

$$F_Y(y) = \begin{cases} 1 & y \geq \frac{1}{2} \\ y^2 & 0 \leq y < \frac{1}{2} \\ 0 & \text{其他} \end{cases}$$

我们注意到该累积分布函数不是连续的，因此 Y 不是连续随机变量。另一方面，该累积分布函数也不是阶梯函数形式，因此它也不是离散随机变量。它确实是一个混合随机变量。

一般来说，混合随机变量 Y 的累积分布函数可以写为一个连续函数和一个阶梯函数之和：

$$F_Y(y) = C(y) + D(y)$$

设 $\{y_1, y_2, y_3, \dots\}$ 是 $D(y)$ 的跳跃点集合，即满足 $P(Y = y_k) > 0$ 的点。那么我们有

$$\int_{-\infty}^{\infty} c(y) dy + \sum_{y_k} P(Y = y_k) = 1$$

2.4.2 Functions that uniquely define a probability distribution

累积分布函数唯一地定义了一个分布。连续变量具有光滑的累积分布函数，离散变量具有“阶梯状”的累积分布函数，而有些累积分布函数是混合的。对于离散变量，概率质量函数唯一地定义了一个分布，因为改变 $f(x_0)$ 也会改变 $F(x_0)$ 。然而，概率密度函数并不唯一地定义连续分布，因为改变其在某一点的值不会改变任何定积分。例如 $X \sim U(0, 1)$ 和 $Y \sim U[0, 1]$ 具有相同的分布但不同的概率密度函数。可以在可数多个点上改变概率密度函数而保持累积分布函数不变。在这方面，加法与积分是不同的。

表 3: 描述概率分布的方法

方式	描述函数
直接	1. 累积分布函数 * (适用于所有随机变量，包括离散、连续和混合) 2. 概率质量函数 * (适用于所有离散随机变量)，概率密度函数 (适用于所有连续随机变量)
间接	3. 概率母函数 *、矩母函数 * (将在后续介绍)

注：* 表示该函数唯一地定义了概率分布，即该概率分布被完整地描述。

对于所有随机变量，

$$\text{随机变量} \implies \text{概率分布} \iff \text{累积分布函数}$$

每个随机变量背后都有一个（概率）分布。不同的随机变量可能具有相同的分布。每个分布都由一个累积分布函数唯一地定义。如果两个分布相同，则它们应有相同的累积分布函数；如果两个随机变量的概率分布具有相同的累积分布函数，则它们具有相同的分布。对于所有离散随机变量，概率质量函数也唯一地定义了它们的分布，因此唯一地对应于其累积分布函数。然而，对于连续随机变量，两个不同的概率密度函数可能对应于相同的累积分布函数，从而对应于相同的分布。因此，分布不能由概率密度函数唯一地定义。对于混合随机变量，它们甚至没有有效的概率质量函数或概率密度函数。如果存在，一个分布也可以由其概率母函数或矩母函数唯一地定义。

2.5 Expectation, variance, and higher moments

2.5.1 Expectation

定义 2.8 (期望/均值) 随机变量 X 的期望 (*expectation*) (或均值 (*mean*)) 存在, 且等于数值

$$\mathbb{E}[X] = \begin{cases} \sum_x x f_X(x) & \text{若 } X \text{ 是离散的} \\ \int_{-\infty}^{\infty} x f_X(x) dx & \text{若 } X \text{ 是连续的} \end{cases}$$

要求离散情况下该求和绝对收敛, 连续情况下 $\int_0^\infty x f(x) dx$ 和 $\int_{-\infty}^0 x f(x) dx$ 不均为无穷。

定理 2.9 (随机变量函数的期望) 对于任何性质良好的函数 $g: \mathbb{R} \rightarrow \mathbb{R}$, $g(X)$ 的期望定义为

$$\mathbb{E}[g(X)] = \begin{cases} \sum_x g(x) f_X(x) & \text{若 } X \text{ 是离散的} \\ \int_{-\infty}^{\infty} g(x) f_X(x) dx & \text{若 } X \text{ 是连续的} \end{cases}$$

定理 2.10 (期望的性质)

1. 如果 $X \geq 0$, 则 $\mathbb{E}[X] \geq 0$ 。
2. 如果 $X \geq 0$ 且 $\mathbb{E}[X] = 0$, 则 $P(X = 0) = 1$ 。
3. 如果 a 和 b 是常数, 则 $\mathbb{E}[a + bX] = a + b\mathbb{E}[X]$ 。
4. 对于任意随机变量 X, Y , 有 $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ 。
* 性质 3 和 4 表明 \mathbb{E} 是一个线性算子。*
5. $\mathbb{E}[X]$ 是使 $\mathbb{E}[(X - c)^2]$ 最小的常数。

2.5.2 Variance

定义 2.9 (方差与标准差) 随机变量 X 的方差 (*variance*) 定义为

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

标准差 (*standard deviation*) 为

$$\sqrt{\text{Var}(X)}$$

定理 2.11 (方差的性质)

- (i) $\text{Var}(X) \geq 0$ 。如果 $\text{Var}(X) = 0$, 则 $P(X = \mathbb{E}[X]) = 1$ 。
- (ii) 如果 a, b 是常数, 则 $\text{Var}(a + bX) = b^2 \text{Var}(X)$ 。
- (iii) $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ 。
- (iv) $\text{Var}(X) = \mathbb{E}[X(X - 1)] - \mathbb{E}(X)\mathbb{E}(X - 1)$ 。

2.5.3 Inequalities involving expectation

在证明收敛结果时, 能够为概率和期望提供界限通常很有用。

(a) 简森不等式

定义 2.10 (凸函数) 函数 $f: (a, b) \rightarrow \mathbb{R}$ 是凸函数, 如果对所有 $x_1, x_2 \in (a, b)$ 和满足 $\lambda_1 + \lambda_2 = 1$ 的 $\lambda_1 \geq 0, \lambda_2 \geq 0$, 有

$$\lambda_1 f(x_1) + \lambda_2 f(x_2) \geq f(\lambda_1 x_1 + \lambda_2 x_2)$$

如果当 $x_1 \neq x_2$ 且 $0 < \lambda_1 < 1$ 时严格不等式成立, 则称其为严格凸函数。

定理 2.12 (简森不等式) 设 $f : (a, b) \rightarrow \mathbb{R}$ 是一个凸函数。那么对于所有 $x_1, \dots, x_n \in (a, b)$ 和满足 $\sum_i p_i = 1$ 的 $p_1, \dots, p_n \in (0, 1)$, 有

$$\sum_{i=1}^n p_i f(x_i) \geq f\left(\sum_{i=1}^n p_i x_i\right)$$

此外, 如果 f 是严格凸的, 则等号成立当且仅当所有 x_i 都相等。

简森不等式表明, 如果 X 取有限多个值, 则

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

(b) 算术-几何平均不等式

定理 2.13 (算术-几何平均不等式) 给定正实数 x_1, \dots, x_n , 有

$$\left(\prod_{i=1}^n x_i\right)^{1/n} \leq \frac{1}{n} \sum_{i=1}^n x_i$$

(c) 柯西-施瓦茨不等式

定理 2.14 (柯西-施瓦茨不等式) 对于任意随机变量 X 和 Y , 有

$$\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$$

(d) 马尔可夫不等式

定理 2.15 (马尔可夫不等式) 如果 X 是一个满足 $\mathbb{E}[|X|] < \infty$ 的随机变量, 且 $a > 0$, 那么

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

(e) 切比雪夫不等式

定理 2.16 (切比雪夫不等式) 如果 X 是一个满足 $\mathbb{E}[X^2] < \infty$ 的随机变量, 且 $\epsilon > 0$, 那么

$$P(|X| \geq \epsilon) \leq \frac{\mathbb{E}[X^2]}{\epsilon^2}$$

将该不等式应用于 $X - \mathbb{E}[X]$ 可得

$$P(|X - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\text{Var}(X)}{\epsilon^2}$$

定理 2.17 (标准化分布的切比雪夫不等式) 如果 X 是一个具有 $\mu = \mathbb{E}[X]$ 和 $\sigma^2 = \text{Var}(X)$ 的随机变量, 且 $\lambda > 0$ 是一个实常数, 那么

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

2.5.4 Moments

我们已经讨论了集中趋势的度量和离散程度的度量。顾名思义, 集中趋势给出了分布中心位置的指示, 而离散程度度量了概率分布的广泛程度。分布的其他可能令人感兴趣的特征包括对称性和尾部概率的程度 (尾部的厚度)。我们可以用矩 (**moments**) 和中心矩 (**central moments**) 来表示常用的集中趋势、离散程度、对称性和尾部厚度的度量。

定义 2.11 (矩/原点矩) 对于随机变量 X 和正整数 r , X 的 r 阶矩记为 μ'_r , 其中

$$\mu'_r = \mathbb{E}[X^r]$$

矩依赖于分布的水平位置。当我们测量像离散程度这样的特征时, 我们希望使用一个在分布沿水平轴左右移动时保持不变的量。这启发了中心矩的定义, 其中我们执行平移以考虑均值的值。

定义 2.12 (中心矩) 对于随机变量 X 和正整数 r , X 的 r 阶中心矩记为 μ_r , 其中

$$\mu_r = \mathbb{E}[(X - \mathbb{E}[X])^r]$$

注意, 现在我们有多种不同的方式来指代同一事物。例如,

$$\mu'_1 = \mathbb{E}[X] = \mu$$

$$\mu_2 = \mathbb{E}[(X - \mathbb{E}[X])^2] = \text{Var}(X) = \sigma^2$$

注意

$$\mu_1 = \mathbb{E}[X - \mu'_1] = 0$$

我们已经看到, 一阶矩 (均值) 是常用的集中趋势度量, 二阶中心矩 (方差) 是常用的离散程度度量。以下定义提供了基于矩的对称性和尾部厚度的度量。

定义 2.13 (偏度系数与峰度系数) 对于方差 $\text{Var}(X) = \sigma^2 < \infty$ 的随机变量 X , 其偏度 (*skewness*) 系数由下式给出:

$$\gamma_1 = \frac{\mathbb{E}[(X - \mu)^3]}{\sigma^3} = \frac{\mu_3}{\mu_2^{3/2}}$$

其峰度 (*kurtosis*) 系数由下式给出:

$$\gamma_2 = \left(\frac{\mathbb{E}[(X - \mu)^4]}{\sigma^4} \right) - 3 = \frac{\mu_4}{\mu_2^2} - 3$$

表 4: 矩与中心矩

r	原点矩 μ'_r	中心矩 μ_r
$r = 1$	$\mathbb{E}[X]$	0
$r = 2$	$\mathbb{E}[X^2]$	$\text{Var}(X)$
$r = 3$	$\mathbb{E}[X^3]$	μ_3
\vdots	\vdots	\vdots

2.6 Common distributions of random variables

2.6.1 Common discrete distributions

定义 2.14 (伯努利分布 (0-1 分布)) 考虑抛掷一次硬币, 可能的结果为 $\Omega = \{H, T\}$ 。对于 $p \in [0, 1]$, 伯努利分布 (0-1 分布), 记为 $B(1, p)$, 其概率质量函数为:

$$f_X(x) = \begin{cases} 1-p & \text{当 } x=0 \\ p & \text{当 } x=1 \\ 0 & \text{其他} \end{cases}$$

期望: $\mathbb{E}[X] = p$, 方差: $\text{Var}(X) = p(1-p)$ 。

定义 2.15 (二项分布) 将上述硬币抛掷 n 次，我们得到一系列伯努利试验。

$$P(\text{次成功}) = f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n$$

这是**二项分布**，记为 $B(n, p)$ 。

期望： $\mathbb{E}[X] = np$ ，方差： $\text{Var}(X) = np(1-p)$ 。

二项分布是 n 次独立伯努利试验中成功次数的分布。

例题 2.1 某工厂生产的产品次品率为 5%，从这批产品中随机抽取 10 件，求恰好有 2 件次品的概率。

解：设 X 为次品数，则 $X \sim B(10, 0.05)$ 。

$$P(X = 2) = \binom{10}{2} (0.05)^2 (0.95)^8 \approx 0.0746$$

定义 2.16 (几何分布) 考虑一个无限的伯努利试验序列，其中 $P(\text{成功}) = 1 - P(\text{失败}) = p$ 。首次成功所需试验次数的概率质量函数为：

$$f_X(x) = (1-p)^{x-1} p, \quad x = 1, 2, \dots$$

这是参数为 p 的**几何分布**。

期望： $\mathbb{E}[X] = \frac{1}{p}$ ，方差： $\text{Var}(X) = \frac{1-p}{p^2}$ 。

几何分布具有**无记忆性** (*memoryless property*)： $P(X > m+n | X > m) = P(X > n)$ 。

例题 2.2 连续抛掷一枚均匀硬币，求第一次出现正面时所需抛掷次数的期望。

解：设 X 为首次出现正面所需的抛掷次数，则 $X \sim \text{Geometric}(0.5)$ 。

$$\mathbb{E}[X] = \frac{1}{0.5} = 2$$

定义 2.17 (泊松分布) 泊松分布常用于模拟在特定时间内某事件发生的次数，例如保险公司一年内的理赔次数。记为 $\text{Pois}(\lambda)$ ，参数 $\lambda > 0$ 的泊松分布为：

$$f_X(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, \dots$$

期望： $\mathbb{E}[X] = \lambda$ ，方差： $\text{Var}(X) = \lambda$ 。

定理 2.18 (泊松极限定理) 假设 $X \sim \text{Bin}(n, p)$ 。如果令 $n \rightarrow \infty$ 且 $p \rightarrow 0$ ，使得 $np \rightarrow \lambda$ ，那么

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \rightarrow \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, \dots$$

例题 2.3 某交叉路口每小时平均发生 3 起交通事故。求一小时内恰好发生 2 起事故的概率。

解：设 X 为一小时内事故数，则 $X \sim \text{Pois}(3)$ 。

$$P(X = 2) = \frac{3^2}{2!} e^{-3} = \frac{9}{2} e^{-3} \approx 0.2240$$

定义 2.18 (离散均匀分布) 离散均匀分布在其支撑集的每个成员上分配相等的概率。因此，如果 X 是一个支撑集为 $\{x_1, \dots, x_n\}$ 的离散均匀分布，则 X 的概率质量函数为：

$$f_X(x) = \begin{cases} \frac{1}{n} & \text{当 } x \in \{x_1, \dots, x_n\} \\ 0 & \text{其他} \end{cases}$$

期望： $\mathbb{E}[X] = \frac{1}{n} \sum_{i=1}^n x_i$ ，方差： $\text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \mathbb{E}[X])^2$ 。

2.6.2 Common continuous distributions

定义 2.19 (连续均匀分布) 区间 $[a, b]$ 上的连续均匀分布的累积分布函数和相应的概率密度函数为：

$$F_X(x) = \frac{x-a}{b-a}, \quad f_X(x) = \frac{1}{b-a}, \quad a \leq x \leq b$$

如果 X 服从此分布，我们记作 $X \sim U[a, b]$ 。

期望： $\mathbb{E}[X] = \frac{a+b}{2}$ ，方差： $\text{Var}(X) = \frac{(b-a)^2}{12}$ 。

例题 2.4 某公交车每 15 分钟一班，乘客随机到达车站。求乘客等待时间不超过 5 分钟的概率。

解：设 X 为等待时间，则 $X \sim U[0, 15]$ 。

$$P(X \leq 5) = \frac{5-0}{15-0} = \frac{1}{3}$$

定义 2.20 (指数分布) 参数为 $\lambda > 0$ 的指数分布的累积分布函数和概率密度函数为：

$$F_X(x) = 1 - e^{-\lambda x}, \quad f_X(x) = \lambda e^{-\lambda x}, \quad 0 \leq x < \infty$$

如果 X 服从此分布，我们记作 $X \sim Exp(\lambda)$ 。注意 X 是非负的。

参数 λ 称为速率。指数分布也常用其尺度参数 θ 表示，其中 $\theta = 1/\lambda$ 。

期望： $\mathbb{E}[X] = \frac{1}{\lambda}$ ，方差： $\text{Var}(X) = \frac{1}{\lambda^2}$ 。

指数分布具有无记忆性 (memoryless property)。如果 $X \sim Exp(\lambda)$ ，则

$$P(X \geq x+z | X \geq z) = \frac{P(X \geq x+z)}{P(X \geq z)} = \frac{e^{-\lambda(x+z)}}{e^{-\lambda z}} = e^{-\lambda x} = P(X \geq x)$$

如果 X 表示某物的寿命，无记忆性表明在已经持续了 z 时间后，剩余寿命的分布与初始时相同。

离散分布中具有无记忆性的是几何分布。即对于正整数 k 和 h ，

$$P(X \geq k+h | X \geq k) = \frac{P(X \geq k+h)}{P(X \geq k)} = \frac{q^{k+h}}{q^k} = q^h = P(X \geq h)$$

例题 2.5 某电子元件的寿命服从参数为 $\lambda = 0.01$ (小时⁻¹) 的指数分布。求该元件在使用 50 小时后还能继续工作 100 小时的概率。

解：由无记忆性，

$$P(X \geq 150 | X \geq 50) = P(X \geq 100) = e^{-0.01 \times 100} = e^{-1} \approx 0.3679$$

定义 2.21 (正态分布) 参数为 μ 和 σ^2 的正态分布 (或高斯分布) 的概率密度函数为：

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

如果 X 服从此分布，我们记作 $X \sim N(\mu, \sigma^2)$ 。

标准正态分布指 $N(0, 1)$ ，其累积分布函数通常记为 $\Phi(z) = F_Z(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$ ，且 $\Phi(-z) = 1 - \Phi(z)$ 。

期望： $\mathbb{E}[X] = \mu$ ，方差： $\text{Var}(X) = \sigma^2$ 。

例题 2.6 某考试分数服从正态分布 $N(70, 100)$ ，求分数超过 85 分的概率。

解： $P(X > 85) = 1 - \Phi\left(\frac{85-70}{10}\right) = 1 - \Phi(1.5) \approx 1 - 0.9332 = 0.0668$

定义 2.22 (Gamma 分布) *Gamma* 分布的随机变量只能取正值。该分布由两个正参数表征，通常称为形状参数 α 和速率参数 λ 。如果 $X \sim \text{Gamma}(\alpha, \lambda)$ ，则 X 的密度函数为：

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x} & \text{当 } 0 < x < \infty \\ 0 & \text{其他} \end{cases}$$

我们可以立即观察到指数分布是 *Gamma* 分布的特例。如果 $Y \sim \text{Exp}(\lambda)$ ，则 $Y \sim \text{Gamma}(1, \lambda)$ 。
期望： $\mathbb{E}[X] = \frac{\alpha}{\lambda}$ ，方差： $\text{Var}(X) = \frac{\alpha}{\lambda^2}$ 。

定义 2.23 (Beta 分布) *Beta* 分布是一种定义在区间 $(0, 1)$ 上的连续概率分布，由两个正形状参数 α 和 β 表征。如果 $X \sim \text{Beta}(\alpha, \beta)$ ，则其概率密度函数为：

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, & \text{当 } 0 < x < 1 \\ 0, & \text{其他} \end{cases}$$

其中 $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ 是 *Beta* 函数。

我们可以观察到，当 $\alpha = \beta = 1$ 时，*Beta* 分布退化为区间 $(0, 1)$ 上的均匀分布。

期望： $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$ ，方差： $\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ 。

定义 2.24 (指数族) 如果一个分布的质量函数或密度函数可以写成以下形式，则它属于**指数族 (exponential family)**：

$$f_Y(y) = h(y)a(\theta) \exp \left[\sum_{i=1}^k \eta_i(\theta)t_i(y) \right]$$

其中 θ 是参数向量，且 $h(y) \geq 0$ 。对于任何给定的族，族成员仅在参数向量 θ 的值上有所不同。特殊情况 $k = 1$, $\eta_1(\theta) = \theta$ ，且 $t_1(y) = y$ 被称为（一阶）**自然指数族**。

构成指数族的离散分布类包括二项分布、泊松分布和负二项分布，连续分布包括 *Beta* 分布、*Gamma* 分布和正态分布。

2.6.3 Parameters and families of distributions

参数是分布的一个感兴趣的特征。分布族是仅在其参数值上有所不同的一组分布。

2.6.4 Univariate distribution relationships

了解分布之间的相互关系是有趣且有用的。例如，泊松分布是二项分布在 $n \rightarrow \infty$ 且 $np \rightarrow \lambda$ 时的极限。如果 X 和 Y 是独立同分布的指数随机变量，则 $X/(X+Y)$ 服从均匀分布。

2.7 Generating functions

2.7.1 Probability generating function

定义 2.25 (概率母函数) 考虑一个取值为 $0, 1, 2, \dots$ 的随机变量 X 。令 $p_r = P(X = r), r = 0, 1, 2, \dots$ 。 X 或分布 $(p_r, r = 0, 1, 2, \dots)$ 的**概率母函数 (probability generating function, p.g.f.)** 为

$$p(z) = \mathbb{E}[z^X] = \sum_{r=0}^{\infty} P(X = r)z^r = \sum_{r=0}^{\infty} p_r z^r$$

因此 $p(z)$ 是一个多项式或幂级数。作为幂级数，通过与几何级数比较，它在 $|z| \leq 1$ 时收敛，且

$$|p(z)| \leq \sum_r p_r |z|^r \leq \sum_r p_r = 1$$

当我们想提醒这是 X 的概率母函数时，可以写为 $p_X(z)$ 。

定理 2.19 (概率母函数的唯一性) X 的分布由概率母函数 $p(z)$ 唯一确定。

定理 2.20 (Abel 引理)

$$\mathbb{E}[X] = \lim_{z \rightarrow 1} p'(z)$$

定理 2.21

$$\mathbb{E}[X(X - 1)] = \lim_{z \rightarrow 1} p''(z)$$

定理 2.22 (独立随机变量和的概率母函数) 如果 X_1, X_2, \dots, X_n 是 n 个独立的随机变量，其概率母函数分别为 $p_1(z), p_2(z), \dots, p_n(z)$ ，则 $X_1 + X_2 + \dots + X_n$ 的概率母函数为

$$p_1(z)p_2(z) \dots p_n(z)$$

2.7.2 Moment generating function

对于许多分布，所有矩 $\mathbb{E}[X], \mathbb{E}[X^2], \dots$ 可以被封装在一个单一函数中。该函数称为矩母函数，它存在于许多常用分布中。它通常为计算矩提供最有效的方法。矩母函数在建立分布结果（例如随机变量和的性质）以及证明渐近结果时也很有用。

定义 2.26 (矩母函数) 随机变量 X 的矩母函数 (*moment generating function, m.g.f.*) 是一个函数 $M_X : \mathbb{R} \rightarrow [0, \infty)$ ，定义为

$$M_X(t) = \mathbb{E}[e^{tX}] = \begin{cases} \sum_x e^{tx} f_X(x) & \text{若 } X \text{ 是离散的} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x) dx & \text{若 } X \text{ 是连续的} \end{cases}$$

为了使函数定义良好，我们要求对于某个 $h > 0$ ，所有 $t \in [-h, h]$ 都有 $M_X(t) < \infty$ 。

矩母函数适用于任何类型的随机变量，但最常用于连续随机变量，而概率母函数最常用于离散随机变量。

X 的矩母函数是 X 的指数函数的期望值。矩母函数的有用性质继承自指数函数 e^x 。在零点的泰勒级数展开将 e^x 表示为 x 的多项式：

$$e^x = 1 + x + \frac{1}{2!}x^2 + \dots + \frac{1}{r!}x^r + \dots = \sum_{j=0}^{\infty} \frac{1}{j!}x^j$$

此展开式（以及任何在零点的泰勒级数展开）通常称为麦克劳林级数展开。 e^x 的所有导数都等于 e^x ：

$$\frac{d^r}{dx^r} e^x = e^x \quad \text{对于 } r = 1, 2, \dots$$

这可以通过对麦克劳林级数展开进行微分来验证。这些观察引出了两个直接用于计算矩的命题。

定理 2.23 (将矩母函数表示为多项式) 矩母函数 $M_X(t)$ 可以表示为 t 的多项式：

$$M_X(t) = 1 + t\mathbb{E}[X] + \frac{t^2}{2!}\mathbb{E}[X^2] + \dots + \frac{t^r}{r!}\mathbb{E}[X^r] + \dots = \sum_{j=0}^{\infty} \frac{\mathbb{E}[X^j]}{j!}t^j$$

定理 2.24 (矩母函数在零点的导数) 矩母函数在零点的 r 阶导数是 r 阶矩:

$$M_X^{(r)}(0) = \frac{d^r}{dt^r} M_X(t) \Big|_{t=0} = \mathbb{E}[X^r]$$

以上两个命题为我们提供了两种从矩母函数计算矩的机制。

1. 矩母函数展开式中 t^r 的系数是 r 阶矩除以 $r!$ 。我们可以通过比较系数来计算 r 阶矩, 即:

$$\text{如果 } M_X(t) = \sum_{j=0}^{\infty} a_j t^j, \text{ 则 } \mathbb{E}[X^r] = r! a_r$$

2. 我们可以直接使用命题 2.24 计算 r 阶矩, 即对矩母函数关于 t 求 r 阶导数并在 $t = 0$ 处求值:

$$\mathbb{E}[X^r] = \frac{d^r}{dt^r} M_X(t) \Big|_{t=0}$$

通常, 比较系数更快。然而, 对于某些矩母函数, 将其展开为 t 的多项式不容易。在这些情况下, 微分为计算矩提供了另一种方法。

矩母函数用于建立随机变量和的性质, 以及证明收敛结果。其有用性的关键在于矩母函数唯一地刻画了一个分布; 如果随机变量具有相同的矩母函数, 则它们具有相同的分布。

定理 2.25 (矩母函数的唯一性) 如果 X 和 Y 是随机变量, 并且我们能找到 $h > 0$ 使得对于所有 $t \in [-h, h]$ 有 $M_X(t) = M_Y(t)$, 那么对于所有 $x \in \mathbb{R}$ 有 $F_X(x) = F_Y(x)$ 。

定理 2.26 (独立随机变量和的矩母函数) 如果 X_1, X_2, \dots, X_n 是 n 个独立的随机变量, 其矩母函数分别为 $M_{X_1}(s), M_{X_2}(s), \dots, M_{X_n}(s)$, 则 $X_1 + X_2 + \dots + X_n$ 的矩母函数为:

$$M_{X_1+X_2+\dots+X_n}(s) = M_{X_1}(s)M_{X_2}(s)\dots M_{X_n}(s)$$

2.7.3 Cumulant generating functions and cumulants

使用矩母函数的对数通常很方便。事实证明, 矩母函数对数的多项式展开系数在矩和中心矩方面有方便的解释。

定义 2.27 (累积量母函数与累积量) 具有矩母函数 $M_X(t)$ 的随机变量 X 的累积量母函数 (*cumulant generating function*) 定义为

$$K_X(t) = \log M_X(t)$$

第 r 个累积量 (*cumulant*) k_r 是累积量母函数 $K_X(t)$ 展开式中 $\frac{t^r}{r!}$ 的系数, 因此

$$K_X(t) = k_1 t + k_2 \frac{t^2}{2!} + \dots + k_r \frac{t^r}{r!} + \dots = \sum_{j=1}^{\infty} k_j \frac{t^j}{j!}$$

从这个定义可以清楚地看出, 累积量母函数与累积量之间的关系与矩母函数与矩之间的关系相同。因此, 为了计算累积量, 我们可以比较系数或进行微分。

累积量可以用矩和中心矩表示。特别有用的是, 第一个累积量是均值, 第二个累积量是方差。为了证明这些结果, 我们将使用以下展开式 (对于 $|x| < 1$):

$$\log(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \dots + \frac{(-1)^{j+1}}{j}x^j + \dots$$

定理 2.27 (累积量与矩的关系) 如果 X 是一个具有矩 $\{\mu'_r\}$ 、中心矩 $\{\mu_r\}$ 和累积量 $\{\kappa_r\}$ 的随机变量，那么

- i. 第一个累积量是均值， $\kappa_1 = \mathbb{E}[X] = \mu'_1 = \mu$ 。
- ii. 第二个累积量是方差， $\kappa_2 = \text{Var}(X) = \mu_2 = \sigma^2$ 。
- iii. 第三个累积量是第三个中心矩， $\kappa_3 = \mu_3$ 。
- iv. 第四和第二个累积量的函数得到第四个中心矩， $\kappa_4 + 3\kappa_2^2 = \mu_4$ 。

2.8 Functions of random variables

2.8.1 Distribution and mass/density for $g(X)$

假设 X 是定义在 (Ω, \mathcal{F}, P) 上的随机变量，且 $g : \mathbb{R} \rightarrow \mathbb{R}$ 是一个性质良好的函数。我们希望推导 $Y = g(X)$ 的累积分布函数的表达式。

定义 2.28 (逆像) 如果 $g : \mathbb{R} \rightarrow \mathbb{R}$ 是一个函数，且 B 是实数的子集，则 B 在 g 下的逆像是那些在 g 下的像位于 B 中的实数集合，即对于所有 $B \subseteq \mathbb{R}$ ，我们定义 B 在 g 下的逆像为

$$g^{-1}(B) = \{x \in \mathbb{R} : g(x) \in B\}$$

例题 2.7 (连续随机变量的平方) 考虑一个连续随机变量 X ，并令 $Y = X^2$ 。我们可以直接推导 Y 的累积分布函数：

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

$$F_Y(y) = \begin{cases} F_X(\sqrt{y}) - F_X(-\sqrt{y}) & \text{当 } y \geq 0 \\ 0 & \text{其他} \end{cases}$$

对 y 求导得到密度函数：

$$f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}} [f_X(\sqrt{y}) + f_X(-\sqrt{y})] & \text{当 } y \geq 0 \\ 0 & \text{其他} \end{cases}$$

2.8.2 Monotone functions of random variables

如果我们假设所关注的函数是单调的，那么前一节的大部分讨论将得到简化。

定理 2.28 (随机变量的单调函数的分布) 如果 X 是一个随机变量， $g : \mathbb{R} \rightarrow \mathbb{R}$ 在 X 的支撑集上是严格单调函数，且 $Y = g(X)$ ，则 Y 的累积分布函数为

$$F_Y(y) = \begin{cases} F_X(g^{-1}(y)) & \text{若 } g \text{ 是递增的} \\ 1 - F_X(g^{-1}(y)-) & \text{若 } g \text{ 是递减的} \end{cases}$$

其中 y 在 g 的值域内。

注意，当 X 是连续型随机变量时，有 $F_X(g^{-1}(y)-) = F_X(g^{-1}(y))$ ；当 X 是离散型随机变量时，有 $F_X(g^{-1}(y)-) = F_X(g^{-1}(y)) - P(X = g^{-1}(y))$ 。

定理 2.29 (连续随机变量的单调函数的密度) 如果 X 是一个连续随机变量， $g : \mathbb{R} \rightarrow \mathbb{R}$ 是一个性质良好的单调函数，且 $Y = g(X)$ ，则 Y 的密度函数为

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| & \text{当 } y \text{ 在 } g \text{ 的值域内} \\ 0 & \text{其他} \end{cases}$$

定理 2.30 设 $U \sim U[0, 1]$ 。对于任何严格递增且连续的分布函数 F , 由 $X = F^{-1}(U)$ 定义的随机变量 X 具有分布函数 F 。

2.9 Sequences of random variables and convergence

定义 2.29 (实数序列的收敛) 设 $\{x_n\}$ 是一个实数序列, x 是一个实数。我们说 x_n 收敛于 x 当且仅当对于每个 $\varepsilon > 0$, 我们能找到一个整数 N , 使得对所有 $n > N$ 有 $|x_n - x| < \varepsilon$ 。在这些条件下, 我们记作 $x_n \rightarrow x$ 当 $n \rightarrow \infty$ 。

现在考虑一个随机变量序列 $\{X_n\}$ 和一个随机变量 X 。我们想知道 $\{X_n\}$ 收敛于 X 是什么意思。使用上述定义是不可能的; 因为 $|X_n - X|$ 是一个随机变量, 直接与实数 ε 比较没有意义。实际上, 对于随机变量, 存在许多不同形式的收敛。我们为随机变量序列定义四种不同的收敛模式。

定义 2.30 (收敛的类型) 设 $\{X_n\}$ 是一个随机变量序列, X 是一个随机变量。

i. 依分布收敛: 如果对于累积分布函数连续的所有 x , 有

$$P(X_n \leq x) \rightarrow P(X \leq x) \text{ 当 } n \rightarrow \infty$$

则 $\{X_n\}$ 依分布收敛于 X 。这记为

$$X_n \xrightarrow{d} X.$$

这也可以写成 $F_{X_n}(x) \rightarrow F_X(x)$ 。依分布收敛有时称为依法律收敛。

ii. 依概率收敛: 如果对于任何 $\varepsilon > 0$, 有

$$P(|X_n - X| < \varepsilon) \rightarrow 1 \text{ 当 } n \rightarrow \infty$$

则 $\{X_n\}$ 依概率收敛于 X 。另一种表达式是

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

这记为 $X_n \xrightarrow{p} X$ 。依概率收敛有时称为依测度收敛。

iii. 几乎必然收敛: 如果对于任何 $\varepsilon > 0$, 有

$$P\left(\lim_{n \rightarrow \infty} |X_n - X| < \varepsilon\right) = 1$$

则 $\{X_n\}$ 几乎必然收敛于 X 。这记为 $X_n \xrightarrow{a.s.} X$ 。几乎必然收敛有时称为以概率 1 收敛。

iv. 均方收敛: 如果

$$\mathbb{E}[(X_n - X)^2] \rightarrow 0 \text{ 当 } n \rightarrow \infty$$

则 $\{X_n\}$ 均方收敛于 X 。这记为 $X_n \xrightarrow{m.s.} X$ 。

$$\begin{array}{ccc} X_n \xrightarrow{a.s.} X & \Leftrightarrow & X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X. \\ & \nearrow \searrow & \\ X_n \xrightarrow{m.s.} X & \Leftrightarrow & \end{array}$$

图 1: 随机变量序列的收敛

定理 2.31 (收敛于退化分布) 假设 $\{X_n\}$ 是一个随机变量序列。如果 $X_n \xrightarrow{d} c$, 其中 c 是一个常数, 那么 $X_n \xrightarrow{p} c$ 。

例题 2.8 设 X_2, X_3, X_4, \dots 是一个随机变量序列, 使得

$$F_{X_n}(x) = \begin{cases} 1 - \left(1 - \frac{1}{n}\right)^{nx} & x > 0 \\ 0 & \text{其他} \end{cases}$$

证明 X_n 依分布收敛于 $Exp(1)$ 。

解:

设 $X \sim Exp(1)$ 。对于 $x \leq 0$, 我们有

$$F_{X_n}(x) = F_X(x) = 0, \quad \text{对于 } n = 2, 3, 4, \dots$$

对于 $x \geq 0$, 我们有

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = \lim_{n \rightarrow \infty} \left(1 - \left(1 - \frac{1}{n}\right)^{nx}\right) = 1 - \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^{nx} = 1 - e^{-x} = F_X(x), \quad \text{对所有 } x$$

因此, 我们得出结论 $X_n \xrightarrow{d} X$ 。

定理 2.32 考虑序列 X_1, X_2, X_3, \dots 和随机变量 X 。假设 X 和 X_n (对所有 n) 是非负且整数值的, 即

$$R_X \subset \{0, 1, 2, \dots\},$$

$$R_{X_n} \subset \{0, 1, 2, \dots\}, \quad \text{对于 } n = 1, 2, 3, \dots$$

那么 $X_n \xrightarrow{d} X$ 当且仅当

$$\lim_{n \rightarrow \infty} f_{X_n}(k) = f_X(k), \quad \text{对于 } k = 0, 1, 2, \dots$$

例题 2.9 设 X_1, X_2, X_3, \dots 是一个随机变量序列, 使得

$$X_n \sim Bin\left(n, \frac{\lambda}{n}\right), \quad \text{对于 } n \in \mathbb{N}, n > \lambda$$

其中 $\lambda > 0$ 是一个常数。证明 X_n 依分布收敛于 $Pois(\lambda)$ 。

解:

$$\lim_{n \rightarrow \infty} f_{X_n}(k) = \lim_{n \rightarrow \infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \lambda^k \lim_{n \rightarrow \infty} \frac{n!}{k!(n-k)!} \left(\frac{1}{n^k}\right) \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{e^{-\lambda} \lambda^k}{k!}$$

依分布收敛最著名的例子是中心极限定理 (CLT), 它指出独立同分布随机变量 X_1, X_2, X_3, \dots 的标准化平均值依分布收敛于标准正态随机变量。

例题 2.10 设 $X_n \sim Exp(n)$, 证明 $X_n \xrightarrow{p} 0$ 。

解:

$$\lim_{n \rightarrow \infty} P(|X_n - X| \geq \varepsilon) = \lim_{n \rightarrow \infty} P(X_n \geq \varepsilon) = \lim_{n \rightarrow \infty} e^{-n\varepsilon} = 0, \quad \text{对于所有 } \varepsilon > 0$$

依概率收敛最著名的例子是弱大数定律 (WLLN)，它指出如果 X_1, X_2, X_3, \dots 是独立同分布随机变量，均值 $\mu < \infty$ ，则定义为

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

的平均值序列依概率收敛于 μ 。

几乎必然收敛的一个重要例子是强大数定律 (SLLN)，它指出如果 X_1, X_2, X_3, \dots 是独立同分布随机变量，均值 $\mu < \infty$ ，则定义为

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

的平均值序列几乎必然收敛于 μ 。

定理 2.33 (连续映射定理) 设 X_1, X_2, X_3, \dots 是一个随机变量序列。再设 $h : \mathbb{R} \rightarrow \mathbb{R}$ 是一个连续函数。那么以下陈述成立：

1. 如果 $X_n \xrightarrow{d} X$ ，则 $h(X_n) \xrightarrow{d} h(X)$ 。
2. 如果 $X_n \xrightarrow{p} X$ ，则 $h(X_n) \xrightarrow{p} h(X)$ 。
3. 如果 $X_n \xrightarrow{a.s.} X$ ，则 $h(X_n) \xrightarrow{a.s.} h(X)$ 。

2.10 Further exercises

1. 随机变量 X 的概率密度函数为

$$f(x) = \begin{cases} 2e^{-2x} & x > 0, \\ 0 & \text{其他.} \end{cases}$$

(i) 证明 X 的分布函数为

$$F(x) = \begin{cases} 1 - e^{-2x} & x > 0, \\ 0 & \text{其他.} \end{cases}$$

(ii) 证明 $\mathbb{E}(X) = \frac{1}{2}$, 并求 X 的中位数。

随机变量 Y 定义为 $Y = X^2$ 。

(iii) 对 $y > 0$, 求 $P(Y < y)$, 并导出 Y 的概率密度函数。

(iv) 用 $\mathbb{E}(X)$ 和 $\text{Var}(X)$ 表示 $\mathbb{E}(Y)$, 并推导出 $\mathbb{E}(Y) = \frac{1}{2}$ 。

(v) 证明 Y 的中位数是 X 的中位数的平方。

解答:

(i) 对于 $x > 0$, 有

$$F(x) = \int_0^x f(t) dt = \int_0^x 2e^{-2t} dt = [-e^{-2t}]_0^x = -e^{-2x} + 1 = 1 - e^{-2x}.$$

对于 $x \leq 0$, $f(x) = 0$, 所以 $F(x) = 0$ 。因此,

$$F(x) = \begin{cases} 1 - e^{-2x} & x > 0, \\ 0 & \text{其他.} \end{cases}$$

(ii) 期望:

$$\mathbb{E}(X) = \int_0^\infty x \cdot 2e^{-2x} dx.$$

使用分部积分: 令 $u = x$, $dv = 2e^{-2x} dx$, 则 $du = dx$, $v = -e^{-2x}$ 。

$$\mathbb{E}(X) = [-xe^{-2x}]_0^\infty + \int_0^\infty e^{-2x} dx = 0 + \left[-\frac{1}{2}e^{-2x} \right]_0^\infty = 0 + \frac{1}{2} = \frac{1}{2}.$$

中位数: 设中位数为 m , 则 $F(m) = 0.5$ 。

$$1 - e^{-2m} = 0.5 \Rightarrow e^{-2m} = 0.5 \Rightarrow -2m = \ln(0.5) = -\ln 2 \Rightarrow m = \frac{\ln 2}{2}.$$

(iii) 对于 $y > 0$, 有

$$P(Y < y) = P(X^2 < y) = P(0 < X < \sqrt{y}) = F(\sqrt{y}) = 1 - e^{-2\sqrt{y}}.$$

因此, Y 的分布函数为

$$G(y) = \begin{cases} 1 - e^{-2\sqrt{y}} & y > 0, \\ 0 & \text{其他.} \end{cases}$$

概率密度函数为

$$g(y) = \frac{d}{dy} G(y) = \frac{d}{dy} (1 - e^{-2\sqrt{y}}) = e^{-2\sqrt{y}} \cdot \frac{1}{\sqrt{y}} = \frac{e^{-2\sqrt{y}}}{\sqrt{y}}, \quad y > 0.$$

(iv)

$$Y = X^2 \Rightarrow \mathbb{E}(Y) = \mathbb{E}(X^2) = \text{Var}(X) + [\mathbb{E}(X)]^2.$$

已知 $\mathbb{E}(X) = \frac{1}{2}$, 且

$$\mathbb{E}(X^2) = \int_0^\infty x^2 \cdot 2e^{-2x} dx.$$

使用公式 $\int_0^\infty x^n e^{-ax} dx = \frac{n!}{a^{n+1}}$, 得

$$\mathbb{E}(X^2) = 2 \cdot \frac{2!}{(2)^3} = 2 \cdot \frac{2}{8} = \frac{1}{2}.$$

因此,

$$\mathbb{E}(Y) = \mathbb{E}(X^2) = \frac{1}{2}.$$

(v) 设 m_Y 为 Y 的中位数, 则 $G(m_Y) = 0.5$ 。

$$1 - e^{-2\sqrt{m_Y}} = 0.5 \Rightarrow e^{-2\sqrt{m_Y}} = 0.5 \Rightarrow -2\sqrt{m_Y} = -\ln 2 \Rightarrow \sqrt{m_Y} = \frac{\ln 2}{2}.$$

而 X 的中位数 $m_X = \frac{\ln 2}{2}$, 所以

$$m_Y = \left(\frac{\ln 2}{2}\right)^2 = m_X^2.$$

2. 连续随机变量 X 的概率密度函数为

$$f(x) = \begin{cases} 1 & 1 \leq x \leq 2, \\ 0 & \text{otherwise,} \end{cases}$$

定义 $Y = \frac{2}{X}$ 。求 Y 的概率密度函数。

解答:

由 $Y = \frac{2}{X}$ 得 $X = \frac{2}{Y}$, 且 $\frac{dX}{dY} = -\frac{2}{Y^2}$ 。

当 $1 \leq X \leq 2$ 时, Y 的范围是 $1 \leq Y \leq 2$ 。

使用变换公式:

$$g(y) = f(x(y)) \left| \frac{dx}{dy} \right| = 1 \cdot \left| -\frac{2}{y^2} \right| = \frac{2}{y^2}, \quad 1 \leq y \leq 2.$$

因此,

$$g(y) = \begin{cases} \frac{2}{y^2} & 1 \leq y \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

3. 连续随机变量 T 的概率密度函数为

$$f(t) = \begin{cases} 0 & t < 2, \\ \frac{2}{(t-1)^3} & t \geq 2. \end{cases}$$

(i) 求 T 的分布函数, 并求 $P(T > 5)$ 。

(ii) 对 T 进行连续独立观测，直到第一次观测值大于 5 为止。随机变量 N 是到首次出现大于 5 的观测为止的总观测次数。求 $P(N > \mathbb{E}(N))$ 。

(iii) 求 $Y = \frac{1}{T-1}$ 的概率密度函数。

解答：

(i) 分布函数：

$$F(t) = \int_2^t \frac{2}{(s-1)^3} ds = \left[-\frac{1}{(s-1)^2} \right]_2^t = 1 - \frac{1}{(t-1)^2}, \quad t \geq 2.$$

所以

$$F(t) = \begin{cases} 0 & t < 2, \\ 1 - \frac{1}{(t-1)^2} & t \geq 2. \end{cases}$$

$$P(T > 5) = 1 - F(5) = 1 - \left(1 - \frac{1}{(5-1)^2} \right) = \frac{1}{16}.$$

(ii) 设 $p = P(T > 5) = \frac{1}{16}$ ，则 $N \sim \text{Geometric}(p)$ ， $\mathbb{E}(N) = \frac{1}{p} = 16$ 。

$$P(N > 16) = 1 - P(N \leq 16) = 1 - \sum_{k=1}^{16} (1-p)^{k-1} p = (1-p)^{16} = \left(\frac{15}{16} \right)^{16}.$$

(iii) $Y = \frac{1}{T-1}$ ，则 $T = 1 + \frac{1}{Y}$ ， $\frac{dT}{dY} = -\frac{1}{Y^2}$ 。

当 $T \geq 2$ 时， $Y \leq \frac{1}{2-1} = 1$ ，且 $Y > 0$ 。

使用变换公式：

$$g(y) = f(t(y)) \left| \frac{dt}{dy} \right| = \frac{2}{(1/y)^3} \cdot \frac{1}{y^2} = 2y^3 \cdot \frac{1}{y^2} = 2y, \quad 0 < y \leq 1.$$

所以

$$g(y) = \begin{cases} 2y & 0 < y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

4. 连续随机变量 X 的概率密度函数为

$$f(x) = \begin{cases} 1+x & -1 \leq x \leq 0, \\ 1-x & 0 < x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

证明其分布函数为

$$F(x) = \begin{cases} 0 & x < -1, \\ \frac{1}{2}(1+x)^2 & -1 \leq x \leq 0, \\ 1 - \frac{1}{2}(1-x)^2 & 0 < x \leq 1, \\ 1 & x > 1. \end{cases}$$

随机变量 $Y = X^2$ 。证明

$$P(Y \leq y) = 2\sqrt{y} - y, \quad 0 \leq y \leq 1.$$

求 $\mathbb{E}(Y)$ 。

解答：

当 $-1 \leq x \leq 0$:

$$F(x) = \int_{-1}^x (1+t) dt = \left[t + \frac{t^2}{2} \right]_{-1}^x = x + \frac{x^2}{2} - (-1 + \frac{1}{2}) = x + \frac{x^2}{2} + \frac{1}{2} = \frac{1}{2}(1+x)^2.$$

当 $0 < x \leq 1$:

$$F(x) = F(0) + \int_0^x (1-t) dt = \frac{1}{2} + \left[t - \frac{t^2}{2} \right]_0^x = \frac{1}{2} + x - \frac{x^2}{2} = 1 - \frac{1}{2}(1-x)^2.$$

对于 $Y = X^2$:

$$P(Y \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = F(\sqrt{y}) - F(-\sqrt{y}).$$

当 $0 \leq y \leq 1$:

$$F(\sqrt{y}) = 1 - \frac{1}{2}(1-\sqrt{y})^2, \quad F(-\sqrt{y}) = \frac{1}{2}(1-\sqrt{y})^2.$$

所以

$$P(Y \leq y) = 1 - \frac{1}{2}(1-\sqrt{y})^2 - \frac{1}{2}(1-\sqrt{y})^2 = 1 - (1-\sqrt{y})^2 = 2\sqrt{y} - y.$$

期望:

$$\begin{aligned} \mathbb{E}(Y) &= \int_0^1 y \cdot \frac{d}{dy} (2\sqrt{y} - y) dy = \int_0^1 y \left(\frac{1}{\sqrt{y}} - 1 \right) dy = \int_0^1 (\sqrt{y} - y) dy \\ &= \left[\frac{2}{3}y^{3/2} - \frac{y^2}{2} \right]_0^1 = \frac{2}{3} - \frac{1}{2} = \frac{1}{6}. \end{aligned}$$

5. 连续随机变量 X 的概率密度函数为

$$f(x) = \begin{cases} \frac{1}{6}x & 2 \leq x \leq 4, \\ 0 & \text{otherwise.} \end{cases}$$

随机变量 $Y = X^3$ 。证明 Y 的概率密度函数为

$$g(y) = \begin{cases} \frac{1}{18}y^{-\frac{1}{3}} & 8 \leq y \leq 64, \\ 0 & \text{otherwise.} \end{cases}$$

求 $\mathbb{E}(Y)$ 。

解答:

由 $Y = X^3$ 得 $X = Y^{1/3}$, $\frac{dx}{dy} = \frac{1}{3}Y^{-2/3}$ 。

当 $2 \leq X \leq 4$ 时, $8 \leq Y \leq 64$ 。

使用变换公式:

$$g(y) = f(x(y)) \left| \frac{dx}{dy} \right| = \frac{1}{6}y^{1/3} \cdot \frac{1}{3}y^{-2/3} = \frac{1}{18}y^{-1/3}, \quad 8 \leq y \leq 64.$$

期望:

$$\begin{aligned} \mathbb{E}(Y) &= \int_8^{64} y \cdot \frac{1}{18}y^{-1/3} dy = \frac{1}{18} \int_8^{64} y^{2/3} dy \\ &= \frac{1}{18} \left[\frac{3}{5}y^{5/3} \right]_8^{64} = \frac{1}{30} (64^{5/3} - 8^{5/3}) = \frac{1}{30} ((4^3)^{5/3} - (2^3)^{5/3}) \\ &= \frac{1}{30}(4^5 - 2^5) = \frac{1}{30}(1024 - 32) = \frac{992}{30} = \frac{496}{15}. \end{aligned}$$

6. 设随机变量 X 的期望 $\mathbb{E}(X) = 3$, 且 X 为非负随机变量。用马尔可夫不等式估计 $P(X \geq 10)$ 的上界。

解答: 由马尔可夫不等式:

$$P(X \geq 10) \leq \frac{\mathbb{E}(X)}{10} = \frac{3}{10} = 0.3.$$

7. 设随机变量 X 的方差 $\text{Var}(X) = 9$, 期望 $\mathbb{E}(X) = 5$ 。用切比雪夫不等式估计 $P(|X - 5| \geq 6)$ 的上界。

解答: 由切比雪夫不等式:

$$P(|X - \mathbb{E}(X)| \geq \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2}.$$

取 $\varepsilon = 6$:

$$P(|X - 5| \geq 6) \leq \frac{9}{36} = 0.25.$$

8. 设 X_1, X_2, \dots, X_n 独立同分布, $\mathbb{E}(X_i) = \mu$, $\text{Var}(X_i) = \sigma^2$ 。记 $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ 。用切比雪夫不等式证明:

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

并说明当 $n \rightarrow \infty$ 时该概率的极限。

解答: 由已知:

$$\mathbb{E}(\bar{X}_n) = \mu, \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

由切比雪夫不等式:

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

当 $n \rightarrow \infty$ 时, 右边趋于 0, 即 $P(|\bar{X}_n - \mu| \geq \varepsilon) \rightarrow 0$, 这是大数定律的一种形式。

9. 设随机变量 X 服从参数为 $\lambda = 2$ 的指数分布。用马尔可夫不等式估计 $P(X \geq 3)$ 的上界, 并与真实概率比较。

解答: 指数分布 $\mathbb{E}(X) = \frac{1}{\lambda} = 0.5$ (注意: 若参数为 λ , 则期望为 $1/\lambda$, 但题给 $\lambda = 2$, 所以 $\mathbb{E}(X) = 0.5$)。

马尔可夫不等式:

$$P(X \geq 3) \leq \frac{0.5}{3} \approx 0.1667.$$

真实概率:

$$P(X \geq 3) = e^{-2 \cdot 3} = e^{-6} \approx 0.00248.$$

可见马尔可夫不等式给出的上界较宽松。

10. 设 X 为非负随机变量, 且 $\mathbb{E}(X^2) = 20$ 。用马尔可夫不等式估计 $P(X \geq 6)$ 的上界。

解答: 由马尔可夫不等式:

$$P(X \geq 6) = P(X^2 \geq 36) \leq \frac{\mathbb{E}(X^2)}{36} = \frac{20}{36} = \frac{5}{9} \approx 0.5556.$$

11. 设随机变量 X 的期望 $\mathbb{E}(X) = 2$, 方差 $\text{Var}(X) = 0.5$ 。用切比雪夫不等式确定最小的常数 c , 使得 $P(|X - 2| \geq c) \leq 0.1$ 。

解答: 由切比雪夫不等式:

$$P(|X - 2| \geq c) \leq \frac{0.5}{c^2}.$$

令 $\frac{0.5}{c^2} \leq 0.1$, 得 $c^2 \geq 5$, 即 $c \geq \sqrt{5} \approx 2.236$ 。所以最小的 c 为 $\sqrt{5}$ 。

12. 设 X 为随机变量, $\mathbb{E}(X) = 0$, $\mathbb{E}(X^2) = \sigma^2$ 。对任意 $a > 0$, 证明:

$$P(X \geq a) \leq \frac{\sigma^2}{\sigma^2 + a^2}.$$

(提示: 考虑 $Y = X + c$ 的马尔可夫不等式, 并优化 c)

解答: 对任意 $c > 0$, 由马尔可夫不等式:

$$P(X \geq a) = P(X + c \geq a + c) \leq \frac{\mathbb{E}[(X + c)^2]}{(a + c)^2} = \frac{\sigma^2 + c^2}{(a + c)^2}.$$

令 $f(c) = \frac{\sigma^2 + c^2}{(a + c)^2}$, 求导得最优 $c = \frac{\sigma^2}{a}$, 代入得:

$$P(X \geq a) \leq \frac{\sigma^2 + (\sigma^4/a^2)}{(a + \sigma^2/a)^2} = \frac{\sigma^2(a^2 + \sigma^2)/a^2}{(a^2 + \sigma^2)^2/a^2} = \frac{\sigma^2}{\sigma^2 + a^2}.$$

13. 设离散随机变量 X 的概率分布为:

$$P(X = k) = \frac{1}{2^{k+1}}, \quad k = 0, 1, 2, \dots$$

求 X 的概率母函数 $G_X(s)$, 并利用它求 $\mathbb{E}(X)$ 和 $\text{Var}(X)$ 。

解答: 概率母函数:

$$G_X(s) = \mathbb{E}(s^X) = \sum_{k=0}^{\infty} s^k \cdot \frac{1}{2^{k+1}} = \frac{1}{2} \sum_{k=0}^{\infty} \left(\frac{s}{2}\right)^k = \frac{1}{2} \cdot \frac{1}{1 - s/2} = \frac{1}{2 - s}, \quad |s| < 2.$$

一阶导数:

$$G'_X(s) = \frac{1}{(2 - s)^2}, \quad G'_X(1) = 1.$$

二阶导数:

$$G''_X(s) = \frac{2}{(2 - s)^3}, \quad G''_X(1) = 2.$$

所以:

$$\mathbb{E}(X) = G'_X(1) = 1,$$

$$\mathbb{E}(X(X - 1)) = G''_X(1) = 2,$$

$$\text{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = [\mathbb{E}(X(X - 1)) + \mathbb{E}(X)] - 1 = (2 + 1) - 1 = 2.$$

14. 设随机变量 X 的矩母函数为:

$$M_X(t) = \frac{1}{(1 - 2t)^3}, \quad t < \frac{1}{2}.$$

求 $\mathbb{E}(X)$ 和 $\text{Var}(X)$ 。

解答: 取对数:

$$\ln M_X(t) = -3 \ln(1 - 2t).$$

一阶导数:

$$\frac{M'_X(t)}{M_X(t)} = \frac{6}{1 - 2t}, \quad \mathbb{E}(X) = M'_X(0) = 6.$$

二阶导数:

$$\frac{M''_X(t)M_X(t) - [M'_X(t)]^2}{[M_X(t)]^2} = \frac{12}{(1 - 2t)^2},$$

在 $t = 0$ 处:

$$\mathbb{E}(X^2) - 36 = 12, \quad \mathbb{E}(X^2) = 48.$$

所以:

$$\text{Var}(X) = 48 - 36 = 12.$$

15. 设 X 和 Y 独立, $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\mu)$ 。求 $Z = X + Y$ 的矩母函数, 并证明 $Z \sim \text{Poisson}(\lambda + \mu)$ 。

解答: 泊松分布的矩母函数:

$$M_X(t) = e^{\lambda(e^t-1)}, \quad M_Y(t) = e^{\mu(e^t-1)}.$$

由独立性:

$$M_Z(t) = M_X(t)M_Y(t) = e^{\lambda(e^t-1)} \cdot e^{\mu(e^t-1)} = e^{(\lambda+\mu)(e^t-1)},$$

这是参数为 $\lambda + \mu$ 的泊松分布的矩母函数, 故 $Z \sim \text{Poisson}(\lambda + \mu)$ 。

16. 设随机变量 X 的概率母函数为:

$$G_X(s) = \frac{1}{3} + \frac{1}{3}s + \frac{1}{3}s^2.$$

求 X 的分布列, 并求 $Y = 2X + 1$ 的概率母函数。

解答: 由概率母函数形式可得:

$$P(X=0) = \frac{1}{3}, \quad P(X=1) = \frac{1}{3}, \quad P(X=2) = \frac{1}{3}.$$

$Y = 2X + 1$ 的概率母函数:

$$G_Y(s) = \mathbb{E}(s^{2X+1}) = s\mathbb{E}((s^2)^X) = sG_X(s^2) = s\left(\frac{1}{3} + \frac{1}{3}s^2 + \frac{1}{3}s^4\right).$$

17. 设 $X \sim N(\mu, \sigma^2)$, 求 $\mathbb{E}[(X - \mu)^3]$ 和 $\mathbb{E}[(X - \mu)^4]$ 。

解答: 令 $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$ 。

对于标准正态分布, 奇数阶中心矩为 0:

$$\mathbb{E}[(X - \mu)^3] = \sigma^3 \mathbb{E}[Z^3] = 0.$$

对于四阶矩:

$$\mathbb{E}[Z^4] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^4 e^{-z^2/2} dz.$$

利用分部积分或已知结果 $\mathbb{E}[Z^4] = 3$:

$$\mathbb{E}[(X - \mu)^4] = \sigma^4 \mathbb{E}[Z^4] = 3\sigma^4.$$

18. 已知随机变量 X 的期望 $\mathbb{E}(X) = 3$, 方差 $\text{Var}(X) = 4$ 。求:

- (a) $\mathbb{E}(2X + 5)$
- (b) $\text{Var}(2X + 5)$
- (c) $\mathbb{E}(X^2)$

解答:

- (a) $\mathbb{E}(2X + 5) = 2\mathbb{E}(X) + 5 = 2 \times 3 + 5 = 11$
- (b) $\text{Var}(2X + 5) = 2^2 \text{Var}(X) = 4 \times 4 = 16$
- (c) $\mathbb{E}(X^2) = \text{Var}(X) + [\mathbb{E}(X)]^2 = 4 + 3^2 = 13$

2.11 Appendix: Proofs

3 Multivariate distributions 多变量分布

我们已经讨论了单变量的概率分布，但显然在处理现实问题的时候我们更多地还是会面对双变量以及多变量的情况，想知道 $X_1 \leq x_1$ 并且 $X_2 \leq x_2$ 时的概率或者 $X_1 \leq X_2$ 的概率，这就需要我们构建多个变量的联立分布。当然，考虑我们有多个变量的时候显然有两种情况，第一种是一个比较理想的特殊情况，就是虽然你有多个变量，但每个变量之间都是相互独立的，这种情况你算联立分布就特别省事，因为我们都知道一串独立事件都发生的概率就是它们每个事件单独发生的概率的乘积。第二种就是可能更常见的情况，变量之间不是独立的，是有相互联系的，这时光靠单变量分布就没法算了，我们就需要多变量模型来捕捉这些内在联系。初次讨论超过一个随机变量的概率分布，本章会介绍联立分布，边际分布以及独立随机变量的特殊性质。

3.1 Joint and marginal distributions

一组随机变量的累积分布函数称为联合累积分布函数。这是一个多元函数。

定义 3.1 (一般联合累积分布函数) 如果 X_1, \dots, X_n 是随机变量，则联合累积分布函数是一个函数 $F_{X_1, \dots, X_n} : \mathbb{R}^n \rightarrow [0, 1]$ ，定义为

$$F_{X_1, \dots, X_n}(x_1, x_2, \dots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n)$$

与多元分布相关的大部分概念通过查看二维情况（即二元分布）即可完全解释。到 n 维的推广在代数上通常是显而易见的，尽管 n 维分布的可视化要困难得多。

定义 3.2 (二元联合累积分布函数) 对于两个随机变量 X 和 Y ，联合累积分布函数是一个函数 $F_{X,Y} : \mathbb{R}^2 \rightarrow [0, 1]$ ，定义为

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$$

定理 3.1 (联合累积分布函数的基本性质) 假设 X 和 Y 是随机变量。如果 $F_{X,Y}$ 是 X 和 Y 的联合累积分布函数，则 $F_{X,Y}$ 具有以下性质：

- i. $F_{X,Y}(-\infty, y) = \lim_{x \rightarrow -\infty} F_{X,Y}(x, y) = 0$
 $F_{X,Y}(x, -\infty) = \lim_{y \rightarrow -\infty} F_{X,Y}(x, y) = 0$
 $F_{X,Y}(\infty, \infty) = \lim_{x,y \rightarrow \infty} F_{X,Y}(x, y) = 1$
- ii. 关于 x 右连续： $\lim_{h \downarrow 0} F_{X,Y}(x + h, y) = F_{X,Y}(x, y)$
关于 y 右连续： $\lim_{h \downarrow 0} F_{X,Y}(x, y + h) = F_{X,Y}(x, y)$

引理 3.1 如果 X 和 Y 是具有联合累积分布函数 $F_{X,Y}$ 的随机变量，则

$$P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = [F_{X,Y}(x_2, y_2) - F_{X,Y}(x_1, y_2)] - [F_{X,Y}(x_2, y_1) - F_{X,Y}(x_1, y_1)]$$

正如术语“联合”指的是多个变量一样，我们使用术语“边缘”来指代单个随机变量的分布。因此，边缘累积分布函数就是单个随机变量的通常累积分布函数。从联合分布生成边缘累积分布函数是很直接的。

定理 3.2 (从联合分布得到边缘分布) 如果 $F_{X,Y}$ 是 X 和 Y 的联合分布函数，则边缘累积分布函数由下式给出：

$$F_X(x) = \lim_{y \rightarrow \infty} F_{X,Y}(x, y) = F_{X,Y}(x, \infty)$$

$$F_Y(y) = \lim_{x \rightarrow \infty} F_{X,Y}(x, y) = F_{X,Y}(\infty, y)$$

注意，虽然我们总是可以从联合分布生成边缘分布，但反过来却并不成立。联合分布包含了边缘分布未捕获的信息。特别是，联合分布告诉我们所考虑随机变量之间关联的性质；这种关联通常被称为相依性（dependence）。

联合分布 \Rightarrow 边缘分布

联合分布 \Leftarrow 仅对独立随机变量成立 \Leftarrow 边缘分布

边缘分布 + “相依性”信息 = 联合分布

3.2 Joint mass and joint density

3.2.1 Mass for discrete distributions

定义 3.3 (联合质量函数) 假设 X 和 Y 是离散随机变量。 X 和 Y 的联合质量函数 (*joint mass function*) 是函数 $f_{X,Y} : \mathbb{R}^2 \rightarrow [0, 1]$ ，定义为

$$f_{X,Y}(x, y) = P(X = x, Y = y)$$

定理 3.3 对于具有联合质量函数 $f_{X,Y}$ 的离散随机变量 X 和 Y ，以及实数 $x_1 < x_2$ 和 $y_1 < y_2$ ，我们有

$$P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = \sum_{x_1 < x \leq x_2} \sum_{y_1 < y \leq y_2} f_{X,Y}(x, y)$$

该定理的一个简单推论是

$$P(X = x, y_1 < Y \leq y_2) = \sum_{y_1 < y \leq y_2} f_{X,Y}(x, y)$$

通过相加互斥结果来从考虑中移除一个变量的原则可以普遍应用。

定理 3.4 (从联合质量得到边缘质量) 对于具有联合质量函数 $f_{X,Y}$ 的离散随机变量 X 和 Y ，边缘质量函数由下式给出：

$$f_X(x) = \sum_y f_{X,Y}(x, y)$$

$$f_Y(y) = \sum_x f_{X,Y}(x, y)$$

表 5: 联合质量函数与边缘质量函数示例

$f_{X,Y}(x, y)$	$x = 0$	$x = 1$	$x = 2$	$f_Y(y)$
$y = 0$	0.71342	0.13273	0.00452	0.85067
$y = 1$	0.13273	0.01207	0.00000	0.14480
$y = 2$	0.00452	0.00000	0.00000	0.00452
$f_X(x)$	0.85067	0.14480	0.00452	1.00000

3.2.2 Density for continuous distributions

定义 3.4 (联合密度函数) 对于具有联合累积分布函数 $F_{X,Y}$ 的联合连续随机变量 X 和 Y , 联合密度函数 (*joint density function*) 是一个函数 $f_{X,Y} : \mathbb{R}^2 \rightarrow [0, \infty)$, 使得

$$F_{X,Y}(x, y) = \int_{-\infty}^y \int_{-\infty}^x f_{X,Y}(u, v) du dv \quad \text{对所有 } x, y \in \mathbb{R}$$

定理 3.5 (从联合分布得到联合密度) 对于具有联合累积分布函数 $F_{X,Y}$ 的联合连续随机变量 X 和 Y , 联合密度函数由下式给出:

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial u \partial v} F_{X,Y}(u, v) \Big|_{u=x, v=y}$$

定理 3.6 如果 $f_{X,Y}$ 是一个联合密度函数, 则 $f_{X,Y}$ 是一个正的实值函数, 且满足性质

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy = 1$$

定理 3.7 (矩形区域的概率) 如果 X 和 Y 是具有联合密度函数 $f_{X,Y}(x, y)$ 的随机变量, 则

$$P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f_{X,Y}(x, y) dx dy$$

定理 3.8 (从联合密度得到边缘密度) 如果 X 和 Y 是联合连续随机变量, 则

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx$$

从纯数学角度来看, 2 个随机变量的联合概率分布就是一个三变量的函数, 因为概率也是一个维度, 所以 2 个随机变量的联合概率分布是三维的。对于连续随机变量来说, 联合概率密度对 x 和 y 微分两次, 仍然是三维, x 的边际概率密度就是抛掉 y 的信息, 所以降维, 变成 2 维, 也就是回归单变量的情况。更一般的, n 个随机变量的联合概率分布在 $n+1$ 维的空间里。

例题 3.1 (两个多项式联合密度函数)

1. (具有矩形支撑的多项式密度) 考虑函数

$$f_{X,Y}(x, y) = \begin{cases} x + y & \text{当 } 0 < x < 1 \text{ 且 } 0 < y < 1 \\ 0 & \text{其他} \end{cases}$$

(a) 证明 $f_{X,Y}$ 是一个有效的密度函数。

解: 在支撑区域 $0 < x < 1, 0 < y < 1$ 上, $f_{X,Y}(x, y) = x + y \geq 0$ 。验证归一化:

$$\int_0^1 \int_0^1 (x + y) dx dy = \int_0^1 \left[\frac{1}{2}x^2 + xy \right]_{x=0}^{x=1} dy = \int_0^1 \left(\frac{1}{2} + y \right) dy = \left[\frac{1}{2}y + \frac{1}{2}y^2 \right]_0^1 = 1$$

因此 $f_{X,Y}$ 是有效的密度函数。

(b) 推导 X 和 Y 的联合累积分布函数。

解: 对 $0 \leq x \leq 1, 0 \leq y \leq 1$:

$$F_{X,Y}(x, y) = \int_0^y \int_0^x (u+v) du dv = \int_0^y \left[\frac{1}{2}u^2 + uv \right]_{u=0}^{u=x} dv = \int_0^y \left(\frac{1}{2}x^2 + xv \right) dv = \left[\frac{1}{2}x^2v + \frac{1}{2}xv^2 \right]_0^y$$

$$= \frac{1}{2}x^2y + \frac{1}{2}xy^2$$

因此：

$$F_{X,Y}(x,y) = \begin{cases} 0 & x \leq 0 \text{ 或 } y \leq 0 \\ \frac{1}{2}x^2y + \frac{1}{2}xy^2 & 0 < x < 1, 0 < y < 1 \\ \frac{1}{2}y + \frac{1}{2}y^2 & x \geq 1, 0 < y < 1 \\ \frac{1}{2}x^2 + \frac{1}{2}x & 0 < x < 1, y \geq 1 \\ 1 & x \geq 1, y \geq 1 \end{cases}$$

(c) 计算 $P(0.2 < X < 0.5, 0 < Y < 0.3)$ 。

解：

$$\begin{aligned} P &= \int_0^{0.3} \int_{0.2}^{0.5} (x+y) dx dy = \int_0^{0.3} \left[\frac{1}{2}x^2 + xy \right]_{x=0.2}^{x=0.5} dy = \int_0^{0.3} \left(\frac{0.25 - 0.02}{2} + (0.5 - 0.2)y \right) dy \\ &= \int_0^{0.3} (0.115 + 0.3y) dy = [0.115y + 0.15y^2]_0^{0.3} = 0.048 \end{aligned}$$

(d) 求 X 的边缘密度。

解：

$$f_X(x) = \int_0^1 (x+y) dy = \left[xy + \frac{1}{2}y^2 \right]_0^1 = x + \frac{1}{2}, \quad 0 < x < 1$$

(e) 计算 $P(2X < Y)$ 。

解：在区域 $0 < x < 1, 0 < y < 1$ 中， $2x < y$ 意味着 $y > 2x$ ，但 $y \leq 1$ ，所以 $x < 0.5$ ：

$$\begin{aligned} P &= \int_0^{0.5} \int_{2x}^1 (x+y) dy dx = \int_0^{0.5} \left[xy + \frac{1}{2}y^2 \right]_{y=2x}^{y=1} dx = \int_0^{0.5} \left(x + \frac{1}{2} - 2x^2 - 2x^2 \right) dx \\ &= \int_0^{0.5} \left(x + \frac{1}{2} - 4x^2 \right) dx = \left[\frac{1}{2}x^2 + \frac{1}{2}x - \frac{4}{3}x^3 \right]_0^{0.5} = \frac{5}{24} \end{aligned}$$

2. (具有三角支撑的多项式密度) 考虑函数

$$f_{X,Y}(x,y) = \begin{cases} 8xy & \text{当 } 0 < x < y < 1 \\ 0 & \text{其他} \end{cases}$$

(a) 证明 $f_{X,Y}$ 是一个有效的密度函数。

解：在支撑区域 $0 < x < y < 1$ 上， $f_{X,Y}(x,y) = 8xy \geq 0$ 。验证归一化：

$$\int_0^1 \int_0^y 8xy dx dy = \int_0^1 8y \left[\frac{1}{2}x^2 \right]_0^y dy = \int_0^1 4y^3 dy = [y^4]_0^1 = 1$$

因此 $f_{X,Y}$ 是有效的密度函数。

(b) 求 X 的边缘密度。

解：

$$f_X(x) = \int_x^1 8xy dy = 8x \left[\frac{1}{2}y^2 \right]_x^1 = 4x(1-x^2), \quad 0 < x < 1$$

(c) 计算 $P(Y < 2X)$ 。

解：在支撑区域 $0 < x < y < 1$ 中， $y < 2x$ 意味着 $x < y < \min(2x, 1)$ 。由于 $y < 1$ ，考虑 $x < 0.5$ 和 $x \geq 0.5$ 的情况：

$$\begin{aligned} P &= \int_0^{0.5} \int_x^{2x} 8xy \, dy \, dx + \int_{0.5}^1 \int_x^1 8xy \, dy \, dx \\ &= \int_0^{0.5} 8x \left[\frac{1}{2}y^2 \right]_x^{2x} \, dx + \int_{0.5}^1 8x \left[\frac{1}{2}y^2 \right]_x^1 \, dx \\ &= \int_0^{0.5} 4x(4x^2 - x^2) \, dx + \int_{0.5}^1 4x(1 - x^2) \, dx \\ &= \int_0^{0.5} 12x^3 \, dx + \int_{0.5}^1 (4x - 4x^3) \, dx \\ &= [3x^4]_0^{0.5} + [2x^2 - x^4]_{0.5}^1 = \frac{3}{16} + \left(1 - \frac{7}{16} \right) = \frac{3}{4} \end{aligned}$$

3.3 Expectation and joint moments

3.3.1 Expectation of a function of several variables

定理 3.9 (两个随机变量函数的期望) 如果 g 是一个性质良好的二元实值函数， $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ ，且 X 和 Y 是具有联合质量/密度函数 $f_{X,Y}$ 的随机变量，则

$$\mathbb{E}[g(X, Y)] = \begin{cases} \sum_y \sum_x g(x, y) f_{X,Y}(x, y) & (\text{离散情形}) \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) \, dx \, dy & (\text{连续情形}) \end{cases}$$

定理 3.10 (随机变量和的期望值) 如果 X 和 Y 是随机变量，则 $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ 。

3.3.2 Covariance and correlation

对于一对随机变量，常用于衡量（线性）关联程度的量是相关系数。相关系数定义的起点是协方差的概念。

定义 3.5 (协方差) 对于两个随机变量 X 和 Y ，我们定义 X 和 Y 之间的**协方差 (covariance)** 为

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

定理 3.11 (协方差的性质)

- 如果 a, b, c 是常数，

$$\text{Cov}(X, c) = 0$$

$$\text{Cov}(X + c, Y) = \text{Cov}(X, Y)$$

$$\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$$

- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$

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

- $\text{Cov}(X + Z, Y) = \text{Cov}(X, Y) + \text{Cov}(Z, Y)$

- $\text{Cov}(U + V, X + Y) = \text{Cov}(U, X) + \text{Cov}(U, Y) + \text{Cov}(V, X) + \text{Cov}(V, Y)$

- $\text{Cov}(X, X) = \text{Var}(X)$

- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$
- 如果 X 和 Y 独立, 则 $\text{Cov}(X, Y) = 0$

定义 3.6 (相关系数) 对于方差满足 $\text{Var}(X) > 0$ 且 $\text{Var}(Y) > 0$ 的随机变量 X 和 Y , 其**相关系数 (correlation)** 为

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

相关系数衡量两个变量之间线性关联的程度。

定理 3.12 (相关系数的取值范围)

$$|\text{Corr}(X, Y)| \leq 1$$

此外, 我们有

$$|\text{Corr}(X, Y)| = 1 \quad \text{当且仅当存在常数 } \alpha \text{ 和 } \beta \neq 0 \text{ 使得 } Y = \alpha + \beta X \text{ 且满足}$$

$$\text{Corr}(X, Y) = 1 \quad \text{若 } \beta > 0$$

$$\text{Corr}(X, Y) = -1 \quad \text{若 } \beta < 0$$

3.3.3 Joint moments

联合矩提供了关于两个随机变量之间相依结构的信息。对于大多数实际目的, 我们只考虑低阶的联合矩。

定义 3.7 (联合矩与联合中心矩) 如果 X 和 Y 是随机变量, 则 X 和 Y 的 (r, s) 阶联合矩 (*joint moment*) 为

$$\mu'_{r,s} = \mathbb{E}[X^r Y^s]$$

X 和 Y 的 (r, s) 阶联合中心矩 (*joint central moment*) 为

$$\mu_{r,s} = \mathbb{E}[(X - \mathbb{E}[X])^r (Y - \mathbb{E}[Y])^s]$$

3.3.4 Joint moment generating function

定义 3.8 (联合矩母函数) 对于随机变量 X 和 Y , 联合矩母函数定义为

$$\begin{aligned} M_{X,Y}(t, u) &= \mathbb{E}[e^{tX+uY}] \\ M_{X,Y}(t, u) &= \mathbb{E}\left(\sum_{i=0}^{\infty} \frac{(tX)^i}{i!} \sum_{j=0}^{\infty} \frac{(uY)^j}{j!}\right) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \mathbb{E}(X^i Y^j) \frac{t^i u^j}{i! j!} \end{aligned}$$

(r, s) 阶联合矩是联合矩母函数多项式展开中 $(t^r u^s)/(r! s!)$ 的系数。一个推论是联合矩可以通过微分来计算:

$$M_{X,Y}^{(r,s)}(0, 0) = \left. \frac{\partial^{r+s}}{\partial t^r \partial u^s} M_{X,Y}(t, u) \right|_{t=0, u=0} = \mathbb{E}[X^r Y^s] = \mu'_{r,s}$$

定义 3.9 (联合累积量母函数与联合累积量) 对于随机变量 X 和 Y , 联合累积量母函数为

$$K_{X,Y}(t, u) = \log M_{X,Y}(t, u)$$

我们定义 (r, s) 阶联合累积量 $\kappa_{r,s}$ 为 $K_{X,Y}(t, u)$ 的展开式中 $(t^r u^s)/(r! s!)$ 的系数。

联合累积量母函数可以写成 t 和 u 的多项式：

$$K_{X,Y}(t, u) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \kappa_{ij} \frac{t^i u^j}{i! j!}$$

累积量可以通过比较系数或在零点处求导来计算。

定理 3.13 (用累积量表示的相关系数)

$$\text{Corr}(X, Y) = \frac{\kappa_{1,1}}{\sqrt{\kappa_{2,0}\kappa_{0,2}}}$$

3.4 Independent random variables

3.4.1 Independent for pairs of random variables

定义 3.10 (独立随机变量) 随机变量 X 和 Y 是独立 (*independent*) 的当且仅当对所有的 x 和 y , 事件 $\{X \leq x\}$ 和 $\{Y \leq y\}$ 是独立的。

这个定义的一个直接推论是, 对于独立随机变量, 可以从边缘分布生成联合分布。

定理 3.14 (独立随机变量的联合分布) 随机变量 X 和 Y 是独立的当且仅当 X 和 Y 的联合累积分布函数是边缘累积分布函数的乘积, 即当且仅当

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) \quad \text{对所有 } x, y \in \mathbb{R}$$

该定理可以用质量函数或密度函数重新表述。

定理 3.15 (独立随机变量的质量/密度函数) 随机变量 X 和 Y 是独立的当且仅当它们的联合质量/密度函数是边缘质量/密度函数的乘积, 即当且仅当

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \quad \text{对所有 } x, y \in \mathbb{R}$$

引理 3.2 (独立随机变量乘积的期望) 如果 X 和 Y 是独立随机变量, 则

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

表述这个引理的另一种方式是：

$$X \text{ 和 } Y \text{ 独立 } (\text{independent}) \implies X \text{ 和 } Y \text{ 不相关 } (\text{uncorrelated})$$

注意, X, Y 独立 $\implies \mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) \iff \text{Cov}(X, Y) = 0 \iff \text{Corr}(X, Y) = 0$ (假设 X, Y 期望存在, 方差非零)。我们做如下几点补充说明:

- 注意后面三个等式互为充要条件, 是同一事实的三种等价表述, 这个事实可以描述为: “ X 与 Y 线性独立 (linearly independent)”, “ X 与 Y 不 (线性) 相关 (uncorrelated)”, “ X 与 Y 之间不存在线性关系” 或 “ X 与 Y 的最佳线性拟合的斜率为 0”。(这里的“最佳线性拟合”是指最小二乘意义下的总体回归)
- 注意 X 和 Y 之间有无确定的非线性关系不影响它们是否可能线性相关。考虑 $Y = X^2$ 的最佳线性拟合线的斜率, 它们是线性不相关的 (uncorrelated); 考虑 $Y = X^3$ 的非线性关系, 它们是线性相关的 (correlated)。如果 X 和 Y 的非线性关系是偶函数形式, 并且 X 的分布关于 0 对称, 那么它们的最佳线性拟合斜率一定是 0。

- X 和 Y 独立 (independent) 则是一个更强的条件, 要求 X 和 Y 在线性或是非线性层面都不具有任何相关性。之后我们会了解到, 若 (X, Y) 服从二元正态分布, 则 X, Y 独立与 $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ 是充要条件。

引理 3.3 (独立随机变量函数的独立性) 如果 X 和 Y 是独立随机变量, 且 g 和 h 是性质良好的实值函数, 则 $g(X)$ 和 $h(Y)$ 是独立随机变量。

这个引理的一个直接推论是, 独立随机变量的联合矩母函数可以分解为边缘矩母函数的乘积。如果 X 和 Y 独立, 则

$$M_{X,Y}(t, u) = \mathbb{E}[e^{tX+uY}] = \mathbb{E}[e^{tX}]\mathbb{E}[e^{uY}] = M_X(t)M_Y(u)$$

3.4.2 Mutual independence

定义 3.11 (相互独立随机变量) 随机变量 X_1, \dots, X_n 是相互独立 (*mutually independent*) 的当且仅当对于所有 x_1, x_2, \dots, x_n 的选择, 事件 $\{X_1 \leq x_1\}, \dots, \{X_n \leq x_n\}$ 是相互独立的。

定理 3.16 (相互独立的等价表述) 如果 X_1, \dots, X_n 是随机变量, 则以下陈述等价:

- 对于所有 x_1, \dots, x_n , 事件 $\{X_1 \leq x_1\}, \dots, \{X_n \leq x_n\}$ 是独立的。
- 对于所有 x_1, \dots, x_n , $F_{X_1, \dots, X_n}(x_1, \dots, x_n) = F_{X_1}(x_1)F_{X_2}(x_2) \dots F_{X_n}(x_n)$ 。
- 对于所有 x_1, \dots, x_n , $f_{X_1, \dots, X_n}(x_1, \dots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2) \dots f_{X_n}(x_n)$ 。

定理 3.17 (独立性的保持) 如果 X_1, \dots, X_n 是独立随机变量, 且 g_1, g_2, \dots, g_n 是性质良好的实值函数, 则 $g_1(X_1), \dots, g_n(X_n)$ 是独立随机变量。

定理 3.18 (乘积的期望) 如果 X_1, \dots, X_n 是独立随机变量且它们的期望都存在, 则:

$$\mathbb{E}\left[\prod_{i=1}^n X_i\right] = \prod_{i=1}^n \mathbb{E}[X_i]$$

定理 3.19 如果 X_1, \dots, X_n 是独立随机变量, g_1, g_2, \dots, g_n 是性质良好的实值函数, 且 $\{\mathbb{E}[g_i(X_i)]\}_i$ 都存在, 则:

$$\mathbb{E}\left[\prod_{i=1}^n g_i(X_i)\right] = \prod_{i=1}^n \mathbb{E}[g_i(X_i)]$$

定理 3.20 如果 X_1, \dots, X_n 是独立随机变量, 则:

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i)$$

3.4.3 Identical distributions

定义 3.12 (同分布随机变量) 随机变量 X_1, \dots, X_n 是同分布 (*identically distributed*) 的当且仅当它们的累积分布函数相同, 即对于所有 $x \in \mathbb{R}$, 有

$$F_{X_1}(x) = F_{X_2}(x) = \dots = F_{X_n}(x)$$

如果 X_1, \dots, X_n 是同分布的, 我们通常只用字母 X 来表示一个与它们所有具有相同分布的随机变量。如果 X_1, \dots, X_n 是独立同分布的, 我们有时将其记为 $\{X_i\} \sim \text{IID}$ 。

定理 3.21 如果 X_1, \dots, X_n 是独立同分布随机变量, 则:

$$\text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \text{Var}(X_i)$$

3.5 Random vectors and random matrices

定义 3.13 (随机向量与随机矩阵) 一个 $n \times 1$ 的随机向量 (*random vector*) X 是一个 $n \times 1$ 的向量，其元素是随机变量：

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

一个 $n \times n$ 的随机矩阵 (*random matrix*) W 是一个 $n \times n$ 的矩阵，其元素是随机变量：

$$W = \begin{pmatrix} W_{1,1} & \cdots & W_{1,n} \\ \vdots & \ddots & \vdots \\ W_{n,1} & \cdots & W_{n,n} \end{pmatrix}$$

定义 3.14 (随机向量的方差) 假设 X 是一个 $n \times 1$ 随机向量。 X 的方差由下式给出：

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T] \\ &= \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_2, X_1) & \cdots & \text{Cov}(X_n, X_1) \\ \text{Cov}(X_1, X_2) & \text{Var}(X_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \text{Cov}(X_1, X_n) & \cdots & \cdots & \text{Var}(X_n) \end{pmatrix} \end{aligned}$$

因此，随机向量的方差通常称为协方差矩阵。协方差矩阵总是对称的。如果 X 中的元素不相关，则 $\text{Var}(X)$ 是一个对角矩阵。如果元素不相关且同分布，我们可以写成 $\text{Var}(X) = \sigma^2 I_n$ 。

定义 3.15 (正定矩阵与非负定矩阵) 设 A 是一个 $n \times n$ 矩阵。

- i. 如果对所有 $b \in \mathbb{R}^n$ 有 $b^T A b > 0$ ，则 A 是正定 (*positive definite*) 的。
- ii. 如果对所有 $b \in \mathbb{R}^n$ 有 $b^T A b \geq 0$ ，则 A 是非负定 (*non-negative definite*) 的（也称为半正定 (*positive semidefinite*)）。

定理 3.22 (协方差矩阵是非负定的) 如果 X 是一个随机向量且 $\Sigma = \text{Var}(X)$ ，则 Σ 是一个非负定矩阵。

3.6 Transformations of continuous random variables

3.6.1 Bivariate transformations

定义 3.16 (一一映射与满射) 考虑一个定义域为 D 、值域为 R 的函数 g 。我们说 g 是：

- i. 一一映射 (*one-to-one*) 的，如果对所有 $x_1, x_2 \in D$ ， $g(x_1) = g(x_2) \Rightarrow x_1 = x_2$ ，
- ii. 满射 (*onto*) 的，如果对所有 $y \in R$ ，我们能找到 $x \in D$ 使得 $y = g(x)$ 。

我们关注将一对随机变量变换为另一对随机变量。考虑随机变量对 (U, V) 和 (X, Y) 。假设 X 和 Y 都是 U 和 V 的函数，即

$$X = g_1(U, V),$$

$$Y = g_2(U, V).$$

逆变换为

$$U = h_1(X, Y),$$

$$V = h_2(X, Y).$$

我们将整体变换标记为 g , 即 $(X, Y) = g(U, V)$, 逆变换标记为 h , 即 $(U, V) = h(X, Y)$ 。假设 g 是一个性质良好的函数 $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ 。假设对于定义域 $D \subseteq \mathbb{R}^2$, 函数 g 是到值域 $R \subseteq \mathbb{R}^2$ 的一一映射。那么, 如果 (U, V) 是一对支撑集为 D 的连续随机变量, 且 $(X, Y) = g(U, V)$, 则 X 和 Y 的联合密度为

$$f_{X,Y}(x, y) = \begin{cases} f_{U,V}(h_1(x, y), h_2(x, y)) |J_h(x, y)| & \text{当 } (x, y) \in R \\ 0 & \text{其他} \end{cases}$$

这通常称为变量变换公式。 $J_h(x, y)$ 是逆变换的雅可比行列式, 定义为

$$J_h(x, y) = \begin{vmatrix} \frac{\partial}{\partial x} h_1(x, y) & \frac{\partial}{\partial x} h_2(x, y) \\ \frac{\partial}{\partial y} h_1(x, y) & \frac{\partial}{\partial y} h_2(x, y) \end{vmatrix}$$

$$J_h(x, y) = [J_g(h_1(x, y), h_2(x, y))]^{-1}$$

涉及变量变换的问题可以用机械的方式回答。

1. 确定变换: 找到 g_1 和 g_2 使得

$$x = g_1(u, v),$$

$$y = g_2(u, v).$$

2. 求逆变换: 找到 h_1 和 h_2 使得

$$u = h_1(x, y),$$

$$v = h_2(x, y).$$

3. 计算雅可比行列式: 求 h_1 和 h_2 关于 x 和 y 的偏导数, 从而计算

$$J_h(x, y) = \begin{vmatrix} \frac{\partial}{\partial x} h_1(x, y) & \frac{\partial}{\partial x} h_2(x, y) \\ \frac{\partial}{\partial y} h_1(x, y) & \frac{\partial}{\partial y} h_2(x, y) \end{vmatrix}$$

4. 构造 X 和 Y 的联合密度:

$$f_{X,Y}(x, y) = f_{U,V}(h_1(x, y), h_2(x, y)) |J_h(x, y)|$$

例题 3.2 (两个连续随机变量的乘积) 假设 U 和 V 是联合密度为 $f_{U,V}$ 的连续随机变量。我们希望用 $f_{U,V}$ 表示它们的乘积 UV 的密度。我们定义一个二元变换 $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$:

$$g_1(u, v) = uv,$$

$$g_2(u, v) = v.$$

令 $X = g_1(U, V) = UV$ 和 $Y = g_2(U, V) = V$ 。我们将按照上述步骤推导 X 和 Y 的联合密度。

1. 变换可以写成:

$$x = uv,$$

$$y = v.$$

如果此变换的定义域是整个 \mathbb{R}^2 , 则值域也是整个 \mathbb{R}^2 。

2. 逆变换 h 定义为:

$$u = x/y,$$

$$v = y.$$

3. 雅可比行列式为：

$$J_h(x, y) = \begin{vmatrix} \frac{1}{y} & 0 \\ -\frac{x}{y^2} & 1 \end{vmatrix} = \frac{1}{y}.$$

雅可比行列式的绝对值为 $\frac{1}{|y|}$ 。

4. X 和 Y 的联合密度为：

$$f_{X,Y}(x, y) = f_{U,V}(x/y, y) \frac{1}{|y|}.$$

我们实际上感兴趣的是乘积的密度，即 X 的密度。我们可以通过通常的方式对 y 积分来计算：

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy \\ &= \int_{-\infty}^{\infty} f_{U,V}(x/y, y) \frac{1}{|y|} dy. \end{aligned}$$

记号

我们使用 g 和 h 表示变换及其逆变换。这些函数通常被视为隐式的；给定逆变换 $u = h_1(x, y)$ 和 $v = h_2(x, y)$ ，雅可比行列式可以写成：

$$J_h(x, y) = \begin{vmatrix} \partial u / \partial x & \partial v / \partial x \\ \partial u / \partial y & \partial v / \partial y \end{vmatrix}.$$

这提供了一种记忆雅可比行列式的有用方法。在这种语境下，我们将 u 和 v 视为 x 和 y 的函数。

3.6.2 Multivariate transformations

我们现在考虑 n 个随机变量的变换。我们将使用之前建立的随机向量记号。令 $X = (X_1, \dots, X_n)^T$ 是一个连续随机向量，并令 $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ 是一个性质良好的函数。实际上，我们将假设，如果 $D \subseteq \mathbb{R}^n$ 是 X 的支撑集，则 g 是一个从 D 到值域 $R \subseteq \mathbb{R}^n$ 的一一映射。如前所述，我们将广泛使用逆变换 $h(y) = g^{-1}(y)$ ，并偶尔考虑向量的各个分量：

$$\begin{aligned} x &= (x_1, \dots, x_n)^T, \\ g(x) &= (g_1(x), \dots, g_n(x))^T, \end{aligned}$$

等等。注意这里对于 $j = 1, \dots, n$ ，每个 g_j 都是 n 个变量的函数， $g_j : \mathbb{R}^n \rightarrow \mathbb{R}$ ，所以我们可以写成：

$$g(x) = (g_1(x_1, \dots, x_n), \dots, g_n(x_1, \dots, x_n))^T.$$

现在定义随机向量 Y 为 $Y = g(X)$ 。 Y 的密度由下式给出：

$$f_Y(y) = \begin{cases} f_X(h(y)) |J_h(y)| & \text{当 } y \in R, \\ 0 & \text{其他.} \end{cases}$$

雅可比行列式定义为：

$$J_h(y) = \begin{vmatrix} \frac{\partial}{\partial y_1} h_1(y) & \frac{\partial}{\partial y_1} h_2(y) & \cdots & \frac{\partial}{\partial y_1} h_n(y) \\ \frac{\partial}{\partial y_2} h_1(y) & \frac{\partial}{\partial y_2} h_2(y) & \cdots & \frac{\partial}{\partial y_2} h_n(y) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial y_n} h_1(y) & \frac{\partial}{\partial y_n} h_2(y) & \cdots & \frac{\partial}{\partial y_n} h_n(y) \end{vmatrix}.$$

雅可比行列式也可以写成：

$$J_h(y) = [J_g(h(y))]^{-1},$$

其中 $J_g(x) = \left| \frac{\partial}{\partial x} g(x) \right|$ 。我们用一个例子来说明。

例题 3.3 (线性变换) 一个在实践中经常出现的随机向量的简单函数是线性变换：

$$Y = AX.$$

我们通常要求 A 是非奇异的，即 $|A| \neq 0$ 。这个条件确保逆矩阵 A^{-1} 是良定义的。我们可以轻易验证 $J_g(x) = |A|$ ，因此 $J_h(y) = |A|^{-1}$ 。我们得出结论：

$$f_Y(y) = \frac{1}{|A|} f_X(A^{-1}y),$$

其中 $|A|$ 是 A 的行列式的绝对值。

3.7 Sums of random variables

在许多实际情况下，感兴趣量的自然模型是随机变量的和。考虑以下示例。

1. 假设在给定季节有 n 个飓风在大西洋盆地形成。每个飓风独立地以概率 p 登陆。如果 Y 是该季节登陆的飓风总数，我们可以写成 $Y = \sum_{j=1}^n X_j$ ，其中 $\{X_j\}$ 是独立伯努利 (p) 随机变量序列。
2. 我们在英国各地的 5 个站点测量 12 月总降雨量。如果随机变量 X_i 代表我们在站点 i 的 12 月总降雨量模型，则各地点的平均总降雨量为：

$$\bar{X} = \frac{1}{5} \sum_{j=1}^5 X_j.$$

计算该平均值的一个关键组成部分是随机变量的和。

我们从考虑一对随机变量的和开始。

3.7.1 Sum of two random variables

我们之前已经有两个关于一对随机变量和的结果。如果 X 和 Y 是随机变量，则：

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y],$$

$$\text{Var}(X + Y) = \text{Var}(X) + 2\text{Cov}(X, Y) + \text{Var}(Y).$$

实际上，利用期望的线性和二项式展开，两个随机变量和的 r 阶矩为：

$$\mathbb{E}[(X + Y)^r] = \sum_{j=0}^r \binom{r}{j} \mathbb{E}[X^j Y^{r-j}].$$

我们可以容易地推导出两个随机变量和的质量函数或密度函数。

定理 3.23 (两个随机变量和的质量/密度函数) 令 X 和 Y 为联合质量/密度函数为 $f_{X,Y}$ 的随机变量。如果 $Z = X + Y$ ，则 Z 的质量/密度函数为：

$$f_Z(z) = \begin{cases} \sum_u f_{X,Y}(u, z-u) & (\text{离散情形}) \\ \int_{-\infty}^{\infty} f_{X,Y}(u, z-u) du & (\text{连续情形}). \end{cases}$$

当 X 和 Y 独立时，我们可以利用质量函数或密度函数的因式分解。

定理 3.24 (两个独立随机变量和的质量/密度函数) 如果 X 和 Y 是独立随机变量且 $Z = X + Y$, 则 Z 的质量或密度函数由下式给出:

$$f_z(z) = \begin{cases} \sum_u f_X(u)f_Y(z-u) & (\text{离散情形}) \\ \int_{-\infty}^{\infty} f_X(u)f_Y(z-u) du & (\text{连续情形}). \end{cases}$$

由该推论定义的将两个函数组合的运算称为卷积 (convolution)。两个函数的卷积用符号 * 表示, 即,

$$f_z = f_x * f_y \iff f_z(z) = \int_{-\infty}^{\infty} f_x(u)f_y(z-u)du.$$

卷积具有交换律, 因此 $f_x * f_y = f_y * f_x$ 。

许多涉及随机变量之和的问题通过使用矩母函数可以得到极大简化。如果 X 和 Y 独立, 则其和的质量函数或密度函数由卷积积分给出, 该积分可能难以计算。然而, 其矩母函数却简单地就是边缘矩母函数的乘积。

定理 3.25 (两个独立随机变量之和的矩母函数) 如果 X 和 Y 是独立随机变量, 且 $Z = X + Y$, 则 Z 的矩母函数和累积量母函数由下式给出:

$$M_z(t) = M_x(t)M_y(t),$$

$$K_z(t) = K_x(t) + K_y(t).$$

3.7.2 Sum of n independent random variables

考虑一个由 n 个独立随机变量 X_1, \dots, X_n 构成的序列。令 S 为其和, 即,

$$S = X_1 + X_2 + \dots + X_n = \sum_{j=1}^n X_j.$$

S 的质量函数或密度函数是边缘质量函数或密度函数的 n 重卷积,

$$f_s = f_{X_1} * f_{X_2} * \dots * f_{X_n}.$$

通常, 计算这个卷积需要重复的求和或积分。矩母函数提供了一个有用的替代方法。我们可以很容易地证明, 和的矩母函数是边缘矩母函数的乘积,

$$M_S(t) = M_{X_1}(t)M_{X_2}(t)\dots M_{X_n}(t).$$

注意, 此结果要求 X_1, \dots, X_n 相互独立。如果此外 X_1, \dots, X_n 是同分布的, 我们可以用 X_j 的公共矩母函数来表示和的矩母函数。

定理 3.26 (n 个独立随机变量之和的矩母函数) 假设 X_1, \dots, X_n 是一个独立同分布的随机变量序列, 且令 $S = \sum_{j=1}^n X_j$ 。则 S 的矩母函数和累积量母函数由下式给出:

$$M_S(t) = [M_X(t)]^n,$$

$$K_S(t) = nK_X(t),$$

其中 M_X 和 K_X 分别是 X_j 的公共矩母函数和累积量母函数。

表 6: 具有可加性的概率分布

分布名称	参数符号	可加性规则（相互独立前提下）
离散型分布		
二项分布	$X_i \sim \text{Bin}(n_i, p)$	若 $X_i \stackrel{\text{ind}}{\sim} \text{Bin}(n_i, p)$, 则 $\sum_{i=1}^k X_i \sim \text{Bin}\left(\sum_{i=1}^k n_i, p\right)$
要求: 成功概率 p 必须相同		
泊松分布	$X_i \sim \text{Poi}(\lambda_i)$	若 $X_i \stackrel{\text{ind}}{\sim} \text{Poi}(\lambda_i)$, 则 $\sum_{i=1}^k X_i \sim \text{Poi}\left(\sum_{i=1}^k \lambda_i\right)$
连续型分布		
正态分布	$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$	若 $X_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2)$, 则 $\sum_{i=1}^k X_i \sim \mathcal{N}\left(\sum_{i=1}^k \mu_i, \sum_{i=1}^k \sigma_i^2\right)$
卡方分布	$X_i \sim \chi^2(\nu_i)$	若 $X_i \stackrel{\text{ind}}{\sim} \chi^2(\nu_i)$, 则 $\sum_{i=1}^k X_i \sim \chi^2\left(\sum_{i=1}^k \nu_i\right)$
本质: 是伽马分布的特例		
伽马分布	$X_i \sim \text{Gamma}(\alpha_i, \beta)$	若 $X_i \stackrel{\text{ind}}{\sim} \text{Gamma}(\alpha_i, \beta)$, 则 $\sum_{i=1}^k X_i \sim \text{Gamma}\left(\sum_{i=1}^k \alpha_i, \beta\right)$
要求: 尺度参数 β 必须相同		
柯西分布	$X_i \sim \text{Cauchy}(x_{0i}, \gamma_i)$	若 $X_i \stackrel{\text{ind}}{\sim} \text{Cauchy}(x_{0i}, \gamma_i)$, 则 $\sum_{i=1}^k X_i \sim \text{Cauchy}\left(\sum_{i=1}^k x_{0i}, \sum_{i=1}^k \gamma_i\right)$

3.8 Multivariate normal distribution

在前面我们讨论了正态分布的一些性质。特别地，显然正态分布由其均值和方差唯一确定。在多变量情形下，我们可以证明正态分布之间的关系完全由它们的相关系数刻画。因此，如果随机变量是（联合）正态分布且不相关，那么它们也是独立的。

3.8.1 Bivariate case

我们的出发点是一对独立的标准正态随机变量。如果 U 和 V 是独立的 $N(0, 1)$ 随机变量，那么它们的联合密度函数和联合矩母函数分别为：

$$f_{U,V}(u, v) = \frac{1}{2\pi} e^{-(u^2+v^2)/2}, \quad \text{其中 } u, v \in \mathbb{R},$$

$$M_{U,V}(s, t) = e^{(s^2+t^2)/2}, \quad \text{其中 } s, t \in \mathbb{R}.$$

这是独立性的一个简单推论：联合密度是边缘密度的乘积，且联合矩母函数是边缘矩母函数的乘积。独立性假设是相当严格的。在实际感兴趣的情形中，所考虑的变量是相关的。我们想要的是一个标准正态分布的二元版本。以下命题指出了如何构造一个标准的二元正态分布。

定理 3.27 (标准二元正态分布的构造) 假设 U 和 V 是独立的 $N(0, 1)$ 随机变量。如果我们令 $X = U$ 且 $Y = \rho U + \sqrt{1 - \rho^2}V$ ，那么

- i. $X \sim N(0, 1)$ 且 $Y \sim N(0, 1)$,
- ii. $\text{Corr}(X, Y) = \rho$,
- iii. $f_{X,Y}(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp[-(x^2 - 2\rho xy + y^2)/(2(1 - \rho^2))]$, 其中 $x, y \in \mathbb{R}$,
- iv. $M_{X,Y}(s, t) = \exp[\frac{1}{2}(s^2 + 2\rho st + t^2)]$, 其中 $s, t \in \mathbb{R}$.

一个一般的二元正态分布可以通过位置-尺度变换从标准二元正态分布构造出来。

3.8.2 n -dimensional multivariate case

我们可以从一个独立标准正态随机变量的向量构造一个一般的 n 维多元正态分布。假设 U 是一个 $n \times 1$ 随机向量，满足 $U \sim N(0, I_n)$ 。由独立性， U 的密度函数为

$$f_U(u) = \prod_{i=1}^n (2\pi)^{-1/2} \exp(-u_i^2/2) = (2\pi)^{-n/2} \exp[-u^T u/2].$$

考虑变换

$$X = \mu + CU,$$

其中 μ 是一个 $n \times 1$ 向量， C 是一个 $n \times n$ 非奇异下三角矩阵。这是一个尺度-位置变换的多元版本。指定 C 为下三角矩阵的原因稍后会变得明显。在此变换下，

$$\mathbb{E}[X] = \mu \quad \text{且} \quad \text{Var}(X) = CC^T.$$

如果我们定义 $\text{Var}(X) = \Sigma$ ，那么方程表明 $\Sigma = CC^T$ 。该下三角矩阵 C 被称为 Σ 的 **Cholesky 分解**。Cholesky 分解可以视为对非负定矩阵进行开平方根运算的推广。如果 Σ 是正定的，则 Cholesky 分解是唯一且非奇异的。

如果我们定义 $g(u) = \mu + Cu$ ，那么方程可以写为 $X = g(U)$ 。此变换的雅可比行列式为 $J_g(u) = |C|$ 。因此，如果我们定义 $h(x) = C^{-1}(x - \mu)$ ，此逆变换的雅可比行列式为 $J_h(x) = |C|^{-1}$ 。使用变量变换公式的多元版本可得

$$\begin{aligned} f_X(x) &= f_U(h(x))|J_h(x)| \\ &= (2\pi)^{-n/2}|C|^{-1} \exp[-(x - \mu)^T(C^{-1})^T(C^{-1})(x - \mu)/2] \\ &= (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp[-(x - \mu)^T\Sigma^{-1}(x - \mu)/2]. \end{aligned}$$

在推导此密度表达式时，我们使用了 $|C| = |\Sigma|^{1/2}$ 以及 $(C^{-1})^T(C^{-1}) = \Sigma^{-1}$ 这一事实，即逆矩阵的 Cholesky 分解是 Cholesky 分解的逆。

一个直接的推论是：任何联合正态分布随机变量的线性组合也服从正态分布。

定理 3.28 如果 $X \sim N(\mu, \Sigma)$ 且 $a = (a_1, \dots, a_n)^T$ 是一个常数向量，则

$$a^T X = a_1 X_1 + \dots + a_n X_n \sim N(a^T \mu, a^T \Sigma a).$$

3.9 Further exercises

1. 掷一个公平的骰子 2 次，得到的结果分别为 X_1, X_2 ，求 $X_1 > X_2$ 的概率。

解答：

利用对称性， $2P(X_1 > X_2) + P(X_1 = X_2) = 1$ ，

$$P(X_1 = X_2) = \frac{1}{6} \times \frac{1}{6} \times 6 = \frac{1}{6}。$$

$$\text{因此 } P(X_1 > X_2) = \frac{5}{12}。$$

2. 掷一个公平的骰子 3 次，得到的结果分别为 X_1, X_2, X_3 ，求 $X_1 > X_2 > X_3$ 的概率。

解答：

用组合方法，总共有 $6^3 = 216$ 种可能，不重复的数字共 $C_6^3 = 20$ 种可能，故概率等于 $\frac{20}{216} = \frac{5}{54}$ 。

3. 随机生成一个变量 X_1 服从连续均匀分布 $[a, b]$ ，再随机生成一个变量 X_2 来自相同的分布，求 $X_1 > X_2$ 的概率。

解答：

考虑 $P(X_2 - X_1 < 0)$ ，把 $X_2 - X_1$ 看作一个随机变量。

$$\text{已知 } P(X_2 - X_1 < 0) + P(X_2 - X_1 > 0) + P(X_2 - X_1 = 0) = 1,$$

$$P(X_2 - X_1 < 0) = P(X_2 - X_1 > 0) \text{ 且 } P(X_2 - X_1 = 0) = 0,$$

$$\text{得 } P(X_2 - X_1 < 0) = \frac{1}{2}。$$

4. 随机生成一个变量 X_1 服从连续均匀分布 $[a, b]$ ，再随机生成两个变量 X_2, X_3 来自相同的分布，求 $X_1 > X_2 > X_3$ 的概率。

解答：

依旧使用对称法的思想，由于是连续的随机变量，只需考虑严格大小排列的情况，一共有 $3! = 6$ 种等可能的情况，因此概率为 $\frac{1}{6}$ 。

同理可以推得 n 个随机变量的一般情况，概率为 $\frac{1}{n!}$ 。

3.10 Appendix: Proofs

4 Conditional distributions 条件分布

4.1 Discrete conditional distributions

定义 4.1 (条件质量函数) 假设 X 和 Y 是具有联合质量函数 $f_{X,Y}$ 的离散随机变量。给定 $X = x$ 时 Y 的条件质量函数定义为

$$f_{Y|X}(y|x) = P(Y = y|X = x), \quad \text{其中 } P(X = x) > 0$$

$$f_{Y|X}(y|x) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_X(x)} & \text{若 } f_X(x) > 0 \\ 0 & \text{其他情况} \end{cases}$$

定义 4.2 (条件累积分布函数 (离散情形)) 假设 X 和 Y 是离散随机变量，并令 $f_{Y|X}$ 为给定 $X = x$ 时 Y 的条件质量函数。给定 $X = x$ 时 Y 的条件累积分布函数定义为

$$F_{Y|X}(y|x) = \sum_{y_i \leq y} f_{Y|X}(y_i|x)$$

在离散情况下，条件分布函数的值可以解释为条件概率，

$$F_{Y|X}(y|x) = \sum_{y_i \leq y} \frac{f_{X,Y}(x,y_i)}{f_X(x)} = \frac{P(Y \leq y, X = x)}{P(X = x)} = P(Y \leq y|X = x)$$

定义 4.3 (条件分布的期望值 (离散情形)) 假设 X 和 Y 是离散随机变量， $f_{Y|X}$ 是给定 $X = x$ 时 Y 的条件质量函数。该条件分布的期望值由下式给出：

$$\mathbb{E}[Y|X = x] = \sum_y y f_{Y|X}(y|x)$$

4.2 Continuous conditional distributions

定义 4.4 (条件密度) 假设 X 和 Y 是联合连续的随机变量，其联合密度为 $f_{X,Y}$ ，且 X 的边缘密度为 f_X 。给定 $X = x$ 时 Y 的条件密度定义为

$$f_{Y|X}(y|x) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_X(x)} & \text{若 } f_X(x) > 0 \\ 0 & \text{其他情况} \end{cases}$$

给定 $X = x$ 时 Y 的累积分布函数和期望值通过积分求得：

$$F_{Y|X}(y|x) = \int_{-\infty}^y f_{Y|X}(u|x) du$$

$$\mathbb{E}[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy$$

4.3 Relationship between joint, marginal, and conditional

联合密度函数也可以通过条件密度和边缘密度表示为：

$$f_{X,Y}(x,y) = f_{Y|X}(y|x)f_X(x) = f_{X|Y}(x|y)f_Y(y)$$

4.4 Conditional expectation and conditional moments

4.4.1 Conditional expectation

我们已明确 $Y|X = x$ 表示一个具有密度函数 $f_{Y|X}(y|x)$ 的随机变量。该随机变量的期望值是 x 的一个函数，记作 $\psi(x)$ 。事实证明，随机变量 $\psi(X)$ 有许多有用的应用。我们称该随机变量为给定 X 时 Y 的 **条件期望**。

定义 4.5 (条件期望) 假设 X 和 Y 是随机变量。我们定义

$$\psi(x) = \mathbb{E}[Y|X = x] = \begin{cases} \sum_y y f_{Y|X}(y|x) & (\text{离散情形}) \\ \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy & (\text{连续情形}) \end{cases}$$

给定 X 时 Y 的条件期望是 $\mathbb{E}[Y|X] = \psi(X)$ ，这是一个随机变量。

我们可以使用条件期望 $\mathbb{E}[Y|X]$ 来求 Y 的期望值。此结果通常被称为**迭代期望定律**。

定理 4.1 (迭代期望定律) 对于任意两个随机变量 X 和 Y ，有

$$\mathbb{E}[\mathbb{E}[X | Y]] = \mathbb{E}[X]$$

定理 4.2 如果 X 和 Y 独立，则

$$\mathbb{E}[X | Y] = \mathbb{E}[X]$$

4.4.2 Conditional moments

定理 4.3 (函数 $g(Y)$ 的条件期望) 假设 $g : \mathbb{R} \rightarrow \mathbb{R}$ 是一个性质良好的函数。令 X 和 Y 为具有条件质量/密度函数 $f_{Y|X}$ 的随机变量。如果我们定义

$$h(x) = \mathbb{E}[g(Y)|X = x] = \begin{cases} \sum_y g(y) f_{Y|X}(y|x) & (\text{离散情形}) \\ \int_{-\infty}^{\infty} g(y) f_{Y|X}(y|x) dy & (\text{连续情形}) \end{cases}$$

则给定 X 时 $g(Y)$ 的条件期望是 $h(x)$ ，这是一个随机变量。

通常，对于性质良好的函数 g 和 h ，有

$$\mathbb{E}[h(X)g(Y)|X] = h(X)\mathbb{E}[g(Y)|X]$$

请始终记住，任何条件期望都是我们所条件化的随机变量的函数。

定义 4.6 (条件矩与条件中心矩) 对于随机变量 X 和 Y ，给定 X 时 Y 的 r 阶**条件矩**是 $\mathbb{E}[Y^r|X]$ ，给定 X 时 Y 的 r 阶**条件中心矩**是 $\mathbb{E}[(Y - \mathbb{E}[Y|X])^r|X]$ 。

定义 4.7 (条件方差) 对于随机变量 X 和 Y ，给定 X 时 Y 的**条件方差**为

$$\text{Var}(Y|X) = \mathbb{E}[(Y - \mathbb{E}[Y|X])^2|X]$$

定理 4.4 (条件方差的另一种表示) 对于随机变量 X 和 Y ，给定 X 时 Y 的条件方差可以写为

$$\text{Var} = \mathbb{E}[Y^2|X] - \mathbb{E}[Y|X]^2$$

定理 4.5 (方差的分解) 对于随机变量 X 和 Y ， Y 的方差由下式给出：

$$\text{Var} = \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\mathbb{E}[Y|X])$$

4.4.3 Conditional moment generating functions

定义 4.8 (条件矩母函数) 对于随机变量 X 和 Y , 给定 X 时 Y 的条件矩母函数为

$$M_{Y|X}(u|X) = \mathbb{E}[e^{uY}|X]$$

条件矩母函数可用于计算 Y 的边缘矩母函数以及 X 和 Y 的联合矩母函数。

定理 4.6 (条件矩母函数的性质) 如果 X 和 Y 是具有条件矩母函数 $M_{Y|X}(u|X)$ 的随机变量, 则

- i. $M_Y(u) = \mathbb{E}[M_{Y|X}(u|X)]$
- ii. $M_{X,Y}(t, u) = \mathbb{E}[e^{tX} M_{Y|X}(u|X)]$

4.5 Hierarchies and mixtures

在许多情况下, 很自然地会将模型构建为层次结构。在**层次模型**中, 我们感兴趣的随机变量 (例如 Y) 的分布依赖于其他随机变量。在这种情况下, 我们称 Y 具有**混合分布**。考虑以下说明。一个两层的层次模型可以总结如下: 如果 Y 是我们关心的量, 起初我们并不直接知道 Y 的边缘分布。相反, 这种情况最自然的描述方式是通过给定 $X = x$ 时 Y 的条件分布以及 X 的边缘分布。

4.6 Random sums

定理 4.7 (随机和的条件结果) 假设 $\{X_j\}$ 是一个独立同分布的随机变量序列, 且 N 是一个取非负整数值的随机变量, 与 $\{X_j\}$ 独立。如果 $Y = \sum_{j=1}^N X_j$, 则

- i. $\mathbb{E}[Y|N] = N\mathbb{E}[X]$
- ii. $\text{Var}(Y|N) = N\text{Var}(X)$
- iii. $M_{Y|N}(t|N) = [M_X(t)]^N$
- iv. $K_{Y|N}(t|N) = NK_X(t)$

定理 4.8 (随机和的边缘结果) 假设 $\{X_j\}$ 是一个独立同分布的随机变量序列, 且 N 是一个取非负整数值的随机变量, 与 $\{X_j\}$ 独立。如果 $Y = \sum_{j=1}^N X_j$, 则

- i. $\mathbb{E}[Y] = \mathbb{E}[N]\mathbb{E}[X]$
- ii. $\text{Var}(Y) = \mathbb{E}[N]\text{Var}(X) + \mathbb{E}[X]^2\text{Var}(N)$
- iii. $M_Y(t) = M_N(\log M_X(t))$
- iv. $K_Y(t) = K_N(K_X(t))$

4.7 Conditioning for random vectors

考虑随机向量 $X = (X_1, \dots, X_m)^T$ 和 $Y = (Y_1, \dots, Y_n)^T$ 。 X 和 Y 的联合分布就是 X 中所有变量与 Y 中所有变量的联合分布,

$$f_{X,Y}(x, y) = f_{X_1, \dots, X_m, Y_1, \dots, Y_n}(x_1, \dots, x_m, y_1, \dots, y_n)$$

我们可以用与单变量情况完全相同的方式来定义条件质量/密度函数,

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)} \quad \text{其中} \quad f_X(x) > 0.$$

我们将利用以下表达式

$$f_{X,Y}(x,y) = f_{Y|X}(y|x)f_X(x).$$

考虑三个随机变量 X_1, X_2 , 和 X_3 的情况。我们可以先将它们分组为 (x_3) 和 (x_1, x_2) 。于是方程意味着

$$f_{X_1, X_2, X_3}(x_1, x_2, x_3) = f_{X_3|X_1, X_2}(x_3|x_1, x_2)f_{X_1, X_2}(x_1, x_2).$$

再次应用可得 $f_{X_1, X_2}(x_1, x_2) = f_{X_2|X_1}(x_2|x_1)f_{X_1}(x_1)$ 。将此结果代入方程可得

$$f_{X_1, X_2, X_3}(x_1, x_2, x_3) = f_{X_3|X_1, X_2}(x_3|x_1, x_2)f_{X_2|X_1}(x_2|x_1)f_{X_1}(x_1).$$

通常, 我们可以将一个随机向量的联合质量/密度函数分解为条件质量/密度函数的乘积。以下命题提供了详细信息。

定理 4.9 (质量/密度函数的分解) 如果我们定义一个 n 维随机向量 $X_n = (X_1, \dots, X_n)^T$, 且 $x_n = (x_1, \dots, x_n)^T \in \mathbb{R}^n$, 那么

$$\begin{aligned} f_{X_n}(x_n) &= f_{X_n|X_{n-1}}(x_n|x_{n-1})f_{X_{n-1}|X_{n-2}}(x_{n-1}|x_{n-2}) \cdots f_{X_2|X_1}(x_2|x_1)f_{X_1}(x_1) \\ &= \prod_{j=1}^n f_{X_j|X_{j-1}}(x_j|x_{j-1}), \end{aligned}$$

其中我们定义 $f_{X_1|X_0}(x_1|x_0) = f_{X_1}(x_1)$ 。

定理 4.10 (多元正态分布的条件分布) 假设随机向量 $X = (X_1, \dots, X_n)^T$ 和 $Y = (Y_1, \dots, Y_m)^T$ (对于某些整数 n 和 m) 是联合正态的, 且 $X \sim N(\mu_X, \Sigma_X)$, $Y \sim N(\mu_Y, \Sigma_Y)$ 。如果此外,

$$\text{Cov}(X, Y) = \Sigma_{XY} = \Sigma_{YX}^T,$$

那么

$$\mathbb{E}[Y|X] = \mu_Y + \Sigma_{YX}\Sigma_X^{-1}(X - \mu_X),$$

$$\text{Var}(Y|X) = \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY},$$

并且给定 $X = x$ 时 Y 的条件分布是多元正态的。

4.8 Further exercises

1. 设 X 与 Y 为独立随机变量, 其分布如下:

$X = x$	0	1	2	$Y = y$	1	2
$p_X(x)$	0.4	0.2	0.4	$p_Y(y)$	0.4	0.6

定义随机变量 $W = 2X$, $Z = Y - X$ 。

- (a) 求 W 和 Z 各自的分布。
- (b) 求 W 与 Z 的联合分布。
- (c) 计算 $P(W = 2 | Z = 1)$ 、 $\mathbb{E}(W | Z = 0)$ 以及 $\text{Cov}(W, Z)$ 。

解答:

(a)

w	0	2	4	
$P(W = w)$	0.4	0.2	0.4	
z	-1	0	1	2

$P(Z = z)$	0.16	0.32	0.28	0.24
------------	------	------	------	------

(b)

注意虽然 X 与 Y 独立, 但 $W = 2X$ 与 $Z = Y - X$ 都依赖于 X , 因此 W 与 Z 不独立。一个常见的错误做法是直接 $P(W = w, Z = z) = P_W(w) \cdot P_Z(z)$ (假定了 W 与 Z 独立)。正确的做法应从 (X, Y) 的联合分布出发, 通过变换得到 (W, Z) 的联合分布。

$X \setminus Y$	1	2
0	0.16	0.24
1	0.08	0.12
2	0.16	0.24

- $(X = 0, Y = 1)$: $W = 0, Z = 1$, 概率 0.16
- $(X = 0, Y = 2)$: $W = 0, Z = 2$, 概率 0.24
- $(X = 1, Y = 1)$: $W = 2, Z = 0$, 概率 0.08
- $(X = 1, Y = 2)$: $W = 2, Z = 1$, 概率 0.12
- $(X = 2, Y = 1)$: $W = 4, Z = -1$, 概率 0.16
- $(X = 2, Y = 2)$: $W = 4, Z = 0$, 概率 0.24

$W \setminus Z$	-1	0	1	2
0	0	0	0.16	0.24
2	0	0.08	0.12	0
4	0.16	0.24	0	0

(c)

$$P(Z = 1) = 0.28, \quad P(W = 2, Z = 1) = 0.12$$

$$P(W = 2 | Z = 1) = \frac{0.12}{0.28} = \frac{3}{7}$$

$$P(W = 2 | Z = 0) = \frac{0.08}{0.32} = 0.25, \quad P(W = 4 | Z = 0) = \frac{0.24}{0.32} = 0.75$$

$$\mathbb{E}(W | Z = 0) = 2 \times 0.25 + 4 \times 0.75 = 0.5 + 3 = 3.5$$

$$\mathbb{E}[W] = 0 \times 0.4 + 2 \times 0.2 + 4 \times 0.4 = 2.0$$

$$\mathbb{E}[Z] = (-1) \times 0.16 + 0 \times 0.32 + 1 \times 0.28 + 2 \times 0.24 = 0.6$$

$$WZ = 2X(Y - X) = 2XY - 2X^2$$

$$\mathbb{E}[XY] = 0 \times 1 \times 0.16 + 0 \times 2 \times 0.24 + 1 \times 1 \times 0.08 + 1 \times 2 \times 0.12 + 2 \times 1 \times 0.16 + 2 \times 2 \times 0.24 = 1.6$$

$$\mathbb{E}[X^2] = 0 \times 0.4 + 1 \times 0.2 + 4 \times 0.4 = 1.8$$

$$\mathbb{E}[WZ] = 2 \times 1.6 - 2 \times 1.8 = -0.4$$

$$\text{Cov}(W, Z) = \mathbb{E}[WZ] - \mathbb{E}[W]\mathbb{E}[Z] = -0.4 - 2.0 \times 0.6 = -1.6$$

2. 随机变量 X 和 Y 的联合概率分布如下：

	$X = -1$	$X = 0$	$X = 1$
$Y = -1$	0.05	0.15	0.10
$Y = 0$	0.10	0.05	0.25
$Y = 1$	0.10	0.05	0.15

(a) 求 X 和 Y 的边缘分布以及给定 $Y = 1$ 时 X 的条件分布。

(b) 计算 $\mathbb{E}(X | Y = 1)$ 以及 X 和 Y 的相关系数。

(c) X 和 Y 是相互独立的随机变量吗？

3. X_1, X_2, \dots, X_n 是独立的伯努利随机变量。 X_i 的概率函数由下式给出：

$$p(x_i) = \begin{cases} (1 - \pi_i)^{1-x_i} \pi_i^{x_i} & \text{当 } x_i = 0, 1 \\ 0 & \text{其他情况} \end{cases}$$

其中：

$$\pi_i = \frac{e^{i\theta}}{1 + e^{i\theta}}$$

对于 $i = 1, 2, \dots, n$ 。推导联合概率函数 $p(x_1, x_2, \dots, x_n)$ 。

4. X_1, X_2, \dots, X_n 是独立的随机变量，具有共同的概率密度函数：

$$f(x) = \begin{cases} \lambda^2 x e^{-\lambda x} & \text{当 } x \geq 0 \\ 0 & \text{其他情况.} \end{cases}$$

推导联合概率密度函数 $f(x_1, x_2, \dots, x_n)$ 。

5. X_1, X_2, \dots, X_n 是独立的随机变量，具有共同的概率函数：

$$p(x) = \binom{m}{x} \frac{\theta^x}{(1+\theta)^m} \quad \text{当 } x = 0, 1, 2, \dots, m$$

否则为 0。推导联合概率函数 $p(x_1, x_2, \dots, x_n)$ 。

4.9 Appendix: Proofs

Appendix A: Cheatsheet for Expectation, Variance, and Covariance

公式	描述
$\mathbb{E}[a + bX] = a + b\mathbb{E}[X]$	期望的线性性质
$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$	期望的可加性
$\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$	柯西-施瓦茨不等式
$\mathbb{E}[\mathbb{E}[X Y]] = \mathbb{E}[X]$	迭代期望定律
$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$	方差定义
$\text{Var}(a + bX) = b^2\text{Var}(X)$	方差的缩放与平移
$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y) \pm 2\text{Cov}(X, Y)$	和/差的方差
$\text{Var}(Y X) = \mathbb{E}[(Y - \mathbb{E}[Y X])^2 X] = \mathbb{E}[Y^2 X] - \mathbb{E}[Y X]^2$	条件方差定义
$\text{Var}(Y) = \mathbb{E}[\text{Var}(Y X)] + \text{Var}(\mathbb{E}[Y X])$	方差分解公式
$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$	协方差定义
$\text{Cov}(X + c, Y) = \text{Cov}(X, Y)$	协方差的平移不变性
$\text{Cov}(X + Z, Y) = \text{Cov}(X, Y) + \text{Cov}(Z, Y)$	协方差的线性性

当 X 和 Y 独立时

公式	描述
$\mathbb{E}[\prod_{i=1}^n X_i] = \prod_{i=1}^n \mathbb{E}[X_i]$	独立随机变量乘积的期望
$\mathbb{E}[X Y] = \mathbb{E}[X]$	条件期望等于无条件期望
$\text{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \text{Var}(X_i)$	独立随机变量和的方差
$\text{Cov}(X, Y) = 0$	独立意味着不相关

当 X 和 Y 独立同分布时

公式	描述
$\text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \text{Var}(X_i)$	样本均值的方差

Appendix B: Cheatsheet for Common Univariate Distributions

Distribution	Notation	Parameters	Support	PMF/PDF	CDF	Mean	Variance	PGF	MGF
Bernoulli	$B(1, p)$	$0 \leq p \leq 1$	$x \in \{0, 1\}$	$f_X(x) = \begin{cases} 1-p & \text{if } x=0 \\ p & \text{if } x=1 \end{cases}$	$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1-p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$	p	$p(1-p)$	$1-p+pz$	$1-p+pe^t$
Binomial	$B(n, p)$	$n \in \{0, 1, 2, \dots\}$ $p \in [0, 1]$	$x \in \{0, 1, \dots, n\}$	$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$	$F_X(x) = \sum_{k=0}^x (n-k)! / k! (n-x)! (1-p)^{n-x}$	np	$np(1-p)$	$[1-p+pz]^n$	$(1-p+pe^t)^n$
Geometric (x trials)	$Geo(p)$	$0 < p \leq 1$	$x \in \mathbb{N}$	$f_X(x) = (1-p)^{x-1} p$	$F_X(x) = \begin{cases} 1 - (1-p)^{ x } & \text{for } x \geq 1 \\ 0 & \text{for } x < 1 \end{cases}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pz}{1-(1-p)z}$	$\frac{pe^t}{1-(1-p)e^t}$ for $t < -\ln(1-p)$
Poisson	$Pois(\lambda)$	$\lambda > 0$	$x \in \mathbb{N}_0$	$f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}$	$F_X(x) = \frac{\Gamma(x +1, \lambda)}{ x !}$	λ	λ	$\exp[\lambda(z-1)]$	$\exp[\lambda(e^t - 1)]$
Discrete Uniform	$Unif[a, b]$	a, b integers, $b \geq a$ $n = b - a + 1$	$x \in \{a, a+1, \dots, b\}$	$f_X(x) = \frac{1}{n}$	$F_X(x) = \frac{ x -a+1}{n}$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2 - 1}{12}$	$\frac{z^a - z^{b+1}}{n(1-z)}$	$\frac{e^{at} - e^{(b+1)t}}{n(1-e^t)}$
(Continuous) Uniform	$Unif[a, b]$	$-\infty < a < b < \infty$	$x \in [a, b]$	$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$	$F_X(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{ x -a}{b-a} & \text{for } x \in [a, b] \\ 1 & \text{for } x > b \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$		$\begin{cases} \frac{e^{tb} - e^{ta}}{t(b-a)} & \text{for } t \neq 0 \\ 1 & \text{for } t = 0 \end{cases}$
Exponential	$Exp(\lambda)$	$\lambda > 0$	$x \in [0, \infty)$	$f_X(x) = \lambda e^{-\lambda x}$	$F_X(x) = 1 - e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$		$\frac{\lambda}{\lambda-t}$, for $t < \lambda$
Normal	$N(\mu, \sigma^2)$	$\mu \in \mathbb{R}$ $\sigma^2 \in \mathbb{R}_{>0}$	$x \in \mathbb{R}$	$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$	$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$	μ	σ^2		$\exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$

图 2: 常见的单变量分布

5 Sample moments and quantiles 样本矩和分位数

5.1 Core Mathematical Models Revisit: Random Variables and Distributions

我们先对前面概率论部分研究的重点概念依次做一个梳理：

随机试验

人类对于概率论的探索来自生活中的随机现象，我们把(1)理论上可在相同条件下重复；(2)每次试验的结果不止一个，并且所有可能的结果在试验前是明确的；(3)在每次试验进行之前，不能预知确切的结果的过程称为**随机试验** (random experiments)。注意，随机试验是我们抽象出来为试验结果、样本空间、事件等后续概念提供场域的广义概念。它可以是诸如抛一枚硬币的**主动型试验** (active experiments)，但我们更多关注的是我们不去干预，只是观察和记录一个自然发生或社会发生的**被动型试验** (passive experiments)，或者叫**观察型试验** (observational experiments)。这个过程本身就在按照其内在规律运行，并产生不确定的结果。概率论并不关心随机性的来源是什么，它只关心结果不确定这一核心特征。这样的抽象便于我们用一套统一的数学工具分析和解决问题。事实上，在引出了随机变量和概率分布的概念后我们会渐渐淡化和忽略对“随机试验”的关注和讨论。

概率

为了量化这些现象的“不确定性”，产生了“**概率** (probability)”的朴素概念。尽管频率学派和贝叶斯学派对于“概率”的哲学内涵有着不用的解读，在数学语言的抽象上双方都同意由柯尔莫哥洛夫等人建立的公理化概率论体系，以及随机变量、分布函数等概念的利用。在公理化体系下，事件域是样本空间子集的集合，而概率是事件（事件域的元素）到 $[0, 1]$ 的映射。例如，事件 $A = \{\omega_1, \omega_2\}$ 的概率写作 $P(A)$ 。

随机变量

这是概率论最重要的一个概念。**随机变量** (random variable) 是一个函数，把随机试验的结果映射到实数，从而把随机试验不同结果的概率转化为了随机变量不同取值的概率。例如， $P(A)$ 可以写作 $P(X = x_1, x_2 | X(\omega_1) = x_1, X(\omega_2) = x_2)$ 。

注意，从严格定义来看，随机变量的概念必然依赖于它背后的随机试验，这正是它描述的对象。然而，在我们真正去使用一个随机变量的场景下，你会发现我们往往并不去讨论它背后具体的样本空间和随机试验是什么。这不仅是因为我们不关心背后随机试验的物理故事，更是因为很多时候面对一个复杂的随机系统，我们很难严格地写出这个随机试验具体是什么。但我们知道，一个随机变量的背后一定对应着一个随机试验，即便有时这个“试验”并不天然存在，需要人为构造。我们通过一个例子来理解“随机试验有时难以明确”这一点：当我们把一个城市某年的犯罪率看作随机变量时，其对应的那个“随机试验”应该是什么？一个可能的答案是：想象我们有一个“世界模拟器”，可以固定该城市的基本面（人口结构、经济水平、法律框架等），但让所有随机的、不可预测的事件（如：某个关键政策执行人的决策、一场罕见的自然灾害、一股无法预测的犯罪潮流、甚至一些偶然的个体行为）重新随机发生一遍。而这个随机试验的一次试验即为运行一次模拟器，生成一个完整的年份，并记录下这一年的犯罪率。可以看到，这个试验不是发生在现实城市中的，而是发生在我们头脑的想象空间或统计空间里。这给了我们两个启示：

首先，在上述例子中，随机试验的形式特别复杂并且是理论假设的，这意味着如果我们在使用“随机变量”这个概念之前要先明确“随机试验”是什么，那么包括上述例子在内的诸多情境我们都不太可能应用随机变量的概念，即便是应用，在第一步也会耗费大量无意义的精力。事实上，我们之所以会常常直接定义一个随机变量正是因为这个世界的很多自然和社会机制非常复杂，我们只能观察到试验的结果，而随机试验的具体形式往往涉及到对于自然和社会具体因果机制的询问，这是

我们不关心的。我们只关心相关性，而不关心因果性。

因此，我们在应用的时候往往会习惯将随机变量视为一个独立且纯粹的数学模型。在建立这个模型的时候，我们不需要从随机试验是什么开始搭建，在设定了一个随机变量后同样也不用去纠结它背后的随机试验是什么，因为我们知道一定可以写出它的形式，并且明确试验的形式不会给我们增加任何直觉洞见或者信息，随机变量的取值已包含了这个试验可能结果的全部信息。在设定了一个随机变量后，我们真正需要关心的，是它的概率分布。

概率分布

一个随机变量一定有其概率分布 (probability distribution)。如果概率分布已知，那么这个随机变量的全部信息就已知了，但往往概率分布是未知的。

5.2 Population, sampling, sample, and observed sample

5.2.1 What is ‘Population’: Population and Data Generating Process (DGP)

从本章开始，我们要将随机变量和其背后的概率分布这个数学模型应用到现实数据中，帮助我们进行参数估计和统计检验。在此之前，我们要先明确一些概念，这些概念在别处可能有诸多含义，但在统计学中有其特殊的含义。首先，我们需要明确什么是统计学中的“总体”。我们先介绍一种常见的对于总体的定义，再来说明这种定义有什么问题，继而给出我们的一种更合适的定义。

社会科学界一种常见的说法是，总体是研究者感兴趣的对象的全体，例如，“高二一班的所有学生”、“上海市全体网民”、“牧场的所有奶牛”。要研究的有关总体对象的具体变量的全部数据继而被称为总体数据。然而，当研究者感兴趣的变量很明晰的时候，统计学中一种更方便的说法是其在现实生活中的所有数据被称为总体，它们是一堆具体数值 $\{a_1, a_2, \dots\}$ ，例如所有美国男性的身高数据。总体的范围取决于研究者感兴趣的范围。注意，在这种定义下，我们必须强调，总体必须是能够存在于这个真实世界的数据，它不一定需要是一组具体的、完全可观测的现实世界数据点，但它需要是能够存在于现实世界中的。它可以是已经存在过的数据（1900 年英国的 GDP），可以是即将存在的数据（2050 年 1 月 1 日剑桥的气温），或者是在未来如果某些条件发生下会存在的数据（如果无限次投放广告会产生的点击率数据、如果无限次抛硬币会产生的结果数据、药物如果给所有人服用得到的药效的数据），但不能是存在于平行世界的数据。

这种将总体定义为感兴趣的数据的全体的理解非常易懂，但却会带来一些困扰。

首先，如果总体是一堆数据，那么理论上存在总体数据的频数分布，它是确定的，并且是不规则的、有锯齿的，因为它完全反映了这组特定数据的实际情况。如果我们完全获知了总体的所有数据，那么我们就在统计学上拥有了总体的全部信息，可以画出总体的频数分布。然而，在现实中我们往往没法知道总体的全部信息（如下图），因此我们采用对总体取样的方法通过样本数据的性质来对总体数据的性质进行推断和估计。抽样方法的分析起点是对随机变量和概率分布模型的应用。我们会把一组随机样本看作一组随机变量，而这些随机变量都服从同一个概率分布。按照之前的定义，其对应的随机试验便是随机从这堆总体数据中抽取一个数据的过程，因此这个概率分布被称为总体的概率分布。基于古典概率的逻辑可以得出总体概率分布与总体频数分布是相同形状的，其图像可仅通过纵轴一定比例的拉伸互相转换。也就是说，在这种总体的定义下，总体的概率与频数分布在信息层面是完全等价的，如果能收集到现实世界中所有的总体数据，那么我们就可以求出总体的概率分布。

这种理解有一个致命的缺陷，思考下面这个问题：我们为什么在乎的是总体的概率/频数分布而不是样本的概率/频数分布？因为我们认为总体才包含最完整的数据和信息，而一个样本的数据只是随机试验一次片面的、偶然的实现，所以其分布不具有意义和价值。我们想了解的是那个最顶层的随机试验的结果的概率分布这个“终极奥义”，而不是“随机从这堆样本数据中抽取一个数据”

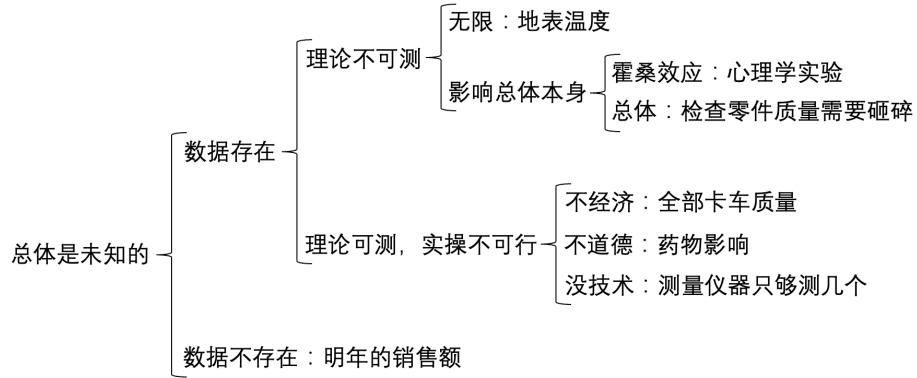


图 3: 总体的未知性 (假设总体为数据)

这个子随机试验的结果的概率分布, 这是 “片面的真理”。但你有没有想过, 如果 “随机从这堆总体数据中抽取一个数据” 也是一个更高的随机试验的子随机试验呢? 也就是说, 你能看到的现实生活中存在的这些 “总体数据”, 不过也是这个更高的随机试验的一次偶然实现罢了, 那么你所谓的 “总体概率分布” 不是也失去了意义? 事实上, 从逻辑上我们几乎可以相信, “终极奥义” 应是一个 “按照某个规则随机生成一个数据” 的最顶层随机试验而非一堆数据, 因为若存在一堆数据, 我们就可以将其看作一个更高的随机试验的一次实现。下面我们通过两个具体例子来体会这一点。

考虑总体为一批总计 5 万颗零件的重量, 则按照定义我们认为其最本质的总体概率分布对应的是 “随机从这 5 万颗零件中随机抽取 1 颗并测量它的重量” 这个随机试验, 但这 5 万颗零件又何尝不具有与抽样行为相似的随机性, 又何尝不是生产过程这个随机试验的一次实现呢? 因此, 最本质的总体不应该是一堆现实世界存在的数据, 甚至是所有平行世界中的数据, 而应该是一个数据生成过程, 一个逻辑最底层的随机试验对应的概率分布。

再考虑蜂鸟的平均寿命这个总体, 如果我抓住了全世界存在的所有蜂鸟并得到了它们寿命的均值数据, 你会更愿意相信这个数据代表着蜂鸟的平均寿命这个真理值, 还是现存所有蜂鸟的寿命作为一个样本的样本均值的观测值呢? 我想显然是后者。

一个比较底层的思考是这样的: 这个世界上的所谓 “数据” 从何而来, 它都是这个世界某个机制和过程的结果被我们观测到。如果同一个变量我们可以观测到不同的结果, 我们就可以将生成它的过程看作一个随机试验, 这个生成过程的规则, 即背后的概率分布只有上帝知道, 我们是无从获知的。这个世界上有的随机试验 (数据生成过程) 可能只留下了一次实现 (例如你某次考试的分数, 如果按照定义, 这个分数就是你的全部总体数据, 但我们仍然可以把它看作一个更底层的随机试验的一次实现值), 有的可能有很多次实现的观测值 (例如一批零件的重量)。这种思路的本质是把上帝的规则看作总体概率分布的真理和数据生成过程的规则, 因此每个平行世界中的任何数据都只是上帝生成数据这个终极随机试验的一次实现, 可以意识到这与频率学派的观点是吻合的, 即明天某地火山爆发的概率应是火山爆发发生的平行世界个数与总平行世界个数比例的极限值。

明白了将总体理解为数据的局限性后, 我们正式给出对总体的定义。我们把**总体**(population)看作一个**数据生成过程**(Data Generating Process, DGP), 即跳过现实存在的全体数据而直接关注背后更底层的对象。生成的随机变量服从的概率分布便称为**总体概率分布**, 简称**总体分布**(population distribution)。相比于纠结总体是什么, 我们更关注的是这个数据生成过程背后的概率分布。这个分布的 μ, σ^2 等特征常量被称为**总体参数**(population parameters)。

在这样的定义下, “总体是未知/不可观测的” 这句话就有了一个更合理的落脚点。

表 7: 总体视角的对比: 静态集合 vs. 动态过程

视角	总体是一堆数据 (静态集合)	总体是 DGP (动态过程)
核心思想	总体是所有现有个体的集合。	总体是生成所有可能数据 (包括已观测和未观测) 的潜在机制。
关注点	是什么: 描述这个集合的特征。	为什么: 理解数据背后的生成规则。
对“新数据”的态度	新数据是总体中尚未被测量的部分。	新数据是 DGP 一次新的、独立的运行结果。
适用范围	有限的、已存在的群体 (如“2023 年本公司所有员工”)。	无限的、理论上的、未来的群体 (如“本生产线生产的所有灯泡”)。
与概率的关系	概率是总体中的比例 (频率)。	概率是 DGP 的内在属性, 决定了不同结果出现的倾向性。

5.2.2 Sampling, sample, and observed sample

如上节所说, 我们可以通过**随机抽样** (random sampling) 的方法来研究总体性质。一个样本在被观测前, 我们可以将它视作一个随机变量序列 $\{X_1, X_2, \dots, X_n\}$ 进行建模, 观测后的样本则为一组具体实数 $\{x_1, x_2, \dots, x_n\}$ 。我们分别称呼“**样本 (sample)**”和“**观测样本 (observed sample)**”作以区分, 后者又称**观测值 (observations)**、**数据 (data)**或者**数据集 (dataset)**。(由于我们不关心所谓“**总体数据**”, 因此这种叫法不会产生歧义。) 我们有时会将样本视为一个随机向量, 并使用记号 $X = (X_1, \dots, X_n)^T$ 。类似地, 将观测样本视为一个实数向量在记号上可能更为方便, 即 $x = (x_1, \dots, x_n)^T$ 。

5.3 Independent and identically distributed (IID) sequences

5.3.1 Structural and distributional assumptions

样本的性质完全由样本成员的联合分布所刻画。然而, 我们通常并非明确地指涉联合分布, 而是根据一些能够推断出联合分布的性质来设定模型。这些假设分为两大类:

-**结构性假设**: 描述样本成员之间的关系。

-**分布性假设**: 描述样本成员的边缘分布。

一个广泛使用的结构性假设是样本成员是**独立同分布 (IID)** 的。有时会使用稍弱的结构性假设。例如, 我们可能假设样本成员是不相关的 (而非独立), 或者假设它们是同分布的 (而不对样本成员之间依赖关系的性质做任何假设)。

请记住, 正态分布之间唯一的依赖关系就是相关性, 因此一个同分布、不相关的联合正态随机变量序列就是一个独立同分布序列。

5.3.2 Random sample

我们通常会假设样本由从总体分布中抽取的独立同分布随机变量组成。这被称为一个**随机样本**。

定义 5.1 (随机样本) 随机变量集合 Y_1, \dots, Y_n 是来自累积分布函数为 F_Y 的总体的一个随机样本, 如果 Y_1, \dots, Y_n 相互独立且对于 $i = 1, \dots, n$ 有 $Y_i \sim F_Y$ 。

随机样本被用作我们对从总体中抽样，且总体中每个成员被选中的概率相等这种情况的模型。事实上，如果总体是一个有形的、有限的个体集合，我们应该对总体进行有放回抽样。当有放回抽样时，我们每次抽样都是从总体分布中抽取。因此，样本 Y_1, \dots, Y_n 由同分布的随机变量组成，因为它们每一个都具有总体分布，即对于 $i = 1, \dots, n$ 有 $Y_i \sim F_Y$ 。此外， Y_1, \dots, Y_n 相互独立。为了说明独立性，假设我们知道观测样本的第一个成员是 y_1 。这个知识不会改变 Y_2 的分布，因为

$$P(Y_2 \leq y_2 | Y_1 = y_1) = P(Y_2 \leq y_2) = F_Y(y_2)$$

如果我们进行无放回抽样，第一个样本成员是从总体分布中抽取的，但是，由于没有放回，每个后续成员是从一个更小的群体中选择的。这个更小群体的组成取决于总体中哪些成员已经被纳入样本。因此，组成样本的随机变量既不是同分布的，也不是独立的。然而，如果总体很大，移除少数个体对剩余群体的分布影响很小。我们得出结论，如果样本量相对于总体规模较小，那么在从总体中进行无放回抽样时，随机样本可能是一个良好的近似模型。

5.4 Functions of a sample

5.4.1 Statistics

样本的任何函数都是一个统计量。该函数可以是标量值的，也可以是向量值的。由于样本是一个随机向量，因此一个统计量也是一个随机向量。

定义 5.2 (统计量) 对于样本 $Y = (Y_1, \dots, Y_n)^T$ ，一个统计量 (*Statistic*) $U = h(Y)$ 是一个随机向量，它仅仅是样本和已知常数的函数。给定一个观测样本 $y = (y_1, \dots, y_n)^T$ ，我们可以计算统计量的观测值 $u = h(y)$ 。

5.4.2 Sampling distribution

统计量 U 的分布通常被称为 U 的抽样分布 (*sampling distribution*)。虽然统计量仅仅是样本和已知常数的函数，但统计量的分布可能依赖于未知参数。统计量的观测值 u 只是一个实数向量。正如观测样本 y 被视为样本 Y 的一个实例一样，值 $u = h(y)$ 被视为统计量 $U = h(Y)$ 的一个实例。

在某些情况下，我们可以使用前面章节中的分布结果来计算出感兴趣的统计量的抽样分布。在其他情况下，所涉及的数学是难以处理的，我们需要诉诸基于模拟的方法。

5.4.3 Pivotal functions

统计量本身完全是样本的函数，自身不包含任何未知参数（样本一旦确定，统计量的值也就定下来了），但是其分布却往往包含未知参数。枢轴函数恰恰相反，我们让其本身包含总体中的未知参数，而好处则是其分布形式一般是确定的，不包含未知参数。比如考虑一个随机样本 Y 服从 $N(\mu, 1)$ 分布， \bar{Y} 是样本均值，这是一个统计量，这个函数本身不包含任何未知参数但是其分布包含未知参数 μ 。相反的，考虑 $\sqrt{n}(\bar{Y} - \mu) \sim N(0, 1)$ ，这本身是一个关于样本 Y 以及未知参数 μ 的函数，但这样它的分布就不取决于 μ ，是固定的了。这个函数就是一个枢轴函数 (*Pivotal Function*)（或者叫“枢轴量”，这样与“统计量”形成对照）。

枢轴函数在置信区间的构造中起着基础性的作用。

定义 5.3 (枢轴函数) 考虑一个样本 Y 和一个标量参数 θ 。令 $g(Y, \theta)$ 是 Y 和 θ 的一个函数，且不涉及 θ 以外的任何未知参数。如果 $g(Y, \theta)$ 的分布不依赖于 θ ，我们称其为枢轴函数。

5.5 Common sampling distributions

5.5.1 Chi-squared distribution

定义 5.4 (卡方分布) 令 Z_1, \dots, Z_k 为独立的 $N(0, 1)$ 随机变量。如果：

$$X = \sum_{i=1}^k Z_i^2$$

则 X 的分布是自由度为 k 的卡方分布，记为 $X \sim \chi_k^2$ 或 $X \sim \chi^2(k)$ 。

χ_k^2 分布是一个连续分布，其取值 $x \geq 0$ 。其均值和方差为：

$$\mathbb{E}[X] = k$$

$$\text{Var}(X) = 2k$$

5.5.2 t-distribution

定义 5.5 (t 分布) 假设 $Z \sim N(0, 1)$ 且 $X \sim \chi_k^2$ ，并且 Z 和 X 独立。则随机变量：

$$T = \frac{Z}{\sqrt{X/k}}$$

的分布是自由度为 k 的（学生） t 分布，记为 $T \sim t_k$ 或 $T \sim t(k)$ 。 t 分布的密度函数为：

$$f_T(t) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})} \left(1 + \frac{t^2}{k}\right)^{-(k+1)/2} \quad \text{其中 } -\infty < t < \infty$$

对于 $T \sim t_k$ ，该分布的均值和方差为：

$$\mathbb{E}[T] = 0 \quad \text{当 } k > 1$$

$$\text{Var}(T) = \frac{k}{k-2} \quad \text{当 } k > 2$$

5.5.3 F distribution

定义 5.6 (F 分布) 令 U 和 V 为两个独立的随机变量，其中 $U \sim \chi_p^2$ 且 $V \sim \chi_k^2$ 。则：

$$F = \frac{U/p}{V/k}$$

的分布是自由度为 (p, k) 的 F 分布，记为 $F \sim F_{p,k}$ 或 $F \sim F(p, k)$ 。

F 分布是一个连续分布，对 $x > 0$ 有非零概率。

5.6 Sample mean

定义 5.7 (样本均值) 对于样本 Y_1, \dots, Y_n ，样本均值 \bar{Y} 由下式给出：

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

5.6.1 Mean and variance of the sample mean

定理 5.1 (样本均值的性质) 假设 Y_1, \dots, Y_n 是来自一个均值为 μ 的总体分布的随机样本，则

i. 样本均值的期望等于总体均值，即 $\mathbb{E}[\bar{Y}] = \mu$ 。

如果总体分布具有有限方差 $\sigma^2 < \infty$ ，则

ii. 样本均值的方差由 $\text{Var}(\bar{Y}) = \frac{\sigma^2}{n}$ 给出。

注意，样本均值的方差随着样本量的增加而减小，也就是说，取平均能够降低变异性。这是统计学中的一个基本概念；要使样本均值取极值，需要样本中的大多数取值都为极值。

5.6.2 Central limit theorem (CLT)

中心极限定理是一个卓越的结果。其推论之一是，大随机样本的样本均值近似服从正态分布，(几乎)无论我们抽取的样本来自何种分布。正态分布在此背景下的神秘出现被广泛用于统计推断。

定理 5.2 (中心极限定理) 假设 $Y = (Y_1, \dots, Y_n)^T$ 是一个随机样本，且 $\mathbb{E}[Y] = \mu$, $\text{Var}(Y) = \sigma^2$ 。若 $0 < \sigma^2 < \infty$ ，则当 $n \rightarrow \infty$ 时，

$$\frac{\bar{Y}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow{d} N(0, 1)$$

其中 $\bar{Y}_n = \frac{1}{n} \sum_{j=1}^n Y_j$ 是样本均值。

证明 (特殊情况): 我们仅证明 Y_j 的矩母函数存在的情形。如果我们定义

$$Z_n = \frac{\bar{Y}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

那么我们可以证明

$$Z_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}$$

其中 $S_n = \sum_{j=1}^n Y_j$ 是部分和。利用 $m_{S_n}(t) = [m_Y(t)]^n$ 这一事实，我们可以用 Y 的矩母函数来表示 Z 的矩母函数：

$$\begin{aligned} m_{Z_n}(t) &= \mathbb{E}[e^{tZ_n}] = \mathbb{E}\left\{\exp\left[t\left(\frac{S_n - n\mu}{\sqrt{n\sigma^2}}\right)\right]\right\} = \exp\left(-\frac{\sqrt{n}\mu t}{\sigma}\right) \mathbb{E}\left[\exp\left(\frac{t}{\sqrt{n\sigma^2}}S_n\right)\right] \\ &= \exp\left(-\frac{\sqrt{n}\mu t}{\sigma}\right) m_{S_n}\left(\frac{t}{\sqrt{n\sigma^2}}\right) = \exp\left(-\frac{\sqrt{n}\mu t}{\sigma}\right) \left[m_Y\left(\frac{t}{\sqrt{n\sigma^2}}\right)\right]^n \end{aligned}$$

取对数得到累积量母函数：

$$K_{Z_n}(t) = -\frac{\sqrt{n}\mu t}{\sigma} + nK_Y\left(\frac{t}{\sqrt{n\sigma^2}}\right)$$

根据累积量的性质：

$$K_Y(t) = \mu t + \frac{\sigma^2 t^2}{2} + (\text{包含 } t^3 \text{ 及更高阶的项})$$

关注结果中 $\frac{1}{n}$ 的幂次，我们有

$$\begin{aligned} K_{Z_n}(t) &= -\frac{\sqrt{n}\mu t}{\sigma} + n\left[\mu \frac{t}{\sqrt{n\sigma^2}} + \sigma^2 \frac{t^2}{2n\sigma^2} + \left(\text{包含 } \left(\frac{1}{n}\right)^{\frac{3}{2}} \text{ 及更高阶的项}\right)\right] \\ &= \frac{t^2}{2} + \left(\text{包含 } \left(\frac{1}{n}\right)^{\frac{1}{2}} \text{ 及更高阶的项}\right) \end{aligned}$$

当 $n \rightarrow \infty$ 时，包含 $(\frac{1}{n})^{\frac{1}{2}}$ 及更高阶的项将趋于零。因此，

$$K_{Z_n}(t) \rightarrow \frac{t^2}{2} \quad \text{当 } n \rightarrow \infty$$

我们得出结论， $K_{Z_n}(t)$ 收敛于标准正态分布的累积量母函数。由于累积量母函数唯一地刻画了一个随机变量的分布，这意味着当 $n \rightarrow \infty$ 时， $Z_n \xrightarrow{d} N(0, 1)$ 。

中心极限定理没有告诉我们任何关于收敛速率的信息；我们不知道需要多大的样本才能使正态分布对样本均值的分布提供合理的近似。事实上，这个问题的答案取决于总体分布。如果我们从正态总体中抽样，那么无论样本量大小，样本均值都服从正态分布。在另一个极端，如果我们从一个高度偏斜的离散分布中抽样，则需要相对较大的样本才能使正态分布对样本均值提供合理的近似。

上述定理是最简单的中心极限定理版本之一；存在各种不同抽象程度和通用性的中心极限定理。我们在矩母函数存在的假设下证明了该定理，这是一个相当强的假设。通过使用特征函数代替矩母函数，我们可以修改证明，使得只需一阶矩和二阶矩存在即可。进一步的推广削弱了 X_i 同分布的假设，并适用于独立随机变量的线性组合。已经有中心极限定理被证明，它们削弱了独立性假设，允许 X_i 是相关的但不是“太”相关。中心极限定理仍然是概率论中一个活跃的研究领域。

5.6.3 A Complete Proof of CLT

定理 5.3 (中心极限定理) 设 Y_1, Y_2, \dots 是独立同分布的随机变量序列，具有有限的期望 $\mathbb{E}[Y_i] = \mu$ 和有限的方差 $\text{Var}(Y_i) = \sigma^2 > 0$ 。定义标准化样本均值为：

$$Z_n = \frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^n (Y_i - \mu)}{\sigma\sqrt{n}}$$

那么，当 $n \rightarrow \infty$ 时， Z_n 依分布收敛于标准正态分布，即：

$$Z_n \xrightarrow{d} N(0, 1)$$

证明 5.1 步骤 1：定义特征函数

随机变量 X 的特征函数定义为：

$$\phi_X(t) = \mathbb{E}[e^{itX}], \quad i = \sqrt{-1}$$

特征函数总是存在，并且唯一地决定了随机变量的分布。

步骤 2：利用特征函数的性质

令 $X_i = \frac{Y_i - \mu}{\sigma}$ 。则 $\mathbb{E}[X_i] = 0$, $\text{Var}(X_i) = 1$, 且 $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ 。

由于 X_i 是独立同分布的， Z_n 的特征函数为：

$$\phi_{Z_n}(t) = \mathbb{E}[e^{itZ_n}] = \mathbb{E}\left[\exp\left(\frac{it}{\sqrt{n}} \sum_{k=1}^n X_k\right)\right] = \prod_{k=1}^n \mathbb{E}\left[\exp\left(\frac{it}{\sqrt{n}} X_k\right)\right] = \left[\phi_X\left(\frac{t}{\sqrt{n}}\right)\right]^n$$

其中 $\phi_X(t)$ 是 X_i 的共同特征函数。

步骤 3：对特征函数进行泰勒展开

由于 $\mathbb{E}[X] = 0$ 且 $\mathbb{E}[X^2] = 1$ ，我们可以对 $\phi_X(s)$ 在 $s = 0$ 处进行泰勒展开：

$$\phi_X(s) = \phi_X(0) + \phi'_X(0)s + \frac{\phi''_X(0)}{2}s^2 + o(s^2)$$

其中 $o(s^2)$ 是满足 $\lim_{s \rightarrow 0} o(s^2)/s^2 = 0$ 的高阶项。

我们知道：

$$\phi_X(0) = \mathbb{E}[e^{i0 \cdot X}] = 1$$

$$\phi'_X(0) = i\mathbb{E}[X] = 0$$

$$\phi''_X(0) = i^2\mathbb{E}[X^2] = -1$$

因此，泰勒展开式为：

$$\phi_X(s) = 1 - \frac{s^2}{2} + o(s^2)$$

步骤 4：应用展开式到 $\phi_{Z_n}(t)$

令 $s = t/\sqrt{n}$, 代入上式：

$$\phi_X\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)$$

因此：

$$\phi_{Z_n}(t) = \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n$$

步骤 5：取极限 $n \rightarrow \infty$

利用极限公式 $\lim_{n \rightarrow \infty} \left(1 + \frac{a_n}{n}\right)^n = e^a$ (当 $a_n \rightarrow a$ 时), 我们得到：

$$\lim_{n \rightarrow \infty} \phi_{Z_n}(t) = \lim_{n \rightarrow \infty} \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n = e^{-t^2/2}$$

步骤 6：结论

我们证明了：

$$\lim_{n \rightarrow \infty} \phi_{Z_n}(t) = e^{-t^2/2}$$

而 $e^{-t^2/2}$ 正是标准正态分布 $N(0, 1)$ 的特征函数。

根据 **Lévy 连续性定理**, 特征函数的逐点收敛意味着相应的分布函数弱收敛。因此：

$$Z_n \xrightarrow{d} N(0, 1)$$

定理得证。

5.7 Higher-order sample moments

定义 5.8 (样本矩与样本中心矩) 对于样本 Y_1, \dots, Y_n , 其 r 阶样本矩为

$$m'_r = \frac{1}{n} \sum_{i=1}^n Y_i^r$$

其 r 阶样本中心矩为

$$m_r = \frac{1}{n} \sum_{i=1}^n (Y_i - m'_1)^r$$

注意, 样本矩是随机变量。

5.7.1 Sample variance

定义 5.9 (样本方差) 对于样本 Y_1, \dots, Y_n (其中 $n > 1$), 其样本方差 S^2 由下式给出：

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

定理 5.4 (样本方差的四种表示) 令 Y_1, \dots 为一随机变量序列。

定义 $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ 和 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$, 则对于 $n = 2, 3, \dots$:

$$i. (n-1)S_n^2 = \sum_{i=1}^n Y_i(Y_i - \bar{Y}_n)$$

$$ii. (n-1)S_n^2 = \sum_{i=1}^n Y_i^2 - n\bar{Y}_n^2$$

$$iii. (n-1)S_n^2 = [\sum_{i=2}^n (Y_i - \bar{Y}_n)]^2 + \sum_{i=2}^n (Y_i - \bar{Y}_n)^2$$

而对于 $n = 3, 4, \dots$:

$$iv. (n-1)S_n^2 = (n-2)S_{n-1}^2 + \frac{n-1}{n}(Y_n - \bar{Y}_{n-1})^2$$

总体分布参数 (未知常数)	样本统计量 (随机变量)
总体期望/均值: $\mu = \mathbb{E}[X]$ (1 阶总体原点矩)	样本均值: $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ (1 阶样本原点矩)
总体方差: $\sigma^2 = \text{Var}(X)$ (2 阶总体中心矩)	样本方差: $S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$ (不是一个矩)
总体矩: · 中心矩: $\mathbb{E}[(X - \mathbb{E}[X])^r]$ · (原点) 矩: $\mathbb{E}[X^r]$	样本矩: · 中心矩: $\frac{1}{n} \sum (X_i - \bar{X})^r$ · (原点) 矩: $\frac{1}{n} \sum X_i^r$

定理 5.5 (样本方差的性质) 假设 Y_1, \dots, Y_n 是来自一个均值为 μ 、有限方差 $\sigma^2 < \infty$ 的总体的随机样本。如果 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$ 是样本方差, 则:

i. 样本方差的期望等于总体方差, 即 $\mathbb{E}[S^2] = \sigma^2$ 。

如果此外, 总体分布具有有限的四阶中心矩 $\mu_4 < \infty$, 则:

ii. 样本方差的方差由下式给出:

$$\text{Var}(S^2) = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \mu_2^2 \right) = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \sigma^4 \right)$$

我们已经推导了样本均值和样本方差的期望与方差。在实践中, 了解样本均值和样本方差之间的关联性也是有用的。

定理 5.6 (样本均值与样本方差的关联) 假设 Y_1, \dots, Y_n 是来自一个具有有限方差 $\sigma^2 < \infty$ 的总体的随机样本, 则:

i. \bar{Y} 与 $(Y_j - \bar{Y})$ 对于 $j = 1, \dots, n$ 是不相关的。

如果此外, 总体分布具有有限的三阶中心矩 $\mu_3 < \infty$, 则:

ii. 样本均值与样本方差之间的协方差是三阶中心矩的函数:

$$\text{Cov}(\bar{Y}, S^2) = \frac{1}{n} \mu_3.$$

我们推断, 当从对称分布 (例如正态分布) 中抽样时, 样本均值和样本方差是不相关的。

5.7.2 Joint sample moments

联合样本矩提供了关于两个变量之间依赖结构的信息。

定义 5.10 (联合样本矩与联合样本中心矩) 假设我们有一个成对数据的随机样本 $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ 。其 (r, s) 阶联合样本矩为

$$m'_{r,s} = \frac{1}{n} \sum_{i=1}^n X_i^r Y_i^s$$

其 (r, s) 阶联合样本中心矩为

$$m_{r,s} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^r (Y_i - \bar{Y})^s$$

定义 5.11 (样本协方差) 给定一个成对数据的随机样本 $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$, 其样本协方差定义为

$$c_{X,Y} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{n}{n-1} m_{1,1}$$

或者, 等价地定义为

$$c_{X,Y} = \frac{1}{n-1} \left(\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y} \right)$$

定理 5.7 (样本协方差的性质) 如果 $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ 是一个成对数据的随机样本, 则样本协方差具有以下性质:

- i. $c_{X,Y} = c_{Y,X}$
- ii. 如果对于 $i = 1, \dots, n$, 有 $U_i = a + bX_i$ 和 $V_i = c + dY_i$, 其中 a, b, c , 和 d 是实常数, 则 $c_{U,V} = bdc_{X,Y}$
- iii. 如果 $W_i = X_i + Y_i$, 则 $S_W^2 = S_X^2 + S_Y^2 + 2c_{X,Y}$, 其中 S_X^2, S_Y^2 , 和 S_W^2 分别表示 X, Y , 和 W 值的样本方差。

定义 5.12 (样本相关系数) 给定一个成对数据的随机样本 $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$, 其样本相关系数定义为

$$r_{X,Y} = \frac{c_{X,Y}}{\sqrt{S_X^2 S_Y^2}}$$

其中 S_X^2 和 S_Y^2 分别表示 X 和 Y 值的样本方差。

样本相关系数衡量了样本中存在的线性关联程度。

定理 5.8 (样本相关系数的取值范围) 给定一个成对数据的随机样本 $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$, 有

$$-1 \leq r_{X,Y} \leq 1$$

当且仅当存在常数 a 和 $b \neq 0$ 使得对于 $i = 1, 2, \dots, n$ 有 $Y_i = a + bX_i$ 时, 有

$$r_{X,Y} = \begin{cases} 1 & \text{若 } b > 0 \\ -1 & \text{若 } b < 0 \end{cases}$$

5.8 Sample mean and variance for a normal population

假设总体分布是正态的。我们可以用已知的参数形式找到样本均值和样本方差的分布。可能更令人惊讶的是, 我们可以证明样本均值和样本方差是独立的。

定理 5.9 (正态总体下样本均值与方差的独立性) 假设 Y_1, \dots, Y_n 是来自一个正态总体的随机样本, 即 $Y \sim N(\mu, \sigma^2)$ 。那么,

- i. 样本均值 \bar{Y} 与形如 $(Y_j - \bar{Y})$ 的项对于 $j = 1, \dots, n$ 是独立的,
- ii. 样本均值与样本方差 S^2 是独立的。

定理 5.10 (标准正态下样本均值与方差的分布) 假设 Z_1, \dots, Z_n 是来自一个标准正态分布总体的随机样本，即 $Z \sim N(0, 1)$ 。如果我们定义 $\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i$ 和 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z})^2$ ，那么

- i. 样本均值服从正态分布， $\bar{Z} \sim N(0, \frac{1}{n})$ ，
- ii. $(n-1)$ 倍的样本方差服从卡方分布， $(n-1)S^2 \sim \chi_{n-1}^2$ 。

定理 5.11 (一般正态下样本均值与方差的分布) 假设 Y_1, \dots, Y_n 是来自一个均值为 μ 、方差为 σ^2 的正态总体的随机样本，即 $Y \sim N(\mu, \sigma^2)$ 。那么

- i. $\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$
- ii. $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$

5.9 Sample quantiles and order statistics

均值和方差并不是衡量集中趋势和离散程度的唯一指标。通常引用的替代指标包括中位数和四分位距，这两者都是基于分位数计算的。基于分位数的统计量通常比基于矩的统计量更具稳健性；换句话说，基于分位数的统计量对极端观测值不那么敏感。

定义 5.13 (次序统计量) 对于样本 Y_1, \dots, Y_n ，其次序统计量 $Y_{(1)}, \dots, Y_{(n)}$ 是将样本值按升序排列后的结果。因此， $Y_{(i)}$ 是我们样本中第 i 小的值。

定义 5.14 (样本分位数) 假设 Y_1, \dots, Y_n 是一个样本，并令 Q_α 表示样本 α 分位数。我们分两部分定义 Q_α ；首先定义 $\alpha = 1/2$ 的情况，然后定义 $\alpha \neq 1/2$ 的情况。

$$Q_{0.5} = \begin{cases} Y_{((n+1)/2)} & \text{若 } n \text{ 为奇数} \\ \frac{1}{2}(Y_{(n/2)} + Y_{((n/2)+1)}) & \text{若 } n \text{ 为偶数} \end{cases}$$

$$Q_\alpha = \begin{cases} Y_{((n\alpha))} & \text{若 } \frac{1}{2n} < \alpha < \frac{1}{2} \\ Y_{(n+1-(n(1-\alpha)))} & \text{若 } \frac{1}{2} < \alpha < 1 - \frac{1}{2n} \end{cases}$$

此处 $\{k\}$ 表示按通常方式将 k 四舍五入到最接近的整数。

5.9.1 Sample minimum and sample maximum

$$Y_{(1)} = \min_{1 \leq i \leq n} (Y_1, \dots, Y_n)$$

$$Y_{(n)} = \max_{1 \leq i \leq n} (Y_1, \dots, Y_n)$$

定理 5.12 (样本最小值与最大值的分布函数) 假设 Y_1, \dots, Y_n 是来自一个累积分布函数为 F_Y 的总体的随机样本，即 $Y \sim F_Y$ 。那么样本最小值和样本最大值的累积分布函数由下式给出：

$$F_{Y_{(1)}}(y) = P(Y_{(1)} \leq y) = 1 - [1 - F_Y(y)]^n$$

$$F_{Y_{(n)}}(y) = P(Y_{(n)} \leq y) = [F_Y(y)]^n$$

5.9.2 Distribution of i^{th} order statistic

定理 5.13 假设 Y_1, \dots, Y_n 是来自一个累积分布函数为 F_Y 的总体的随机样本，即 $Y \sim F_Y$ 。那么第 i 个次序统计量的累积分布函数由下式给出：

$$F_{Y_{(i)}}(y) = \sum_{j=i}^n \binom{n}{j} F_Y(y)^j (1 - F_Y(y))^{n-j}$$

定理 5.14 (第 i 个次序统计量的密度函数) 假设 Y_1, \dots, Y_n 是来自一个累积分布函数为 F_Y 、密度函数为 f_Y 的总体的随机样本。第 i 个次序统计量的密度函数由下式给出：

$$f_{Y_{(i)}}(y) = \frac{n!}{(i-1)!(n-i)!} f_Y(y) [F_Y(y)]^{i-1} [1 - F_Y(y)]^{n-i}$$

$f_{Y_{(i)}}$ 的支撑集与 f_Y 的支撑集相同。

定理 5.15 (次序统计量的联合密度函数) 假设 Y_1, \dots, Y_n 是来自一个累积分布函数为 F_Y 、密度函数为 f_Y 的总体的随机样本。第 i 个和第 j 个次序统计量的联合密度函数由下式给出：

$$f_{Y_{(i)}, Y_{(j)}}(v, w) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} f_Y(v) f_Y(w) [F_Y(v)]^{i-1} [F_Y(w) - F_Y(v)]^{j-i-1} [1 - F_Y(w)]^{n-j}$$

其支撑集是满足 $v < w, f_Y(v) > 0$ 且 $f_Y(w) > 0$ 的区域。

5.10 Further exercises

5.11 Appendix: Proofs

6 Estimation, testing, and prediction 估计, 检验和预测

6.1 Point estimation

考虑一个标量参数 θ 。任何基于大小为 n 的样本来估计 θ 值的方法都可以表示为一个函数 $h : \mathbb{R}^n \rightarrow \mathbb{R}$ 。当将此函数应用于观测样本时, 它产生一个点估计值 (**point estimate**), $h(y)$ 。这个估计值只是一个数字。为了深入了解估计方法的性质, 我们考虑将该函数应用于样本。由此产生的随机变量 $h(Y)$ 被称为点估计量 (**point estimator**)。

显然, 任何点估计量都是一个统计量。事实上, 这种关联是双向的。估计量显然是一个统计量, 而且是作用很明显的统计量 (可以用来估计分布的未知参数)。但反过来看, 任何一个统计量我们都可以看作是某个参数的估计量, 只是具体是哪个参数的问题, 以及估计的偏误有多少的问题。

定义 6.1 (点估计量) 任何标量统计量都可以被视为参数 θ 的一个点估计量。该统计量的观测值被称为点估计值。

在下表中, 我们列出了一些常用的统计量、它们的观测样本值以及它们用作点估计量的总体特征。

	统计量 (U)	观测值 (u)	估计目标
样本均值	$\bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j$	$\bar{y} = \frac{1}{n} \sum_{j=1}^n y_j$	总体均值
样本方差	$S^2 = \frac{1}{n-1} \sum_{j=1}^n (Y_j - \bar{Y})^2$	$s^2 = \frac{1}{n-1} \sum_{j=1}^n (y_j - \bar{y})^2$	总体方差
第一次序统计量	$Y_{(1)} = \min_{1 \leq j \leq n} Y_j$	$y_{(1)} = \min_{1 \leq j \leq n} y_j$	总体最小值
第 n 次序统计量	$Y_{(n)} = \max_{1 \leq j \leq n} Y_j$	$y_{(n)} = \max_{1 \leq j \leq n} y_j$	总体最大值
样本中位数	$Y_{((n+1)/2)}$	$y_{((n+1)/2)}$	总体中位数

通常, 我们可能不止关心一个参数。这些参数被组合成一个向量 $\theta = (\theta_1, \dots, \theta_r)^T$ 。这个向量既可以视为单个向量参数, 也可以视为标量参数的向量。用作 θ 的点估计量的统计量是一个 $r \times 1$ 随机向量, 相应的点估计值将是一个 $r \times 1$ 的实数向量 (\mathbb{R}^r 中的一个点)。 θ 的方向通常并不重要; 行向量和列向量同样能满足我们的目的。在这些情况下, 我们可以省略转置, 写作 $\theta = (\theta_1, \dots, \theta_r)$ 。

6.1.1 Bias, variance, and mean squared error

定义 6.2 (偏差) 假设 U 是一个统计量。 U 作为 θ 的点估计量的偏差 (**Bias**) 为

$$Bias(U) = \mathbb{E}_\theta[U - \theta]$$

如果 $Bias(U) = 0$, 即 $\mathbb{E}_\theta[U] = \theta$, 我们称 U 是 θ 的一个无偏估计量。

定义 6.3 (均方误差) 假设 U 是一个统计量。 U 作为 θ 的估计量的均方误差 (**Mean Squared Error, MSE**) 由下式给出:

$$MSE(U) = \mathbb{E}_\theta[(U - \theta)^2]$$

定理 6.1 (MSE、偏差与方差的关系) 假设 U 是一个统计量, 则对于所有 $\theta \in \Theta$,

$$MSE(U) = [Bias(U)]^2 + \text{Var}_\theta(U)$$

对于一个无偏估计量, 其均方误差等于其方差。

该定理有一个直观的解释：一个具有低偏差和低方差的估计量是吸引人的。低方差意味着概率质量集中在分布的中心附近，而考虑到我们同时具有低偏差，该分布的中心又接近参数的真值 θ 。总之，一个具有低均方误差的估计量，其分布的概率质量将集中在 θ 附近。在实践中，低偏差和低方差常常是相互竞争的要求；为了减少偏差，我们可能不得不接受更高的方差，而为了获得一个低方差的估计量，我们可能不得不引入一些偏差。这被称为**偏差-方差权衡 (Bias-Variance Tradeoff)**。

定理 6.2 (样本均值与样本方差的 MSE)

$$MSE(\bar{Y}) = \frac{\sigma^2}{n}$$

$$MSE(S^2) = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \sigma^4 \right)$$

6.1.2 Consistency

定义 6.4 (一致性) 估计量序列 $\{U_n : n = 1, 2, \dots\}$ 是参数 θ 的一个一致 (*Consistent*) 估计量序列，如果它依概率收敛于 θ ，即对于每个 $\delta > 0$ 和 $\theta \in \Theta$ ，有

$$\lim_{n \rightarrow \infty} P_\theta(|U_n - \theta| < \delta) = 1$$

一种可能的直观解释如下。假设 $\{U_n\}$ 是 θ 的一致估计量。这意味着，对于我们选择的参数值周围的任何开区间，例如 $(\theta - \delta, \theta + \delta)$ ，我们都能找到一个数 N ，使得只要样本量大于 N ，我们的估计量取值落在该区间之外的概率就任意小。所谓任意小，是指如果我们想要降低估计量取值落在区间外的概率，我们只需要增加 N 的值即可。

定理 6.3 (MSE 与一致性的关系) 假设 $\{U_n\}$ 是一个统计量序列， θ 是一个未知参数。如果对于所有 $\theta \in \Theta$ ，有

$$\lim_{n \rightarrow \infty} MSE(U_n) = 0$$

那么 $\{U_n\}$ 是 θ 的一个一致估计量序列。

定理 6.4 (一致性、偏差与方差的关系) 假设 $\{U_n\}$ 是一个统计量序列， θ 是一个未知参数。如果对于所有 $\theta \in \Theta$ ，有

$$\lim_{n \rightarrow \infty} \text{Bias}(U_n) = 0 \quad \text{且} \quad \lim_{n \rightarrow \infty} \text{Var}(U_n) = 0$$

那么 $\{U_n\}$ 是 θ 的一个一致估计量序列。

定理 6.5 (弱大数定律) 如果 $Y = (Y_1, \dots, Y_n)^T$ 是一个随机样本，且 $E[Y] = \mu < \infty$, $\text{Var}(Y) = \sigma^2 < \infty$ ，那么样本均值是总体均值的一个一致估计量，即

$$\bar{Y}_n \xrightarrow{P} \mu \quad \text{当} \quad n \rightarrow \infty$$

如果我们做一些稍强的假设，我们也可以证明样本均值几乎必然收敛于总体均值。这个结果被称为**强大数定律 (Strong Law of Large Numbers)**。此外，我们可以推断，只要四阶中心矩是有限的，样本方差就是总体方差的一个一致估计量。

6.1.3 The method of moments (MM)

通常，总体矩是我们想要估计的参数的函数。样本矩是统计量，即样本的函数。通过令总体矩等于样本矩，并求解由此产生的联立方程组，我们可以生成总体参数的估计量。这些估计量被称为矩估计量。

这应该是最直接能想到的用样本统计量估计总体参数的方式：用对应的样本矩去估计总体矩，例如总体的一阶原点矩期望就用样本的一阶原点矩样本均值来估计。当然，这样做显然未必能保证是无偏的、最小方差的，很容易想到如果用样本的二阶中心矩来估计总体的二阶中心矩也就是方差，那么这个估计量是有偏的。（别忘了样本方差 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$ 是方差的一个无偏估计量，但是样本的二阶中心矩是 $\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2$ ！）另外我们想估计的总体未知参数未必是总体矩本身，但只要总体矩里包含参数，我们就可以通过联立方程来解出未知参数的矩估计量。

矩估计的合理性来自大数定律。大数定律有很多版本，我们讨论过弱大数定律 (WLLN)，它说的是样本均值是总体均值的一个一致估计量，以及强大数定律。但其实大数定律也可以推广到任意阶矩，即：如果总体矩存在（有限），那么相应的样本矩会依概率收敛到总体矩，

$$\frac{1}{n} \sum_{i=1}^n X_i^k \xrightarrow{P} \mathbb{E}[X^k] \quad (\text{当 } n \rightarrow \infty)$$

这是 Marcinkiewicz-Zygmund 大数定律。这个定律在说，只要总体矩存在，任意高阶样本矩都是它们对应总体矩的一致估计量。但注意，虽然一阶样本原点矩作为其总体矩的估计量时是无偏的，但无偏性并不成立于任意阶，一个直接的反例就是我们刚才说过的二阶中心矩。但注意，对于原点矩来说，样本原点矩不仅是总体原点矩的一致估计量还一定是无偏估计量，而样本中心矩可能是有偏的，比如样本方差就需要修正，但也一定是一致的（相较于原点矩，中心矩涉及 \bar{X} ，但大数定律仍然适用，因为 \bar{X} 也是一致估计量）。

$$\begin{aligned} \mathbb{E}[M_k] &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n X_i^k\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i^k] = \mathbb{E}[X^k] \\ \mathbb{E}[M_k^c] &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(X_i - \bar{X})^k] \neq \mathbb{E}[(X - \mathbb{E}[X])^k] \end{aligned}$$

这里的问题是 $X_i - \bar{X}$ 不是独立的，会引入偏差。

考虑一个由参数 $\theta = (\theta_1, \dots, \theta_r)^T$ 参数化的样本 $Y = (Y_1, \dots, Y_n)^T$ 。 $\theta_1, \dots, \theta_r$ 的矩估计量将通过求解以下方程组生成：

$$\mu'_i(\hat{\theta}_1, \dots, \hat{\theta}_r) = m'_i(Y_1, \dots, Y_n) \quad \text{对于 } i = 1, \dots, r$$

从而得到用 Y_1, \dots, Y_n 表示的 $\hat{\theta}_1, \dots, \hat{\theta}_r$ 的表达式。这里， $\mu'_i(\hat{\theta}_1, \dots, \hat{\theta}_r)$ 是第 i 阶总体矩，而 $m'_i(Y_1, \dots, Y_n)$ 是第 i 阶样本矩。

矩估计法的优点是易于实现。然而，由此产生的估计量通常具有不理想的特性。除了非常简单的案例之外，矩估计法的主要用途是为其他估计程序提供初始值。

定理 6.6 (正态分布的矩估计) 假设 $Y = (Y_1, \dots, Y_n)^T$ 是来自 $N(\mu, \sigma^2)$ 分布的样本，其中 μ 和 σ^2 都是未知参数。矩估计量通过求解以下方程得到：

$$\mu'_1(\hat{\mu}, \hat{\sigma}^2) = m'_1(Y_1, \dots, Y_n)$$

$$\mu'_2(\hat{\mu}, \hat{\sigma}^2) = m'_2(Y_1, \dots, Y_n)$$

这得到：

$$\hat{\mu} = \bar{Y}$$

$$\hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n} \sum_{i=1}^n Y_i^2 \implies \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

注意，方差的矩估计量不是样本方差 S^2 。如果 $\hat{\sigma}^2$ 是矩估计量，那么

$$\hat{\sigma}^2 = \frac{n-1}{n} S^2$$

$\hat{\sigma}^2$ 是总体方差的一个有偏估计量。

6.1.4 Ordinary least squares (OLS)

我们现在考虑一种最常用于线性回归框架的估计技术。

$$Y = X\beta + \sigma\varepsilon$$

其中 Y 是一个 $n \times 1$ 响应向量， X 是一个 $n \times (p+1)$ 的解释变量矩阵（包含常数项）， β 是一个 $(p+1) \times 1$ 参数向量，而 ε 是一个 $n \times 1$ 误差向量。给定 β 的一个估计值 b ，响应变量第 i 个观测值的估计值为

$$\hat{Y}_i = x_i^T b$$

其中 x_i^T 是 X 的第 i 行。在普通最小二乘 (OLS) 估计中，我们寻求最小化误差平方和，

$$S(b) = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = (Y - Xb)^T (Y - Xb)$$

β 的最小二乘估计量于是为

$$\hat{\beta} = \arg \min_b S(b)$$

这有一个简单的几何解释：在简单回归情况下，我们寻求最小化每个点 (X_i, Y_i) 与回归线之间的垂直距离的平方。在更高维度中，这被替换为 \mathbb{R}^{p+1} 中一个点与一个 p 维超平面之间的距离。

我们通过令 $S(b)$ 关于 b 的导数为零来计算 $\hat{\beta}$ 。我们需要一些矩阵微积分的标准结果；如果 A 是一个 $r \times c$ 常数矩阵， w 是一个 $c \times 1$ 向量，我们有

$$\frac{\partial Aw}{\partial w} = A, \quad \frac{\partial w^T A}{\partial w} = A^T, \quad \text{且} \quad \frac{\partial w^T Aw}{\partial w} = w^T (A + A^T)$$

$S(b)$ 关于 b 的导数于是为

$$\begin{aligned} \frac{\partial S(b)}{\partial b} &= \frac{\partial}{\partial b} (Y - Xb)^T (Y - Xb) = \frac{\partial}{\partial b} (Y^T Y - Y^T Xb - b^T X^T Y + b^T X^T Xb) = \\ &0 - Y^T X - (X^T Y)^T + b^T [X^T X + (X^T X)^T] = 2(-Y^T X + b^T X^T X) \end{aligned}$$

由于 $X^T X$ 是对称的。令导数为零得到

$$\hat{\beta}^T X^T X = Y^T X \Leftrightarrow X^T X \hat{\beta} = X^T Y \Leftrightarrow \hat{\beta} = (X^T X)^{-1} X^T Y$$

假设矩阵 $X^T X$ 是满秩的且可逆。二阶导数为

$$\frac{\partial^2 S(b)}{\partial b^T \partial b} = \frac{\partial}{\partial b^T} 2(-Y^T X + b^T X^T X) = 2X^T X$$

这是一个正定矩阵；对于向量值函数，这等价于二阶导数为正，因此我们得出结论 $\hat{\beta}$ 最小化了 $S(b)$ 。

注意 $\hat{\beta}$ 是 Y 的线性函数。我们可以写

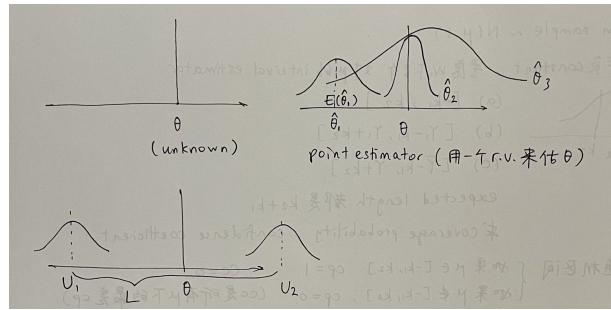
$$\hat{Y} = X \hat{\beta} = X(X^T X)^{-1} X^T Y = HY$$

其中 $H = X(X^T X)^{-1} X^T$ 被称为帽子矩阵 (hat matrix)；它将观测值 Y 映射到估计值 \hat{Y} 。帽子矩阵是幂等 (idempotent) 的，即 $H^2 = H$ 。

6.2 Interval estimation

在上一节中，我们描述了点估计量，它为感兴趣的参数提供一个单一的值。另一类重要的推断方法是区间估计量 (**interval estimators**)。顾名思义，区间估计量为我们的未知参数提供一个可能取值的范围，而不仅仅是一个点。区间估计被广泛使用，但也常被误解。区间估计量可以看作是置信集 (**confidence set**) 的一个特例。

区间估计量是一个随机区间；该区间的端点是统计量。当我们用观测到的样本替换样本时，我们最终得到一个区间估计，它只是实数轴上的一个区间。



定义 6.5 (区间估计量) 假设我们有一个由 θ 参数化的样本 Y 。令 $U_1 = h_1(Y)$ 和 $U_2 = h_2(Y)$ 为满足 $U_1 \leq U_2$ 的统计量。随机区间 $[U_1, U_2]$ 是 θ 的一个区间估计量。如果观测样本为 y ，则统计量的观测值为 $u_1 = h_1(y)$ 和 $u_2 = h_2(y)$ 。区间 $[u_1, u_2]$ 是 θ 的一个区间估计值。

一个明显的问题是什么构成了一个好的区间估计量。我们必须平衡两个相互竞争的要求：我们希望我们的区间估计尽可能窄，但我们也希望它覆盖真实值的概率很高。区间长度和覆盖概率之间的平衡将在后面讨论。

6.2.1 Coverage probability and length

定义 6.6 (覆盖概率) 对于 θ 的区间估计量 $[U_1, U_2]$ ，其覆盖概率是区间估计量覆盖 θ 的概率，即 $P(U_1 \leq \theta \leq U_2)$ 。

定义 6.7 (置信系数) 对于 θ 的区间估计量 $[U_1, U_2]$ ，其置信系数是覆盖概率关于 θ 的下确界，即 $\inf_{\theta \in \Theta} P(U_1 \leq \theta \leq U_2)$ 。实际应用中，覆盖概率可能依赖未知的 θ （例如， t 分布区间依赖于方差）。取所有 θ 情况下的最差覆盖概率（下确界），能确保无论如何，置信系数是最保守的可靠性度量。

重要的是要清楚，在 $P(U_1 \leq \theta \leq U_2)$ 中的随机变量是 U_1 和 U_2 ，因此，

$$P(U_1 \leq \theta \leq U_2) = P[(U_1 \leq \theta) \cap (U_2 \geq \theta)] = 1 - P(U_1 > \theta) - P(U_2 < \theta)$$

你经常会看到置信系数为 $(1 - \alpha)$ 的区间估计量被称为 $100(1 - \alpha)\%$ 置信区间。例如，如果 $\alpha = 0.05$ ，那么得到的区间估计量通常被称为 95% 置信区间。在这种情况下，置信系数（以百分比表示）通常被称为置信水平。我们互换使用置信区间和区间估计量这两个术语。

定义 6.8 (期望长度) 考虑一个区间估计量 $[U_1, U_2]$ 。该区间的期望长度定义为 $\mathbb{E}[U_2 - U_1]$ 。

区间估计量的一个理想特性是对于所有 θ 值，覆盖概率都很大。置信系数代表了最坏情况；根据定义，对于任何 θ 值，覆盖概率至少与置信系数一样大。

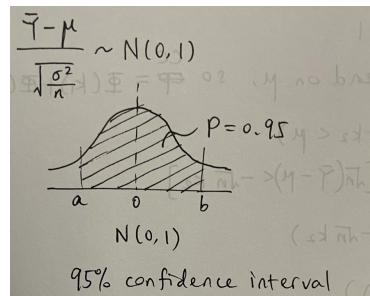
$$\left\{ \begin{array}{l} \text{期望区间长度: } \mathbb{E}[U_2 - U_1] = \mathbb{E}[U_2] - \mathbb{E}[U_1] \\ \text{覆盖概率: } P(U_1 \leq \theta \leq U_2) = P[(U_1 \leq \theta) \cap (U_2 \geq \theta)] = 1 - P(U_1 > \theta) - P(U_2 < \theta) \end{array} \right.$$

假设我们有一个来自 $N(\mu, \sigma^2)$ 分布的随机样本 $Y = (Y_1, \dots, Y_n)^T$ 。我们关心的是在假设 σ^2 已知的情况下， μ 的区间估计量。我们知道样本均值也服从正态分布，

$$\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$$

因此，

$$\frac{\bar{Y} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0, 1)$$



$$P \left(a \leq \frac{\bar{Y} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq b \right) = 0.95$$

$$P \left(\bar{Y} - b\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{Y} - a\sqrt{\frac{\sigma^2}{n}} \right) = 0.95$$

μ 的一个合理的区间估计量是

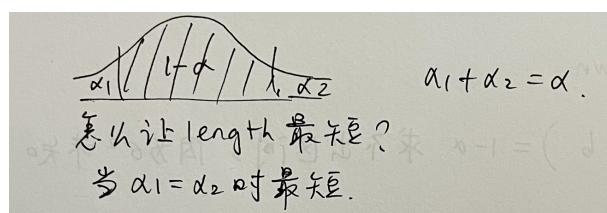
$$\left[\bar{Y} - b\sqrt{\frac{\sigma^2}{n}}, \bar{Y} - a\sqrt{\frac{\sigma^2}{n}} \right]$$

这个随机区间里含 μ 的概率是 0.95。

对应的区间估计值是

$$\left[\bar{y} - b\sqrt{\frac{\sigma^2}{n}}, \bar{y} - a\sqrt{\frac{\sigma^2}{n}} \right]$$

在实践中，我们会固定一个置信水平，并找到对应的最小区间长度。



6.2.2 Constructing interval estimators using pivotal functions

枢轴函数为生成具有给定置信系数的区间估计量提供了一个简单的机制。假设我们想要一个置信系数为 $1 - \alpha$ 的 θ 的区间估计量。我们可以使用以下步骤：

1. 找到一个枢轴函数 $g(Y, \theta)$ 。
2. 找到 w_1, w_2 使得 $P(w_1 \leq g(Y, \theta) \leq w_2) = 1 - \alpha$ 。
3. $P(\dots \leq \theta \leq \dots) = 1 - \alpha$ 。

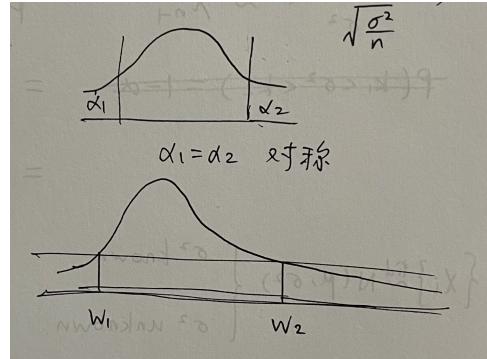
注意这个步骤给到我们一个符合置信系数条件的估计量，但未必是最短长度。

如果 $W = g(Y, \theta)$ 是 θ 的线性函数，且是单峰的（比如之前的 $\frac{\bar{Y} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$ ），我们想要 $\min(w_2 - w_1)$ ：

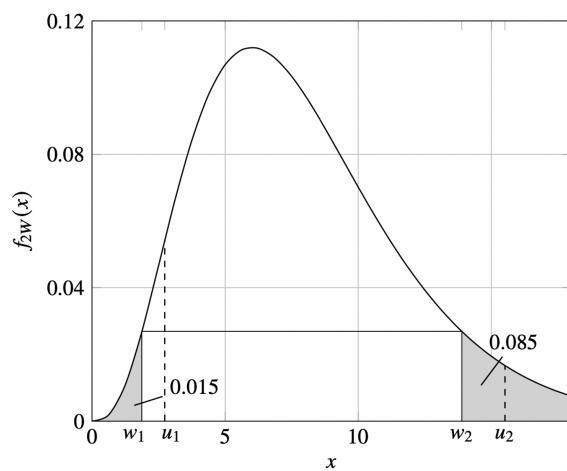
当 $f_W(w_1) = f_W(w_2)$ 时，得到最短长度。

如果 f_W 还是对称的，那么得到的估计量是一个等尾区间，即满足：

$$P(W < w_1) = P(W > w_2) = \frac{\alpha}{2}.$$



定义 6.9 (单峰性) 设 X 是一个具有质量/密度函数 $f_X(x)$ 和支撑集 D 的随机变量。如果存在一个值 x^* ，使得 $f(x)$ 在 $x < x^*$ 时（严格）递增，在 $x > x^*$ 时（严格）递减，其中 $x \in D$ ，则称 X 是（严格）单峰的。值 x^* 是分布的众数。



6.2.3 Constructing interval estimators using order statistics

在目前我们考虑过的例子中，我们已经给出了来自一些常见分布的均值、方差及其他特定参数的置信区间。另一类有用的总体参数是分位数。在前文中，我们展示了顺序统计量（order statistics）

	总体分布已知正态： $\{X_i\} \sim iid N(\mu, \sigma^2)$	σ^2 已知	枢轴量用 $\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$	
		σ^2 未知	枢轴量用 $\frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}} \sim t_{n-1}$ (如果是大样本, $\frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}} \xrightarrow{d} N(0,1)$, 因为 $S^2 \xrightarrow{p} \sigma^2$, 按 $N(0,1)$ 得到的是一个近似的置信区间。)	
总体均值 μ	总体分布未知： $\{X_i\} \sim iid (\mu, \sigma^2)$	大样本 $(\bar{X} \xrightarrow{d} N(\mu, \frac{\sigma^2}{n}))$	σ^2 已知	$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow{d} N(0,1)$
			σ^2 未知	$\frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}} \xrightarrow{p} \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow{d} N(0,1)$
		小样本	需借助参数方法 (如 Bootstrap) 或假设分布形式	
总体方差 σ^2	总体分布已知正态	μ 已知	枢轴量用 $\frac{\sum(X_i - \mu)^2}{\sigma^2} \sim \chi^2(n)$	
		μ 未知	枢轴量用 $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$	
	总体分布未知			

为关于总体分位数的推断提供了一个自然的机制。因此，分位数的区间估计量也基于顺序统计量就不足为奇了。

我们首先建立基于第一个和最后一个顺序统计量的区间作为一个标量参数 θ 的区间估计量时所关联的覆盖概率 (coverage probability)。

定理 6.7 (基于首末顺序统计量的区间估计量) 考虑来自一个具有标量参数 θ 的分布的随机样本 $Y = (Y_1, \dots, Y_n)^T$ 。令 F_Y 为 Y 的累积分布函数, $Y_{(i)}$ 为第 i 个顺序统计量。随机区间 $[Y_{(1)}, Y_{(n)}]$ 覆盖 θ 的概率为

$$1 - [1 - F_Y(\theta)]^n - [F_Y(\theta)]^n.$$

假设感兴趣的参数是 β -分位数, 即 $\theta = q_\beta$, 其中 q_β 是满足 $F_Y(q_\beta) = \beta$ 的点。那么由上述定理给出的覆盖概率就是一个置信系数。下面的推论给出了细节。

定理 6.8 ($[Y_{(1)}, Y_{(n)}]$ 作为分位数的区间估计量) 考虑一个随机样本 $Y = (Y_1, \dots, Y_n)^T$ 。令 q_β 为 Y 的 β -分位数。随机区间 $[Y_{(1)}, Y_{(n)}]$ 是 q_β 的一个区间估计量, 其置信系数为

$$1 - (1 - \beta)^n - \beta^n.$$

在实践中, 我们希望能够为任意给定的置信系数提供 q_β 的区间估计量。通过考虑与一般顺序统计量 $Y_{(i)}$ 和 $Y_{(j)}$ (其中 $i < j$) 相关联的区间估计量, 我们获得了更大的灵活性。下面的定理说明了如何计算覆盖概率。

定理 6.9 ($[Y_{(i)}, Y_{(j)}]$ 作为区间估计量) 考虑来自一个具有标量参数 θ 的分布的随机样本 $Y = (Y_1, \dots, Y_n)^T$ 。令 $F_{Y_{(i)}}$ 为第 i^{th} 个顺序统计量的累积分布函数。如果 $i < j$, 则随机区间 $[Y_{(i)}, Y_{(j)}]$ 覆盖 θ 的概率为

$$F_{Y_{(i)}}(\theta) - F_{Y_{(j)}}(\theta).$$

6.2.4 Confidence sets

在我们的大部分例子中, 我们将构建置信区间。然而, 重要的是要理解, 置信区间可以被视为置信集 (confidence set) 的一个特例。置信集在两种情况下非常有用:

- i. 如果我们不确定某个程序的结果是否是一个区间,

ii. 如果我们有一个参数向量，在这种情况下，我们可能将我们的置信集称为置信区域（confidence region）。

一个对于向量参数 $\theta \in \Theta$ 具有置信系数 $1 - \alpha$ 的置信集定义为随机集 $C(Y) \subseteq \Theta$ ，其中

$$P(\theta \in C(Y)) = 1 - \alpha.$$

请注意，在上述表达式中，变量出现的顺序可能引起混淆；这里的随机变量是 $C(Y)$ （它是样本的一个函数）。在这种情况下，对于一个观测到的样本 y ，我们将得到观测到的置信集 $C(y)$ 。

6.3 Hypothesis testing

假设检验始于一个关于总体的陈述；该陈述被称为原假设（null hypothesis）。假设检验判断样本是否提供了拒绝原假设的证据。做出此推断的基础是，将检验统计量（test statistic）的观测值与我们假设原假设为真时该统计量的抽样分布进行比较。一个完美的假设检验应当总是拒绝错误的原假设，并且从不拒绝正确的原假设。

在实践中，我们必须在拒绝一个错误假设的概率与拒绝一个正确假设的概率之间取得平衡。

假设检验的形式化设置受到了广泛的批评，并且在现代统计学中有些过时。然而，即使没有明确采用其精确的形式，这里所提出的思想仍然被广泛使用。新闻网站上充满了关于食品安全恐慌、考试成绩、疾病原因等的报道。这些报道中包含诸如“没有证据表明对健康存在风险”、“表现的变化具有统计学意义”、或“因素 x 与疾病 y 相关”等形式的陈述。这些陈述植根于假设检验的思想，并 loosely 地借鉴了其术语。即使我们对形式化检验方法的价值存疑，正确理解这些方法及其局限性也至关重要。专业统计学家的一项重要职责就是警示对假设检验的幼稚解读（并识别严重的误解）。

6.3.1 Statistical hypotheses

本节我们考虑的假设检验是参数检验（parametric），即假设是关于分布参数值的陈述；而分布本身的形式被假定为已知。相比之下，非参数检验（nonparametric）则不需要对分布形式做任何假定。我们在本书前文已经遇到过一些非参数技术，例如前文中基于分位数的置信区间。

通常，一个参数假设检验的形式为

$$H_* : \theta \in \Theta_*, \text{ 其中 } \Theta_* \subset \Theta.$$

这里的 H_* 只是我们为假设所取的一个方便标签。在陈述 $\theta \in \Theta_*$ （其中 $\Theta_* \subset \Theta$ ）时，我们是在提议真实的参数值 θ 位于参数空间 Θ 的一个特定子集 Θ_* 中。对于标量参数 θ 和已知常数 k ，我们可能考虑的假设形式包括 $\theta = k$ 、 $\theta < k$ 、 $\theta \leq k$ 、 $\theta > k$ 以及 $\theta \geq k$ 。请注意，与上述形式一致， $\theta = k$ 可以写作 $\theta \in \{k\}$ ， $\theta < k$ 可以写作 $\theta \in (-\infty, k)$ ， $\theta \leq k$ 可以写作 $\theta \in (-\infty, k]$ ，依此类推。在这个简短的说明中，我们实际上描述了两种不同形式的假设。下面的定义予以阐明。

定义 6.10 (简单假设与复合假设) 一个简单假设（simple hypothesis）给出了参数值 θ 的精确规定。因此，一个简单假设将采取 $\theta = k$ 的形式，其中 k 是一个与 θ 维度相同的已知常数向量。任何非简单的假设则被称为复合假设（composite hypothesis）。

在标量参数的情况下，显然 $\theta = k$ 是一个简单假设。其他四个例子并未给出 θ 的具体值，因此 $\theta < k$ 、 $\theta \leq k$ 、 $\theta > k$ 和 $\theta \geq k$ 都是复合假设。通常，一个简单假设会提议 θ 是一个单元素集合的成员，即 $\theta \in \{k\}$ 。而一个复合假设则提议 θ 是一个多于一个元素的集合的成员，即 $\theta \in \Theta_*$ 且

$|\Theta_*| > 1$ 。 Θ_* 的基数可能是有限的，例如 $\theta \in \{k_1, k_2\}$ 或 $\theta \in \{k_1, \dots, k_n\}$ ；可能是可数无限的，例如 $\theta \in \{k_1, k_2, \dots\}$ ；也可能是不可数无限的，例如 $\theta \in (k_1, k_2)$ 、 $\theta \in (-\infty, k]$ 或 $\theta \in \{x : |x| \leq k\}$ 。

在经典假设检验中，会提出两个假设。**原假设** (null hypothesis) 是我们想要检验的假设；**备择假设** (alternative hypothesis) 则指导我们构建检验原假设的方式。

1. 原假设通常是保守的，体现在它反映了既有的观点。原假设常采取“无变化”、“无差异”或“无效应”的形式。我们将遵循惯例，将原假设标记为 H_0 。

2. 备择假设反映了我们对参数可能取值的怀疑。备择假设通常是复合的，并且常采取可以解释为“存在变化”、“存在差异”或“存在效应”的形式。我们将使用 H_1 来标记备择假设。

现在我们可以给出一个一般假设检验的正式陈述。对于一个参数 θ ，假设检验的形式为

$$H_0 : \theta \in \Theta_0,$$

$$H_1 : \theta \in \Theta_1.$$

其中 $\Theta_0 \subset \Theta$, $\Theta_1 \subset \Theta$, 且 $\Theta_0 \cap \Theta_1 = \emptyset$ 。请注意，我们坚持 Θ_0 和 Θ_1 是互斥的，因此 H_0 和 H_1 不可能同时为真。然而，并不要求这些集合是穷尽的，也就是说，我们不要求 $\Theta_0 \cup \Theta_1 = \Theta$ 。

6.3.2 Decision rules

经典假设检验涉及一个二元选择。我们可用的选项是：

- a. 拒绝原假设,
- b. 不拒绝原假设。

统计学教师喜欢指出，“不拒绝原假设”的陈述不应被视为等同于“接受原假设”（而学生们也同样喜欢忽略这个建议）。这虽然有点学究气，但重要的是要认识到，统计程序所能做的只是检查现有证据是否与原假设一致。同理，“拒绝原假设”也不应被解释为“接受备择假设”。

我们用来决定是否拒绝原假设的证据是观测到的样本 $y = (y_1, \dots, y_n)^T$ 。样本被用来构建一个**检验统计量** (test statistic)，而关于 H_0 的决定就是基于这个统计量做出的。目前，我们将用一个单一的函数 ϕ 来总结这个决策过程，其中

$$\phi(y) = \begin{cases} 1 & \text{当拒绝 } H_0 \text{ 时,} \\ 0 & \text{当不拒绝 } H_0 \text{ 时.} \end{cases}$$

函数 ϕ 被称为**决策规则** (decision rule)。 ϕ 的作用是将可能观测到的样本值空间划分为两个集合： $R = \{y : \phi(y) = 1\}$ 和 $R^c = \{y : \phi(y) = 0\}$ 。集合 R 对应于那些将导致 H_0 被拒绝的观测样本值，通常被称为**拒绝域** (rejection region) 或**临界区域** (critical region)。对于给定的检验，决策函数和临界区域是等价的，因为只要知道其中一个，另一个就完全确定了。如果指定了临界区域 R ，我们可以写出决策函数：

$$\phi(y) = \begin{cases} 1 & \text{若 } y \in R, \\ 0 & \text{若 } y \notin R. \end{cases}$$

类似地，对于特定的决策函数 ϕ ，临界区域由下式给出：

$$R = \{y \in \mathbb{R}^n : \phi(y) = 1\}.$$

6.3.3 Types of error and the power function

在任何非平凡的决策问题中，我们都可能做出错误的决定。在假设检验中，有两种可能的错误决定。通常表述为：

- a. 第一类错误：当 H_0 为真时拒绝 H_0 ，
- b. 第二类错误：当 H_0 为假时未能拒绝 H_0 。

评估检验性能的一个可能机制是考虑检验犯错误的概率。第一类错误的概率与检验的**显著性水平** (significance level) 和**检验水平** (size) 相关。这些术语经常互换使用，尽管它们的含义不同，如下述定义所示。

定义 6.11 (显著性水平与检验水平) 考虑检验原假设 $H_0 : \theta \in \Theta_0$ 。若满足

$$\sup_{\theta \in \Theta_0} P_\theta(H_0 \text{ 被拒绝}) \leq \alpha,$$

则该检验具有显著性水平 α 。若满足

$$\sup_{\theta \in \Theta_0} P_\theta(H_0 \text{ 被拒绝}) = \alpha,$$

则该检验具有检验水平 α 。

两者之间的区别相当微妙，并且在我们考虑的最初例子中并不重要，因为检验水平可以很容易地计算出来。然而，在一些更复杂的情况下，检验水平不易计算，我们不得不满足于使用显著性水平。以下是关于检验水平与显著性水平的几点说明：

1. 一个检验水平为 α 的检验，其显著性水平也是 α 。反之则不成立。
2. 一个显著性水平为 α 的检验，总是至少与一个检验水平为 α 的检验一样保守。
3. 如果原假设是简单的（情况通常如此），那么检验的水平就是犯第一类错误的概率。

上述第 3 点涉及到，只有当原假设是简单假设时，我们才能为第一类错误指定一个精确的概率。类似地，只有当备择假设是简单假设时，我们才能为第二类错误指定一个精确的概率。备择假设通常是复合的这一事实，促使了**势函数** (power function) 的定义。

定义 6.12 (势函数) 对于 $\theta \in \Theta$ ，势函数定义为

$$\beta(\theta) = P_\theta(H_0 \text{ 被拒绝}),$$

即，当真实参数值为 θ 时，拒绝原假设的概率。

在特定备择假设 $\theta_1 \in \Theta_1$ 下的检验势 (power) 定义为 $\beta(\theta_1)$ ，即势函数在该特定值 θ_1 处的取值。它与第二类错误的关系就很清楚了，因为

$$P_{\theta_1}(\text{第二类错误}) = P_{\theta_1}(H_0 \text{ 未被拒绝}) = 1 - \beta(\theta_1).$$

对于原假设下的特定参数值，势函数给出了犯第一类错误的概率。如果我们的检验水平是 α ，那么从检验水平的定义可以清楚地看出，对于 $\theta \in \Theta_0$ ，有 $\beta(\theta) \leq \alpha$ 。事实上，显著性水平和检验水平可以用势函数来定义；如果一个检验的显著性水平是 α ，那么

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha,$$

并且，如果检验的检验水平是 α ，那么

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha.$$

6.3.4 Basic ideas in constructing tests

一个完美的假设检验的直观描述很容易构思；这个检验应当从不拒绝一个真实的原假设，并且总是拒绝一个错误的正假设。换句话说，我们希望拒绝一个真实的 H_0 的概率为 0，而拒绝一个错误的 H_0 的概率为 1。这个检验的势函数将是：

$$\beta(\theta) = \begin{cases} 0 & \text{若 } \theta \in \Theta_0, \\ 1 & \text{若 } \theta \notin \Theta_0, \end{cases}$$

这个完美的检验具有水平 0 和势 1。显然，在任何对参数值存在不确定性的情况下，这种理想状态是不可能实现的。在实践中，需要在检验水平与势之间进行权衡。

假设检验的传统方法是控制检验的水平。这等价于控制拒绝一个真实原假设的概率。如果拒绝一个真实原假设的后果非常严重，那么我们可能会将检验水平固定在一个非常小的值。如果我们对拒绝一个真实原假设持更宽松的态度，则可能会为检验水平选择一个较大的值。通常，检验水平固定为 0.05，这种情况被称为“在 5% 的水平下进行检验”。水平为 0.1 和 0.01 的检验也经常被使用。对于一个给定水平的检验，我们希望最大化势，即最大化在原假设为假时拒绝它的概率。

6.3.5 Conclusions and p -values from tests

原则上，假设检验有两种可能的结论：我们要么拒绝原假设，要么不拒绝它。在实践中，需要一些关于反对原假设的证据力度的信息。这可能以检验统计量的观测样本值的形式呈现。另一种呈现此信息的有用方式是使用 p 值。

定义 6.13 (p 值) 考虑一个针对标量参数 θ 的假设检验，其检验统计量为 $U = h(Y)$ 。不失一般性，假设当 U 取值较大时拒绝原假设。对于一个观测样本 y ，相应的 p 值定义为

$$p(y) = \sup_{\theta \in \Theta_0} P(U \geq h(y)).$$

显然，有 $0 \leq p(y) \leq 1$ 。

对于一个简单的原假设，我们可以从定义中移除 $\sup_{\theta \in \Theta_0}$ ，并说 p 值是我们从样本中得到所观测结果或更极端结果的概率。如果我们有一个固定的显著性水平 α ，那么我们可以将检验的临界区域描述为

$$R = \{y : p(y) \leq \alpha\}.$$

简而言之，如果观测到与样本中观测结果至少一样极端的结果的概率很小，那么我们就拒绝原假设。

6.3.6 Hypothesis test for μ and σ^2

下面我们具体看一个例子。已知总体服从正态分布，已知方差 σ^2 ，想要检验 μ 。拿双边检验举例， $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$ 。现在我抽取一个样本，观测到的样本均值为 \bar{X} ，与我声称的总体均值 μ_0 大概率是不相等的，那这个差异有可能是来自随机性，也有可能是因为声称的总体均值是错的，该如何判断呢？我们必须借助确定的概率分布，如果这个观测值出现的概率比较低，我们就认为这不止是随机性导致的差异。具体我们对观测值极端程度的容忍度到哪，取决于我们决定用多大的显著性水平 (significance level)。如果计算发现假设 H_0 成立的情况下出现我们收集到的这个观测值或者更极端的观测值的概率 (p 值) 比我们规定的显著性水平还要小，那我们就拒绝原假设，即 μ 已经在统计学上显著不等于 μ_0 了。

确定的概率分布在哪呢，我们利用前面说过的枢轴变量来作为检验统计量，这个例子中我们选用 U 或者叫 Z 检验统计量：

$$U = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0, 1)$$

注意这是一个分布完全已知的随机变量，但当我们代入观测值 \bar{x} 以后算出的就是检验统计量的一个观测值，是一个实数。我们可以先根据某个给定的显著性水平 α 查出临界值，再拿我们的检验统计量的观测值与临界值做比较；也可以先根据观测值查出 p 值，再拿 p 值与显著性水平 α 做比较。后者的好处在于如果想更换若干个显著性水平观察假设检验的结论无需额外的计算。

注意，置信水平 (confidence level) 等于 1 减去显著性水平 α 。因此，这两个概念是相通的。在 $\alpha = 0.05$ 的显著性水平下，如果未能拒绝 H_0 ，这等价于：假设值 μ_0 会落在基于样本计算的 95% 置信区间之内。反之，如果在 $\alpha = 0.05$ 的显著性水平下拒绝了 H_0 ，这等价于：假设值 μ_0 会落在基于样本计算的 95% 置信区间之外。

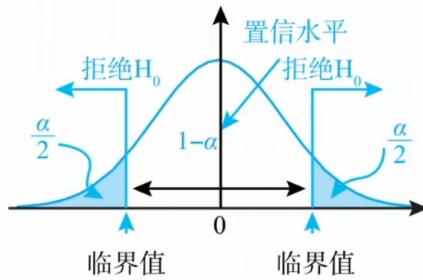


图 4: 双边假设检验中检验统计量的概率分布

注意，第一类错误的概率显然为显著性水平 α ，本质上来说它是一个事先设定的值，也意味着它是我们可以直接控制的风险。事先确定显著性水平为 5% 就等价于事先确定第一类错误的概率为 5%。

$$P(\text{Type 1 error}) = \alpha$$

第二类错误的概率 β 需要计算，且依赖于 α 以及参数的真实值，因此想计算它我们往往需要假设一个真实参数值。与之相关联的概念是检验的功效 (power)，它等于 $1 - \beta$ ，代表已知 H_1 为真时，正确拒绝 H_0 的能力。注意， α 与 β 此消彼长，要同时降低两者，最有效的方法是增加样本量 n 。

注意，“无法拒绝 H_0 ”不代表“ H_0 为真”，它仅仅意味着现有的样本数据没有提供足够的证据来反对 H_0 。事实上，在经典的频率学派统计框架下，我们永远无法知道总体参数的真实值，因此也永远无法通过假设检验来“证明”或“接受”原假设 H_0 为真。一个极端的例子是考虑样本均值的观测值正好等于原假设所声称的总体期望，那么我们绝对会无法拒绝原假设，但这并不意味着 $\mu = \mu_0$ ，这是最典型的“证据不足”的情况。

注意，假设检验的意义，正来自于用未知的、随机的数据，去检验一个已知的、确定的假设。如果把假设也变得随数据而变，那么整个游戏就失去了意义。因此假设必须在数据收集之前确立，且必须有独立于样本数据的来源。

6.4 Prediction

到目前为止，我们一直专注于对未知参数进行推断。在统计学的许多问题中，我们真正的兴趣在于为未观测到的潜在样本成员生成合理的估计；这些可能包括缺失值和未来值。我们将首先假设

【参数检验：单样本】		
单样本均值(μ)	总体正态, 方差 σ^2 已知	$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$
	总体正态, 方差 σ^2 未知	$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$
	总体分布未知或非正态, 大样本($n \geq 30$), 方差 σ^2 未知	$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \rightarrow N(0,1)$
单样本方差(σ^2)	总体正态, 均值 μ 已知	$\chi^2 = \frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\sigma_0^2} \sim \chi_n^2$
	总体正态, 均值 μ 未知	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi_{n-1}^2$
【参数检验：双样本】		
两独立样本均值差($\mu_1 - \mu_2$)	两总体正态, 两方差已知	$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$
	两总体正态, 方差未知但相等 (方差齐性)	$t = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$ $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$
	两总体正态, 方差未知且不等 (方差不齐)	$t = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{\frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2 + (s_2^2/n_2)^2}}$
	两总体分布未知, 大样本($n_1, n_2 \geq 30$), 方差未知	$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \rightarrow N(0,1)$
两配对样本均值差	差值 $D = X_1 - X_2$ 服从正态分布	$t = \frac{\bar{D} - \delta_0}{s_D/\sqrt{n}} \sim t_{n-1}$
两独立样本方差比(σ_1^2/σ_2^2)	两总体均服从正态分布	$F = \frac{s_1^2}{s_2^2} \sim F_{n_1-1, n_2-1}$

图 5: 假设检验的检验统计量

我们有一个随机样本 $X = (X_1, \dots, X_n)^T$, 并且我们有兴趣基于该样本生成一个随机变量 Y 的估计量。换句话说, 我们希望找到一个统计量 $h(X)$, 它能为我们提供一个对 Y 的合理估计量。显然, X 和 Y 之间必须存在某种关联, 这个问题才有意义。一个显而易见的问题是: 我们如何判断什么构成了一个好的估计量? 我们可以借鉴点估计中的一个思想, 使用均方误差:

$$\text{MSE}(h(X)) = \mathbb{E}[(Y - h(X))^2],$$

其中期望是关于 X 和 Y 的联合分布计算的。

一个直观上合理的选择是条件期望 $\mathbb{E}[Y|X]$ 。事实上, 这在均方误差意义上被证明是最优的。在证明最优化之前, 我们先建立条件期望的一个重要性质, 该性质具有几何解释。

引理 6.1 (条件期望作为投影) 对于任何性质足够好的函数 $g : \mathbb{R}^n \rightarrow \mathbb{R}$, 随机变量 $g(X)$ 与 $Y - \mathbb{E}[Y|X]$ 不相关。

定理 6.10 (最小均方误差估计量) 条件期望 $\mathbb{E}[Y|X]$ 是作为 Y 的估计量时, 具有最小均方误差的 X 的函数。

定理 6.11 (MMSE 的性质) Y 的最小均方误差估计量 $\mathbb{E}[Y|X]$ 具有以下性质:

- i. 是无偏的, 即 $\mathbb{E}[Y - \mathbb{E}[Y|X]] = 0$ 。

ii. 其均方误差为 $MSE_Y[\mathbb{E}[Y|X]] = Var(Y) - Var(\mathbb{E}[Y|X])$ 。

在许多情况下，条件期望 $\mathbb{E}[Y|X]$ 难以构造。然而，如果我们将注意力限制在样本的线性函数类估计量上，那么最小化均方误差的估计量可以用 X 和 Y 的联合矩来表示。该估计量被称为**最小均方误差线性估计量** (minimum-mean-square linear estimator, MMSLE)。我们将使用以下记号：

$$\mu_X = \mathbb{E}(X), \quad \mu_Y = \mathbb{E}(Y),$$

$$\Sigma_X = \text{Var}(X), \quad \Sigma_Y = \text{Var}(Y), \quad \Sigma_{YX} = \text{Cov}(Y, X).$$

注意 $\text{Cov}(X, Y) = \Sigma_{XY} = \Sigma_{YX}^T$ 。我们首先考虑零均值的情况。

引理 6.2 (MMSLE 零均值情况) 假设 $\mu_X = 0$ 且 $\mu_Y = 0$ 。如果

$$\tilde{Y} = \Sigma_{YX} \Sigma_X^{-1} X,$$

则 \tilde{Y} 是基于 X 的 Y 的最小均方误差线性估计量。该估计量的均方误差为

$$MSE(\tilde{Y}) = \Sigma_Y - \Sigma_{YX} \Sigma_X^{-1} \Sigma_{XY}.$$

定理 6.12 (MMSLE 一般情况) 基于样本 X 的 Y 的最小均方误差线性估计量由下式给出：

$$\hat{Y} = \mu_Y + \Sigma_{YX} \Sigma_X^{-1} (X - \mu_X).$$

该估计量的均方误差为

$$MSE(\hat{Y}) = \Sigma_Y - \Sigma_{YX} \Sigma_X^{-1} \Sigma_{XY}.$$

此外，如果 X 和 Y 是联合正态分布的，则 MMSLE 及其 MSE 与给定 $X = x$ 时 Y 的条件均值和条件方差相同。换句话说，在正态情况下，MMSLE 同时也是 MMSE。

上述定理与线性回归模型有紧密联系。对于模型 $Y = X\beta + \sigma\varepsilon$ ， Y 的最小二乘估计量为

$$\hat{Y} = X\hat{\beta} = X\underbrace{(X^T X)}_{S_X}^{-1} \underbrace{X^T Y}_{S_{XY}},$$

其中 S_X 和 S_{XY} 分别是 Σ_X 和 Σ_{XY} 的样本类似量。

6.5 Further exercises

6.6 Appendix: Proofs

7 Statistical models 统计模型

7.1 Linear regression

考虑一个我们有两个变量 X 和 Y 的情况。我们认为 Y 值的变化可以通过 X 值的变化来解释。在此背景下， X 被称为**解释变量**（explanatory variable）， Y 被称为**响应变量**（response）。这两个术语都有其他名称。解释变量（根据略有不同的上下文）可能被称为**自变量**（independent variable）、**协变量**（covariate）、**因子**（factor）、**处理**（treatment）或**预测变量**（predictor）；响应变量有时被称为**因变量**（dependent variable）或**目标变量**（target variable）。我们可以为 Y 和 X 之间的关系建模的一种方式是将其视为一条直线。这个简单的想法是一类强大且广泛使用的模型的基础。我们将此类模型称为**线性模型**（linear models）。讨论线性模型的起点是**线性回归**（linear regression）。

7.1.1 Simple linear regression

假设我们有一个随机样本对 $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ 。对于 $i = 1, \dots, n$ ，我们可以写成：

$$Y_i = \alpha + \beta X_i + \sigma \varepsilon_i \quad \text{其中 } \{\varepsilon_i\} \sim \text{IID}(0, 1).$$

通常的做法是通过以变量 X 的观测值为条件来进行推断。因此，方程变为：

$$(Y_i | X_i = x_i) = \alpha + \beta x_i + \sigma \varepsilon_i.$$

考虑该模型的一种便捷方式是使用**条件均值函数**（conditional mean function）和**条件方差函数**（conditional variance function）：

$$\mu(x_i) = \mathbb{E}(Y_i | X_i = x_i) = \alpha + \beta x_i,$$

$$\sigma^2(x_i) = \text{Var}(Y_i | X_i = x_i) = \sigma^2.$$

由此定义的模型只有一个解释变量，因此通常被称为**简单线性回归**（simple linear regression）。

您可能遇到过用更常见的符号表示的此类模型：

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

其中每个误差项 ε_i 均值为 0，方差为 σ^2 。这与前面的方程是等价的，但我们的符号具有明确表示未知参数 σ^2 存在的优点。

7.1.2 Multiple linear regression

现在假设我们有 p 个解释变量 X_1, \dots, X_p 。请注意符号表示与简单回归情况不同；这里的 X_1, \dots, X_p 是 p 个不同变量的集合，而不是来自同一分布的 p 个变量的样本。一个大小为 n 的样本包含 n 个响应值，其中每个响应值都与 p 个解释变量值相关联。我们可以将其表示为：

$$\{(X_{1,1}, \dots, X_{1,p}, Y_1), \dots, (X_{n,1}, \dots, X_{n,p}, Y_n)\},$$

其中 $X_{i,j}$ 表示第 j 个变量的第 i 个观测值。那么， Y 与这 p 个解释变量之间线性关联的模型为：

$$Y_i = \beta_0 + X_{i,1}\beta_1 + \dots + X_{i,p}\beta_p + \sigma \varepsilon_i \quad \text{对于 } i = 1, \dots, n.$$

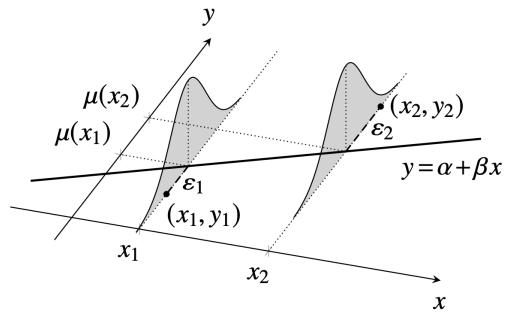


图 6: 简单线性回归可视化

由此定义的 n 个方程可以简洁地用向量形式总结为:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon},$$

其中

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{1,1} & \cdots & \cdots & X_{1,p} \\ 1 & X_{2,1} & \cdots & \cdots & X_{2,p} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots & \cdots & X_{n,p} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad \text{和} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

当我们开始讨论估计 $\boldsymbol{\beta}$ 中的参数时, 这种表述形式的优点将变得清晰。

7.2 Time series models

7.2.1 Autoregressive models

7.2.2 Moving-average models

7.2.3 Autocovariance, autocorrelation, and stationarity

7.3 Poisson processes

7.3.1 Stochastic processes and counting processes

7.3.2 Definitions of the Poisson process

7.3.3 Thinning and superposition

7.3.4 Arrival and interarrival times

7.3.5 Compound Poisson process

7.3.6 Non-homogeneous Poisson process

7.4 Markov chains

7.4.1 Classification of states and chains

7.4.2 Absorption

7.4.3 Periodicity

7.4.4 Limiting distribution

7.4.5 Recurrence and transience

7.4.6 Continuous-time Markov chains

7.5 Further exercises

7.6 Appendix: Proofs

8 Likelihood-based inference 基于似然的估计

在参数统计 (parametric statistics) 中, 我们有时用 $f_Y(y; \theta)$ 表示样本的联合密度 (joint density), 以明确其依赖于未知参数 θ 。我们可以将该密度视为 θ 的函数, 称为似然函数 (likelihood)。尽管密度函数与似然函数在函数形式上完全相同, 但使用似然函数通常更为方便, 尤其是在我们关注不同参数取值下的变化时。

在本章中, 我们将采用一种更具技术性的方法来处理参数估计 (parameter estimation) 与假设检验 (hypothesis testing), 该方法基于似然函数及其相关函数, 例如得分函数 (score) 与信息量 (information)。我们将介绍最大似然估计 (maximum-likelihood estimation), 这是一种重要的推断方法, 并探讨由此得到的估计量 (estimators) 的性质。此外, 我们还将讨论多种似然最大化技术, 例如牛顿-拉弗森方法 (Newton-Raphson method) 与 EM 算法 (EM algorithm)。

我们也将介绍几种基于似然的假设检验方法——似然比检验 (likelihood-ratio test)、得分检验 (score test) 与 Wald 检验 (Wald test)。这些检验方法适用性广泛, 因为它们不要求我们精确推导检验统计量 (test statistics) 的分布。

8.1 Likelihood function and log-likelihood function

在参数统计中, 我们首先考虑来自一个具有已知参数形式 (分布) 的总体样本 $Y = (Y_1, \dots, Y_n)^T$, 该总体包含一个未知的标量参数 θ 。

定义 8.1 (似然函数 (Likelihood function)) 考虑一个密度函数为 $f_Y(y; \theta)$ 的样本 $Y = (Y_1, \dots, Y_n)^T$ 。似然函数与联合概率质量/密度函数具有相同的函数形式, 但被视为参数 θ 的函数而非样本观测值 y 的函数。我们使用符号 L_Y 表示似然函数:

$$L_Y(\theta; y) = f_Y(y; \theta).$$

似然函数表示在给定参数 θ 下, 观察到数据 Y 的概率 (或概率密度)。

定义 8.2 (对数似然函数 (Log-likelihood function)) 如果 L_Y 是一个似然函数, 我们定义对数似然函数 ℓ_Y 为:

$$\ell_Y(\theta; y) \equiv \log L_Y(\theta; y),$$

此定义在 $L_Y(\theta; y) \neq 0$ 处成立。

取对数是一种单调变换; 函数的极大值和极小值的位置不会因对数变换而改变。然而, 对数似然函数的函数形式通常比似然函数本身更便于处理。

例题 8.1 (伯努利分布) 设 $Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} \text{Bernoulli}(\theta)$, 则样本的联合概率质量为:

$$f_Y(y; \theta) = \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i} = \theta^{\sum y_i} (1 - \theta)^{n - \sum y_i}.$$

对应的对数似然函数为:

$$\ell_Y(\theta; y) = \left(\sum_{i=1}^n y_i \right) \log \theta + \left(n - \sum_{i=1}^n y_i \right) \log(1 - \theta).$$

例题 8.2 (正态分布 (方差已知)) 设 $Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} N(\theta, \sigma^2)$, 其中 σ^2 已知。样本的联合密度为:

$$f_Y(y; \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \theta)^2}{2\sigma^2}\right).$$

对应的对数似然函数为：

$$\ell_Y(\theta; y) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta)^2.$$

例题 8.3 (泊松分布) 设 $Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} \text{Poisson}(\theta)$, 则样本的联合概率质量为：

$$f_Y(y; \theta) = \prod_{i=1}^n \frac{e^{-\theta}\theta^{y_i}}{y_i!} = e^{-n\theta}\theta^{\sum y_i} \left(\prod_{i=1}^n y_i!\right)^{-1}.$$

对应的对数似然函数为：

$$\ell_Y(\theta; y) = -n\theta + \left(\sum_{i=1}^n y_i\right) \log \theta - \sum_{i=1}^n \log(y_i!).$$

例题 8.4 (指数分布) 设 $Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} \text{Exp}(\theta)$, 则样本的联合密度为：

$$f_Y(y; \theta) = \prod_{i=1}^n \theta e^{-\theta y_i} = \theta^n \exp\left(-\theta \sum_{i=1}^n y_i\right).$$

对应的对数似然函数为：

$$\ell_Y(\theta; y) = n \log \theta - \theta \sum_{i=1}^n y_i.$$

8.2 Score and information

定义 8.3 (得分函数 (Score function)) 与对数似然函数 $\ell_Y(\theta; y)$ 相关的得分函数定义为：

$$s_Y(\theta; y) = \frac{\partial}{\partial \theta} \ell_Y(\theta; y).$$

根据对数性质与链式法则，我们有：

$$s_Y(\theta; y) = \frac{\frac{\partial}{\partial \theta} L_Y(\theta; y)}{L_Y(\theta; y)}.$$

得分函数的定义明确显示了其对 y 的依赖性。当然，我们也可以将此函数应用于随机变量 Y 。事实上， $s_Y(\theta; Y)$ 的期望值为零。在证明此结果及后续证明中，我们假设 Y 是连续型随机变量。离散情形下有类似的证明。

引理 8.1 (得分函数的一个性质) 假设具备足够的正则性条件，允许我们在积分号下求导，则有：

$$\mathbb{E}[s_Y(\theta; Y)] = 0.$$

对于给定的样本 y ，我们可以绘制似然函数 $L_Y(\theta; y)$ （或等价地，对数似然函数 $\ell_Y(\theta; y)$ ）随 θ 变化的图形。考虑两种极端情况：

- i. 似然函数有一个尖锐的峰：这表明有一小部分 θ 的取值远比其它值更合理。
- ii. 似然函数是平坦的：这意味着很大一个范围内的 θ 取值都同等合理。

在第一种情况下，似然函数为我们提供了大量信息；它允许我们将 θ 的合理取值范围缩小到一个很小的子集。在第二种情况下，观察似然函数无助于我们对 θ 进行推断。第一种情况的特点是似然函数有一个尖锐的峰。当似然函数随 θ 变化而快速变化时，即一阶导数的绝对值很大时，就会出现这种情况。这些想法可以通过定义 Fisher 信息量 (Fisher information) 的概念来形式化。实际上，Fisher 信息量是通过考察对数似然函数导数的平方（而非绝对值），并对样本取值求期望得到的。

定义 8.4 (Fisher 信息量 (Fisher information)) 假设 $Y = (Y_1, \dots, Y_n)^T$ 是来自一个由参数 θ 参数化的总体分布的样本，且 $\ell_Y(\theta; y)$ 是对数似然函数。则样本 Y 中关于参数 θ 的 **Fisher 信息量** 定义为：

$$I_Y(\theta) = \mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \ell_Y(\theta; Y) \right)^2 \right].$$

许多文献会省略 “Fisher”，直接称此量为信息量。

上述定义所定义的量有时被称为 **总 Fisher 信息量** (total Fisher information) (或简称总信息量)。这表明我们考察的是整个样本中的信息。我们很快就会遇到与单个样本成员相关的 Fisher 信息量。

Fisher 信息量是一个可以用多种不同方式表达的量。根据得分函数的定义，我们有：

$$I_Y(\theta) = \mathbb{E} [s_Y(\theta; Y)^2].$$

可能更有用的是，我们可以利用 $\mathbb{E}[s_Y(\theta; Y)] = 0$ 这一事实，从而有：

$$\text{Var}[s_Y(\theta; Y)] = \mathbb{E}[s_Y(\theta; Y)^2],$$

因此可以写作：

$$I_Y(\theta) = \text{Var}[s_Y(\theta; Y)].$$

Fisher 信息量的另一种形式，通常在计算上更为方便，由以下引理给出。

引理 8.2 (Fisher 信息量的另一种形式) 考虑 $Y = (Y_1, \dots, Y_n)^T$ ，这是一个来自由参数 θ 参数化的总体分布的样本，并假设 $\ell_Y(\theta; y)$ 是对数似然函数。则样本 Y 中关于参数 θ 的 Fisher 信息量的另一种表示为：

$$I_Y(\theta) = -\mathbb{E} \left[\frac{\partial^2}{\partial \theta^2} \ell_Y(\theta; Y) \right].$$

$$\hat{I}_Y(\theta; Y) = -\frac{\partial^2}{\partial \theta^2} \ell_Y(\theta; Y)$$

被称为**观测信息量** (observed information)。它是 Fisher 信息量的样本类比，在这个意义上，它是随机样本 Y 的一个函数。注意：

$$\mathbb{E} [\hat{I}_Y(\theta; Y)] = I_Y(\theta).$$

在上一小节中我们看到，如果 $Y = (Y_1, \dots, Y_n)^T$ 是一个随机样本，我们可以将似然表示为个体似然的乘积，将对数似然表示为个体对数似然的和。对于得分函数和信息量也有类似的结果。我们将与单次观测相关的得分函数和信息量分别记为 $s_Y(\theta; y)$ 和 $I_Y(\theta)$ 。

$$s_Y(\theta; y) = \sum_{i=1}^n s_Y(\theta; y_i).$$

我们也可以很容易地证明：

$$I_Y(\theta) = n I_Y(\theta).$$

上述方程表明，对于随机样本，总 Fisher 信息量就是单次观测的 Fisher 信息量乘以样本量；换句话说，信息量具有可加性，并且每个观测值都包含关于未知参数的相同信息量。

定义 8.5 (得分向量与信息矩阵 (Score vector and information matrix)) 考虑样本 $Y = (Y_1, \dots, Y_n)^T$ 及其对数似然函数 $\ell_Y(\theta; Y)$ 。得分向量定义为：

$$s_Y(\theta; y) = \nabla_\theta \ell_Y(\theta; y) = \left(\frac{\partial}{\partial \theta_1} \ell_Y(\theta; y), \dots, \frac{\partial}{\partial \theta_r} \ell_Y(\theta; y) \right)^T,$$

而信息矩阵由下式给出：

$$\mathbf{I}_Y(\theta) = \mathbb{E}[s_Y(\theta; Y)s_Y(\theta; Y)^T].$$

符号 ∇ 表示 del 算子 (del operator)，它是导数在高维空间中的推广。

考虑函数 $g : \mathbb{R}^k \rightarrow \mathbb{R}$ ，若 $x = (x_1, \dots, x_k)^T$ 是一个 k 维向量，则 $g(x)$ 是一个标量。我们使用 del 算子表示 $g(x)$ 对 x 各分量的偏导数，即：

$$\nabla_x g(x) = \left(\frac{\partial}{\partial x_1} g(x), \dots, \frac{\partial}{\partial x_k} g(x) \right)^T.$$

注意 $\nabla_x g(x)$ 与 x 具有相同的维度。

信息矩阵的第 (i, j) 个元素为：

$$[\mathbf{I}_Y(\theta)]_{i,j} = \mathbb{E} \left[\frac{\partial}{\partial \theta_i} \ell_Y(\theta; Y) \frac{\partial}{\partial \theta_j} \ell_Y(\theta; Y) \right].$$

定理 8.1 考虑由参数 θ 参数化的样本 $Y = (Y_1, \dots, Y_n)^T$ 及其对数似然函数 $\ell_Y(\theta; Y)$ 。下列关系成立：

- i. $\mathbb{E}[s_Y(\theta; Y)] = 0$,
- ii. $\mathbf{I}_Y(\theta) = \text{Var}[s_Y(\theta; Y)]$,
- iii. $\mathbf{I}_Y(\theta) = -\mathbb{E} [\nabla_\theta \nabla_\theta^T \ell_Y(\theta; Y)] = -\mathbb{E} [\nabla_\theta s_Y(\theta; Y)]$.

第三个关系表明，信息矩阵的第 (i, j) 个元素可以写为：

$$[\mathbf{I}_Y(\theta)]_{i,j} = -\mathbb{E} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f_Y(Y; \theta) \right].$$

例题 8.5 (正态分布 $N(\mu, \sigma^2)$, σ^2 已知) 设 $Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$ ，其中 σ^2 已知。对数似然函数为：

$$\ell(\mu; \mathbf{y}) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2.$$

得分函数为：

$$s(\mu; \mathbf{y}) = \frac{\partial}{\partial \mu} \ell(\mu; \mathbf{y}) = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu).$$

Fisher 信息量为：

$$I(\mu) = -\mathbb{E} \left[\frac{\partial^2}{\partial \mu^2} \ell(\mu; \mathbf{Y}) \right] = -\mathbb{E} \left[-\frac{n}{\sigma^2} \right] = \frac{n}{\sigma^2}.$$

例题 8.6 (泊松分布 Poisson(λ)) 设 $Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} \text{Poisson}(\lambda)$ 。对数似然函数为：

$$\ell(\lambda; \mathbf{y}) = -n\lambda + \log(\lambda) \sum_{i=1}^n y_i - \sum_{i=1}^n \log(y_i!).$$

得分函数为：

$$s(\lambda; \mathbf{y}) = \frac{\partial}{\partial \lambda} \ell(\lambda; \mathbf{y}) = -n + \frac{1}{\lambda} \sum_{i=1}^n y_i.$$

Fisher 信息量为：

$$I(\lambda) = -\mathbb{E} \left[\frac{\partial^2}{\partial \lambda^2} \ell(\lambda; \mathbf{Y}) \right] = -\mathbb{E} \left[-\frac{1}{\lambda^2} \sum_{i=1}^n Y_i \right] = \frac{n}{\lambda}.$$

例题 8.7 (指数分布 $\text{Exp}(\theta)$) 设 $Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} \text{Exp}(\theta)$ 。对数似然函数为：

$$\ell(\theta; \mathbf{y}) = n \log \theta - \theta \sum_{i=1}^n y_i.$$

得分函数为：

$$s(\theta; \mathbf{y}) = \frac{\partial}{\partial \theta} \ell(\theta; \mathbf{y}) = \frac{n}{\theta} - \sum_{i=1}^n y_i.$$

Fisher 信息量为：

$$I(\theta) = -\mathbb{E} \left[\frac{\partial^2}{\partial \theta^2} \ell(\theta; \mathbf{Y}) \right] = -\mathbb{E} \left[-\frac{n}{\theta^2} \right] = \frac{n}{\theta^2}.$$

例题 8.8 (二项分布 $\text{Binomial}(m, p)$, m 已知) 设 $Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} \text{Binomial}(m, p)$ 。对数似然函数为：

$$\ell(p; \mathbf{y}) = \sum_{i=1}^n \left[\log \binom{m}{y_i} + y_i \log p + (m - y_i) \log(1 - p) \right].$$

得分函数为：

$$s(p; \mathbf{y}) = \frac{\partial}{\partial p} \ell(p; \mathbf{y}) = \frac{\sum_{i=1}^n y_i}{p} - \frac{nm - \sum_{i=1}^n y_i}{1-p}.$$

Fisher 信息量为：

$$I(p) = -\mathbb{E} \left[\frac{\partial^2}{\partial p^2} \ell(p; \mathbf{Y}) \right] = \mathbb{E} \left[\frac{\sum_{i=1}^n Y_i}{p^2} + \frac{nm - \sum_{i=1}^n Y_i}{(1-p)^2} \right] = \frac{nm}{p(1-p)}.$$

8.3 Maximum-likelihood estimation

定义 8.6 (最大似然估计值 (Maximum-likelihood estimate)) 考虑一个由参数 θ 参数化的样本 Y ，并令 L_Y 为似然函数。给定一个观测样本 y ， θ 的 **最大似然估计值 (MLE)** 是使得 $L_Y(\theta; y)$ 作为 θ 的函数取得最大值的那个值，记为 $\hat{\theta}$ 。

关于最大似然估计和最大似然估计量的一些说明如下：

1. 符号的使用有些不幸；上面定义的 $\hat{\theta}$ 是最大似然估计值，只是一个数值。然而，按照惯例， $\hat{\theta}$ 也用来表示相应的统计量。我们尝试在下面澄清：

- 如果 $\hat{\theta} = h(y)$ ，那么 $\hat{\theta}$ 是 θ 的最大似然估计值。这是我们在实践中给定观测样本 y 后计算得到的点估计（一个数值）。

- 如果 $\hat{\theta} = h(Y)$ ，那么 $\hat{\theta}$ 是 θ 的最大似然估计量。这是样本的一个函数。是我们在确定最大似然法的性质时所考虑的理论量。

根据上下文应该能清楚 $\hat{\theta}$ 何时是估计值，何时是估计量。

2. 令 Θ 为参数空间。给定观测样本 y ， θ 的最大似然估计值是满足下式的值 $\hat{\theta}$ ：

$$L_Y(\hat{\theta}; y) = \sup_{\theta \in \Theta} L_Y(\theta; y).$$

注意，如果 Θ 是一个开集，这个定义允许 $\hat{\theta}$ 可能不是 Θ 的成员。我们将处理不会出现此问题的常规情况。

3. 最大似然估计值的定义确保了没有其他可能的参数值具有更大的似然值。如果 $\hat{\theta}$ 是基于观测样本 y 的最大似然估计值，那么对于所有 $\theta \in \Theta$ ，有：

$$L_Y(\theta; y) \leq L_Y(\hat{\theta}; y).$$

4. 我们也可以通过最大化对数似然函数来生成最大似然估计量。取对数是一种单调变换，因此使对数似然函数 $\ell_Y(\theta; Y)$ 最大化的 θ 值，与使似然函数 $L_Y(\theta; Y)$ 最大化的值相同。因此，如果 $\hat{\theta}$ 是基于观测样本 y 的 θ 的最大似然估计，那么：

$$\ell_Y(\hat{\theta}; y) = \sup_{\theta \in \Theta} \ell_Y(\theta; y).$$

5. 最大似然估计值是似然函数达到其最大值的点。通常（但并非总是）我们可以通过求解以下方程来找到给定 y 的 θ 的最大似然估计 $\hat{\theta}$ ：

$$\frac{\partial}{\partial \theta} L_Y(\theta; y) \Big|_{\theta=\hat{\theta}} = 0.$$

根据上面的第 4 点，我们也可以最大化对数似然函数。我们利用得分函数是对数似然函数的导数这一定义。在常规情况下，最大似然估计值 $\hat{\theta}$ 满足：

$$s_Y(\hat{\theta}; y) = 0.$$

上面给出的定义自然转化为我们有多于一个参数的情况。考虑一个由 $\theta = (\theta_1, \dots, \theta_r)$ 参数化的样本 $Y = (Y_1, \dots, Y_n)^T$ ，其中 θ 在参数空间 Θ 中取值。最大似然估计值 $\hat{\theta}$ 是满足下式的值：

$$L_Y(\hat{\theta}; y) = \sup_{\theta \in \Theta} L_Y(\theta; Y),$$

或者等价地，

$$\ell_Y(\hat{\theta}; y) = \sup_{\theta \in \Theta} \ell_Y(\theta; Y).$$

在常规情况下，最大似然估计值满足：

$$s_Y(\hat{\theta}; y) = 0.$$

上述方程实际上定义了一个包含 r 个方程的方程组。在大多数实际感兴趣的案例中，这个方程组没有封闭形式的解，我们需要数值方法来寻找最大似然估计值。

表 8: 矩估计法与最大似然估计法对比

方面	矩估计法	最大似然估计法
必须知道的条件	总体的矩与参数的关系式（例如： $\mathbb{E}(X) = \mu, \text{Var}(X) = \sigma^2$ ）	总体的具体概率分布形式（例如： $X \sim N(\mu, \sigma^2)$ ，并写出其 PDF）
核心思想	用样本矩去“匹配”总体矩。（替换原理）	寻找最可能产生当前样本观测值的参数值。（似然性最大化）
需要求解的方程	矩方程：样本矩 = 总体矩（用参数表示）	似然方程：通常通过求解 $\frac{\partial \ln L(\theta)}{\partial \theta} = 0$ 得到
计算复杂度	通常较低，直接解方程。	可能较高，常涉及求导、对数运算，方程可能无解析解。
对信息的利用	仅利用了矩的信息，信息利用不充分。	利用了分布的全部信息（通过 PDF），信息利用更充分。
估计量的性质	通常更简单直观，但可能不是最优的（如方差不是最小）。	通常具有更优良的统计性质（如一致性、渐近正态性、有效性）。

例题 8.9 (正态分布 $N(\mu, \sigma^2)$, 均值和方差未知) 设 $Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$, 参数 $\theta = (\mu, \sigma^2)$ 。对数似然函数为:

$$\ell(\mu, \sigma^2; \mathbf{y}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2.$$

求偏导得得分函数:

$$\begin{aligned}\frac{\partial \ell}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu), \\ \frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \mu)^2.\end{aligned}$$

令得分函数为零:

$$\begin{aligned}\frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu) &= 0 \Rightarrow \sum_{i=1}^n (y_i - \mu) = 0 \Rightarrow \hat{\mu} = \bar{y}, \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \hat{\mu})^2 &= 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2.\end{aligned}$$

故最大似然估计为 $\hat{\mu} = \bar{Y}$, $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2$ 。

注意到它们与矩估计法得到的估计量是相同的, 总体期望的估计量是无偏的, 而总体方差的估计量是有偏的。

例题 8.10 (均匀分布 $U(0, \theta)$) 设 $Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} U(0, \theta)$ 。似然函数为:

$$L(\theta; \mathbf{y}) = \prod_{i=1}^n \frac{1}{\theta} I(0 \leq y_i \leq \theta) = \theta^{-n} I(0 \leq y_{(1)}) I(y_{(n)} \leq \theta),$$

其中 $y_{(n)} = \max\{y_1, \dots, y_n\}$ 。似然函数在 $\theta \geq y_{(n)}$ 时为 θ^{-n} , 是 θ 的减函数。故在 $\theta = y_{(n)}$ 处取得最大值。因此, 最大似然估计为 $\hat{\theta} = Y_{(n)}$ 。

例题 8.11 (泊松分布 Poisson(λ)) 设 $Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} \text{Poisson}(\lambda)$ 。对数似然函数为:

$$\ell(\lambda; \mathbf{y}) = -n\lambda + \log(\lambda) \sum_{i=1}^n y_i - \sum_{i=1}^n \log(y_i!).$$

求导得得分函数:

$$s(\lambda; \mathbf{y}) = \frac{\partial \ell}{\partial \lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^n y_i.$$

令其为零:

$$-n + \frac{1}{\lambda} \sum_{i=1}^n y_i = 0 \Rightarrow \hat{\lambda} = \bar{y}.$$

故最大似然估计为 $\hat{\lambda} = \bar{Y}$ 。

例题 8.12 (Gamma 分布 Gamma(α, β), 形状参数 α 已知) 设 $Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} \text{Gamma}(\alpha, \beta)$, 其中 α 已知。对数似然函数为:

$$\ell(\beta; \mathbf{y}) = n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^n \log y_i - \beta \sum_{i=1}^n y_i.$$

求导得得分函数：

$$s(\beta; \mathbf{y}) = \frac{\partial \ell}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^n y_i.$$

令其为零：

$$\frac{n\alpha}{\beta} - \sum_{i=1}^n y_i = 0 \Rightarrow \hat{\beta} = \frac{n\alpha}{\sum_{i=1}^n y_i} = \frac{\alpha}{\bar{y}}.$$

故最大似然估计为 $\hat{\beta} = \alpha/\bar{Y}$ 。

8.3.1 Properties of maximum-likelihood estimates

定理 8.2 (最大似然估计量的一致性) 在弱的正则性条件下，最大似然估计量 $\hat{\theta}$ 是 θ 的一个一致估计量。

此处所需的正则性条件及该定理的证明并不特别具有启发性，故不在此列出。我们将满足于假设所处理的所有情况都足够正则，从而保证最大似然估计量的一致性成立。

一致性并未提供关于最大似然估计量抽样分布的线索。然而，我们可以利用中心极限定理和大数定律来确立最大似然估计量的渐近正态性。

定理 8.3 (最大似然估计量的渐近正态性) 在弱的正则性条件下，最大似然估计量 $\hat{\theta}$ 满足：

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I_Y^{-1}(\theta)).$$

该结果的一个直接推论是，在大样本中，我们可以使用正态分布来近似最大似然估计量的抽样分布。

定理 8.4 (最大似然估计量的大样本分布) 如果样本量 n 很大，则最大似然估计量 $\hat{\theta}$ 近似服从：

$$N(\theta, I_Y^{-1}(\theta)).$$

我们常常对参数的某个函数 $g(\theta)$ 感兴趣。给定 θ 的一个估计量 $\hat{\theta}$ ，很自然地会使用 $g(\hat{\theta})$ 作为 $g(\theta)$ 的估计量。这有时被称为 **插件估计量** (plug-in estimator)。通常，插件估计量并不保持原估计量的性质；例如，样本方差 S^2 是 σ^2 的无偏估计，但 S 不是 σ 的无偏估计。最大似然法的一个有用性质是，插件估计量也是最大似然估计量。换句话说，如果 $\hat{\theta}$ 是 θ 的最大似然估计量，那么 $g(\hat{\theta})$ 就是 $g(\theta)$ 的最大似然估计量。

如果 g 是一一对应的，这很容易证明。假设 $\lambda = g(\theta)$ ，且 g 是单射。我们可以利用 $\theta = g^{-1}(\lambda)$ 这一事实进行重新参数化。然后，似然函数可以用 λ 表示为：

$$L_Y(\theta; Y) = L_Y(g^{-1}(\lambda); Y).$$

该函数在 $\hat{\theta}$ 处取得最大值，因此有 $g^{-1}(\hat{\lambda}) = \hat{\theta}$ ，这意味着 $\hat{\lambda} = g(\hat{\theta})$ 。

通常我们感兴趣的是非一一对应的函数。在这种情况下，我们必须明确 $\lambda = g(\theta)$ 时， λ 的最大似然估计量究竟指的是什么。下面的定义予以澄清。

定义 8.7 (诱导似然) 假设 $Y = (Y_1, \dots, Y_n)^T$ 是一个样本，其似然函数 L_Y 由参数 θ 参数化，并令 $\lambda = g(\theta)$ ，其中 $g : \Theta \rightarrow \Lambda$ 。给定观测样本 y ，关于 λ 的 **诱导似然** 定义为：

$$L'_Y(\lambda; y) = \sup_{\{\theta: g(\theta)=\lambda\}} L_Y(\theta; y), \quad \forall \lambda \in \Lambda.$$

诱导似然是一个简单的概念：如果 $g : \Theta \rightarrow \Lambda$ 不是一一对应函数，那么对于给定的值 $\lambda \in \Lambda$ ，可能存在不同的参数值 $\theta_1, \theta_2, \dots, \theta_m \in \Theta$ 使得 $g(\theta_1) = g(\theta_2) = \dots = g(\theta_m) = \lambda$ 。诱导似然 $L'_Y(\lambda; y)$ 就是似然 $L_Y(\theta; y)$ 在集合 $\{\theta_1, \theta_2, \dots, \theta_m\}$ 上的最大值。

注意，这里 Θ 是参数空间， Λ 是函数 g 的值域。那么，很自然地可以将最大似然估计量 $\hat{\lambda}$ 定义为使 $L'_Y(\lambda; y)$ 作为 λ 的函数取得最大值的那个值。利用这个框架，我们可以很容易地证明最大似然估计量的一般不变性。

定理 8.5 (最大似然估计的不变性) 如果 $\hat{\theta}$ 是 θ 的最大似然估计量，那么对于任意函数 $g(\theta)$ ， $g(\theta)$ 的最大似然估计量就是 $g(\hat{\theta})$ 。

8.3.2 Numerical maximization of likelihood

在实践中，通常需要数值优化技术来寻找最大似然估计。我们将考虑参数向量 θ 以及在常规情况下，给定观测样本 y 的 θ 的最大似然估计 $\hat{\theta}$ 满足：

$$\mathbf{s}_Y(\hat{\theta}; y) = \mathbf{0}.$$

得分函数在本节中占据重要地位。为了使符号更简洁，我们将省略得分函数中的下标 Y 和参数 y ，即：

$$\mathbf{s}(\theta) = \mathbf{s}_Y(\theta, y) = \nabla_{\theta} \ell_Y(\theta; y),$$

其中 ∇_{θ} 表示对向量求导。我们将使用 $\dot{\mathbf{s}}$ 表示得分函数转置的一阶导数矩阵，即：

$$\dot{\mathbf{s}}(\theta) = \nabla_{\theta} \mathbf{s}(\theta)^T = \nabla_{\theta} \nabla_{\theta}^T \ell_Y(\theta; y).$$

假设我们当前的估计是 θ^* 。我们希望找到最大似然估计量 $\hat{\theta}$ 。我们知道，在常规情况下， $\hat{\theta}$ 是方程 $\mathbf{s}(\hat{\theta}) = \mathbf{0}$ 的解。

我们对 $\mathbf{s}(\hat{\theta})$ 在 θ^* 处进行泰勒级数展开，以了解如何改进我们的估计。忽略二阶及更高阶项，有：

$$\mathbf{s}(\hat{\theta}) \approx \mathbf{s}(\theta^*) + \dot{\mathbf{s}}(\theta^*)(\hat{\theta} - \theta^*).$$

由于 $\mathbf{s}(\hat{\theta}) = \mathbf{0}$ ，我们可以整理得到：

$$\hat{\theta} \approx \theta^* - \dot{\mathbf{s}}(\theta^*)^{-1} \mathbf{s}(\theta^*).$$

显然，如果 θ^* 距离 $\hat{\theta}$ 很远，高阶项（即包含 $(\hat{\theta} - \theta^*)^2$ 的项）将不可忽略，此时上式给出的近似效果会很差。在实践中，上式被用作一个迭代过程的一部分，该过程称为 Newton-Raphson 方法。

假设 θ_k^* 是我们当前的估计；我们将下一个估计定义为：

$$\theta_{k+1}^* = \theta_k^* - \dot{\mathbf{s}}(\theta_k^*)^{-1} \mathbf{s}(\theta_k^*).$$

迭代过程从某个合理的初始估计 θ_0^* 开始。例如，我们可以使用矩估计法得到的估计值作为该过程的初始值。然后对 $k = 1, 2, \dots$ 重复迭代步骤，希望序列 $\theta_1^*, \theta_2^*, \dots$ 能够收敛到最大似然估计。

评分法是另一种数值过程。它是 Newton-Raphson 方法的一种改进，利用了前面建立的结果：

$$\mathbb{E}(-\dot{\mathbf{s}}(\theta)) = \mathbf{I}_Y(\theta).$$

将 Newton-Raphson 方法中的项 $-\dot{\mathbf{s}}(\theta_k^*)$ 替换为其期望值，迭代变为：

$$\theta_{k+1}^* = \theta_k^* + \mathbf{I}_Y(\theta_k^*)^{-1} \mathbf{s}(\theta_k^*).$$

信息矩阵不依赖于观测样本，并且通常比得分函数的导数更容易计算。

数值方法不能保证找到最大似然估计，也就是说，我们无法确定能找到似然函数的全局最大值。找到最大似然估计的一个必要条件是迭代过程收敛。收敛性是通过连续估计值之间的差异来判断的。我们设定一个任意的容差 $\delta > 0$ ，并且如果能够找到一个 j 使得下式成立，则认为该过程已经收敛：

$$\|\boldsymbol{\theta}_j^* - \boldsymbol{\theta}_{j-1}^*\| < \delta.$$

我们将 $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_j^*$ 视为最大似然估计。所有数值方法都存在一些潜在的缺点，主要与以下两个关键问题有关：

- i. 该过程可能不收敛。当处理相对平坦的似然曲面时，这是一个特别的问题。
- ii. 即使该过程收敛，它也可能不收敛到正确的值。例如，我们可能找到的是一个局部最大值而不是全局最大值。

显然这里需要权衡：如果我们将 δ 设置得太小，该过程可能需要很长时间才能收敛或者根本不收敛。另一方面，如果我们将 δ 设置得太大，我们会增加收敛到一个不接近最大值的值的可能性。 δ 的选择需要考虑多种因素，包括所用计算机的精度、所处理问题的性质，以及非常重要的——过去行之有效的经验。

比较 Newton-Raphson 方法和评分法的相对优点远非易事。如果两种方案都收敛，那么 Newton-Raphson 方法通常收敛得更快。然而，评分法通常对初始值的选择不那么敏感。在实践中，会使用稍微复杂一些的 Newton-Raphson 方法的变体（例如 Broyden-Fletcher-Goldfarb-Shanno 算法及相关方法）来寻找最大似然估计。

8.3.3 EM algorithm

EM 算法是一种用于寻找最大似然估计的迭代过程。该算法的每次迭代包含两个阶段。字母“E”和“M”分别指代一个阶段：E 代表期望（Expectation），M 代表最大化（Maximization）。在描述 EM 算法时，我们将广泛使用条件化的概念。在此上下文中，密度函数记号 $f_Y(y; \boldsymbol{\theta})$ 比似然函数记号 $L_Y(\boldsymbol{\theta}; y)$ 更清晰。我们将使用前者，并理解似然函数与密度函数的函数形式是相同的。

在许多问题中，一个自然的模型设定涉及潜在变量或隐藏变量。顾名思义，这些是我们无法直接观测的变量。为方便起见，我们将所有未观测变量归为一个变量 Z 。我们称 (Y, Z) 为完全数据。通常，为完全数据 (Y, Z) 寻找最大似然估计量比为样本 Y 寻找更容易。然而，这些估计量是我们未观测到的变量的函数。EM 算法通过对隐藏变量取期望来规避这个问题。

我们首先简要描述该算法，然后继续证明该算法将产生一个参数估计序列，该序列永远不会降低似然函数的值。我们使用 $\boldsymbol{\theta}$ 表示通用参数值，并使用 $\boldsymbol{\theta}_j^*$ 表示迭代过程第 j 步的估计值。符号必然相当繁琐，但可以通过首先定义完全数据情况下的对数似然比来稍微简化：

$$K(\boldsymbol{\theta}; \boldsymbol{\theta}_j^*, \mathbf{y}, \mathbf{z}) = \log \frac{f_{Y,Z}(\mathbf{y}, \mathbf{z}; \boldsymbol{\theta})}{f_{Y,Z}(\mathbf{y}, \mathbf{z}; \boldsymbol{\theta}_j^*)},$$

以及给定观测数据条件下，完全数据对数似然比的期望：

$$\begin{aligned} J(\boldsymbol{\theta}; \boldsymbol{\theta}_j^*, \mathbf{y}) &= \mathbb{E}[K(\boldsymbol{\theta}; \boldsymbol{\theta}_j^*, Y, Z) | Y = \mathbf{y}] \\ &= \int_{\mathbb{R}^n} K(\boldsymbol{\theta}; \boldsymbol{\theta}_j^*, \mathbf{y}, \mathbf{z}) f_{Z|Y}(\mathbf{z} | \mathbf{y}; \boldsymbol{\theta}_j^*) d\mathbf{z}. \end{aligned}$$

注意，我们主要将 K 和 J 视为 $\boldsymbol{\theta}$ 的函数。函数 J 在迭代过程的每一步为我们提供了目标函数（即我们希望最大化的函数）。给定某个初始参数值 $\boldsymbol{\theta}_1^*$ ，EM 算法的第 j 步包括两个阶段：

- i. 期望步（E 步）：计算 $J(\boldsymbol{\theta}; \boldsymbol{\theta}_j^*, \mathbf{y})$ 。

ii. 最大化步 (M 步): 令 $\boldsymbol{\theta}_{j+1}^* = \arg \max_{\boldsymbol{\theta}} J(\boldsymbol{\theta}; \boldsymbol{\theta}_j^*, \mathbf{y})$ 。

我们可以用更非正式的方式描述该算法。给定参数 $\boldsymbol{\theta}$ 的一个初始估计, 我们:

- i. 寻找缺失数据的期望值, 将参数估计值视作真实值;
- ii. 估计参数, 将缺失数据的期望值视作真实值。

重复这两个步骤直到收敛。我们现在证明 EM 算法产生一个似然值非递减的估计序列。我们首先证明一个有用的结果。

定理 8.6 (吉布斯不等式 (Gibbs inequality)) 如果 X 是一个密度函数为 f_X 、支撑集为 $A \subset \mathbb{R}$ 的随机变量, 且 f_Y 是任何其他具有相同支撑集的密度函数, 那么:

$$D(f_X, f_Y) = \int_A \log \left(\frac{f_X(x)}{f_Y(x)} \right) f_X(x) dx \geq 0.$$

这个量被称为 f_Y 相对于 f_X 的 **Kullback-Leibler 散度 (Kullback-Leibler divergence)**。

Kullback-Leibler 散度是衡量 f_Y 与“真实”密度 f_X 差异的度量。由于 $g(y) = -\log y$ 对所有 $y > 0$ 是严格凸函数, 仅当两个分布相同时该散度为零。

定理 8.7 (EM 算法估计) 考虑 EM 算法估计序列 $\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \dots$ 。如果 $L_Y(\boldsymbol{\theta}; \mathbf{y})$ 是似然函数, 那么对于 $j = 1, 2, \dots$, 有:

$$L_Y(\boldsymbol{\theta}_{j+1}^*; \mathbf{y}) \geq L_Y(\boldsymbol{\theta}_j^*; \mathbf{y}).$$

虽然上述定理告诉我们 EM 算法永远不会给出比初始值更差的估计, 但它并不保证收敛。事实上, 可以确定 EM 算法估计会收敛到 (不完全数据的) 最大似然估计。然而, 在某些情况下, 收敛速度可能很慢。在实践中, 参数估计通常结合多种方法。例如, 我们可以使用矩估计法生成初始估计, 使用 EM 算法接近最大似然估计, 然后使用某种 Newton-Raphson 方法的变体进行最终优化。

8.4 Likelihood-ratio test (LRT)

考虑假设检验问题:

$$H_0 : \boldsymbol{\theta} \in \Theta_0,$$

$$H_1 : \boldsymbol{\theta} \notin \Theta_0.$$

我们可以将原假设视为对参数取值施加了约束。在此设定中, $\Theta_1 = \Theta_0^c$ 且 $\Theta_0 \cup \Theta_1 = \Theta$, 即原假设和备择假设覆盖了整个参数空间。似然比为这种情况提供了一个明确的检验统计量; 顾名思义, 该统计量基于似然函数。

定义 8.8 (似然比检验统计量 (Likelihood-ratio test statistic)) 考虑一个具有似然函数 L_Y 的样本 Y 。与检验 $H_0: \boldsymbol{\theta} \in \Theta_0$ 相关的 似然比检验统计量 定义为:

$$r(Y) = \frac{\sup_{\boldsymbol{\theta} \in \Theta} L_Y(\boldsymbol{\theta}; Y)}{\sup_{\boldsymbol{\theta} \in \Theta_0} L_Y(\boldsymbol{\theta}; Y)},$$

其中 Θ 是整个参数空间。

似然比检验统计量与最大似然估计密切相关。对于观测样本 \mathbf{y} , 似然比检验统计量的观测值为:

$$r(\mathbf{y}) = \frac{\sup_{\boldsymbol{\theta} \in \Theta} L_Y(\boldsymbol{\theta}; \mathbf{y})}{\sup_{\boldsymbol{\theta} \in \Theta_0} L_Y(\boldsymbol{\theta}; \mathbf{y})}.$$

分子是似然函数的全局最大值；这在最大似然估计 $\hat{\boldsymbol{\theta}}$ 处达到。分母是在约束 $\boldsymbol{\theta} \in \Theta_0$ 下似然函数的最大值；我们用 $\hat{\boldsymbol{\theta}}_0$ 表示该约束似然达到最大值时的值。因此，我们可以写作：

$$r(\mathbf{y}) = \frac{L_Y(\hat{\boldsymbol{\theta}}; \mathbf{y})}{L_Y(\hat{\boldsymbol{\theta}}_0; \mathbf{y})}.$$

似然比检验统计量具有直观的吸引力。我们可以证明，对于所有 \mathbf{y} ，观测值 $r(\mathbf{y})$ 必须大于或等于 1。然而，仅当观测值远大于 1 时，我们才拒绝原假设。粗略地说，我们知道受约束的参数估计 $\hat{\boldsymbol{\theta}}_0$ 会比无约束的估计 $\hat{\boldsymbol{\theta}}$ “差”，体现在其相关的似然值会更小。但是，仅当 $\hat{\boldsymbol{\theta}}_0$ 比 $\hat{\boldsymbol{\theta}}$ “差很多” 时，我们才拒绝约束条件。一个直接的问题是如何判断什么是“差很多”。假设检验的明显解决方案是控制检验的势，即选择一个临界值 c ，使得：

$$P_{\boldsymbol{\theta}_0}(r(Y) > c) = \alpha.$$

似然比检验统计量的观测值是使用似然函数的值计算的。这些值通常必须使用数值优化程序来寻找；因此，似然比的精确抽样分布通常未知也就不足为奇了。然而，在某些条件下，我们可以给出 $2 \log r(Y)$ 的近似分布。

8.4.1 Testing in the presence of nuisance parameters

似然比检验在一个相当特定但非常常见的情况下很有用。我们有一个（现在已熟悉的）由参数 $\boldsymbol{\theta}$ 参数化的样本 Y 的设定。假设 $\boldsymbol{\theta}$ 可以分成两组：主要感兴趣参数 ψ ，以及不太感兴趣的其余参数 λ 。参数 λ 通常被称为 **冗余参数** (nuisance parameters)，这有点不客气。因此，我们可以写作：

$$\boldsymbol{\theta} = \begin{pmatrix} \psi \\ \lambda \end{pmatrix} = (\psi^T, \lambda^T)^T,$$

其中最后表达式中的转置确保 $\boldsymbol{\theta}$ 是一个列向量。在许多情况下，参数的方向并不重要；我们仅将 $\boldsymbol{\theta}, \psi$ 和 λ 用作标量参数的列表。在本节的剩余部分，我们将经常省略转置，并指代：

$$\boldsymbol{\theta} = (\psi, \lambda).$$

有时将 $\boldsymbol{\theta}, \psi$ 和 λ 中的每一个视为单独的（向量）参数会很方便。

给定这个设定，其中 ψ 是感兴趣的参数，我们可能想要检验 ψ 取特定值的假设：

$$\begin{aligned} H_0 : \psi &= \psi_0, \\ H_1 : \psi &\neq \psi_0. \end{aligned}$$

很明显，这里的备择假设是复合假设；可能不那么明显的是，原假设也是复合假设。回想一下，模型由 $\boldsymbol{\theta} = (\psi, \lambda)$ 参数化，因此上式可以表述为：

$$\begin{aligned} H_0 : \psi &= \psi_0, \lambda \in \Lambda, \\ H_1 : \psi &\neq \psi_0, \lambda \in \Lambda, \end{aligned}$$

其中 Λ 是 λ 的参数空间。

原假设可以视为对我们正在拟合的模型施加了约束。在前面的式子中，这个约束是参数 ψ 必须取特定值 ψ_0 。似然比检验统计量比较两个模型，每个模型都有某些优点：

- **无约束模型**拟合得更好，体现在其相关的似然值更大。我们将无约束最大似然估计记为 $\hat{\boldsymbol{\theta}} = (\hat{\psi}, \hat{\lambda})$ 。

- 约束模型需要估计的参数更少。如果 ψ 的维度是 k , 那么 ψ 中包含的 k 个参数的值是固定的; 我们只需要估计 λ , 其中 $\dim(\lambda) < \dim(\theta)$ 。在约束 $\psi = \psi_0$ 下, λ 的最大似然估计是通过关于 λ 最大化 $L_Y(\psi_0, \lambda; y)$ 生成的。我们将得到的估计记为 $\hat{\lambda}_0$ 。因此, θ 的约束最大似然估计是 $\hat{\theta}_0 = (\psi_0, \hat{\lambda}_0)$ 。

在许多情况下, 我们检验 $H_0 : \psi = \mathbf{0}$; 将 ψ 设为 $\mathbf{0}$ 实际上从模型中移除了 k 个参数。这里存在参数数量和拟合优度之间的权衡。我们总是可以通过使用具有更多参数的模型来提高拟合优度。然而, 参数过多的模型在用于预测时可能表现不佳。统计学的指导原则之一是 **简约性** (parsimony), 即我们应使用能以最少参数提供足够拟合的模型。似然比检验在实际情况下用于帮助确定模型的哪些特征是重要的。

8.4.2 Properties of the likelihood ratio

定理 8.8 (似然比的渐近分布——简单情况) 考虑一个由标量参数 θ 参数化的随机样本 Y 。令 $r(Y)$ 为与检验 $H_0 : \theta = \theta_0$ 相关的似然比检验统计量。在原假设下, 若满足某些正则性条件, 则有:

$$2 \log r(Y) \xrightarrow{d} \chi_1^2.$$

定理 8.9 (似然比的渐近分布——存在冗余参数情况) 考虑一个由 $\theta = (\psi, \lambda)$ 参数化的随机样本 Y , 其中 ψ 的维度为 k 。令 $r(Y)$ 为与检验 $H_0 : \psi = \psi_0$ 相关的似然比检验统计量。在原假设下, 若满足某些正则性条件, 则有:

$$2 \log r(Y) \xrightarrow{d} \chi_k^2.$$

8.4.3 Approximate tests

计算似然比检验统计量涉及在原假设模型和备择假设模型下最大化似然函数。在实践中, 这并非总是可行的。**得分检验** (Score test) 和 **Wald 检验** (Wald test) 提供了有吸引力的替代方案, 它们可以被视为对似然比检验的近似。基本设定与似然比检验中使用的相同, 即我们在存在冗余参数 λ 的情况下检验 $H_0 : \psi = \psi_0$ 与 $H_1 : \psi \neq \psi_0$ 。此外, 我们假设 ψ 的长度为 k , 并且所有参数可以组合为 $\theta = (\psi, \lambda)$ 。

得分检验

假设在原假设下的最大似然估计 $\hat{\theta}_0$ 是已知的; 我们有时称此为约束 MLE。得分检验背后的思想是, 在 $\hat{\theta}$ (无约束 MLE) 处评估的得分函数等于零, 因此如果在 $\hat{\theta}_0$ 处的得分远离零, 则有理由拒绝 H_0 。

我们可以基于 $\hat{\theta}_0$ 构建一个检验统计量, 其在原假设下的分布近似于 $-2 \log r(Y)$ 。令 $s_1(\theta; Y)$ 表示得分向量的前 k 个元素, 即对应于参数 ψ 的元素。得分检验统计量定义为:

$$S = h(Y; \hat{\theta}_0) = s_1(\hat{\theta}_0; Y)^T [\mathcal{I}^{-1}(\hat{\theta}_0)]_{11} s_1(\hat{\theta}_0; Y),$$

其中 $[\mathcal{I}^{-1}(\theta)]_{11}$ 是信息矩阵的逆矩阵 $\mathcal{I}^{-1}(\theta)$ 的 $k \times k$ 左上块。这个块是 $s_1(\psi; Y)$ 的方差, 因此得分检验统计量等于在 $\hat{\theta}_0$ 处的得分函数的大小, 并经过其方差标准化。

S 的分布近似为 χ_k^2 , 我们在 S 取大值时拒绝 H_0 。注意, 执行此检验我们不需要知道无约束 MLE。

Wald 检验

Wald 检验在相反的情况下很有用，即当我们只有无约束 MLE $\hat{\boldsymbol{\theta}}$ 时。该检验基于这样的思想：当 H_0 为真时，我们期望无约束 MLE $\hat{\boldsymbol{\psi}}$ 不会离原假设值 $\boldsymbol{\psi}_0$ 太远。Wald 检验统计量定义为：

$$W = (\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^T [\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})]_{11}^{-1} (\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0).$$

根据定理， $(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)$ 的方差近似等于 $[\mathcal{I}^{-1}(\boldsymbol{\theta})]_{11}$ 。我们可以通过在 MLE $\hat{\boldsymbol{\theta}}$ 处评估信息来估计这个方差。因此，Wald 检验统计量等于差值 $(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)$ 的大小，并经过其（估计的）方差标准化。与得分检验统计量类似，Wald 检验统计量在原假设下的分布也近似为 χ_k^2 ，我们在 W 取大值时拒绝 H_0 。

8.5 Further exercises

1. 已知一个大小为 n 的随机样本，总体分布的密度函数为

$$f_Y(y; \theta) = \frac{2y}{\theta^2}, \quad 0 < y \leq \theta$$

求 θ 的最大似然估计量。提示：不要用微积分，画出对数似然函数的图像。

解答：

联合密度函数：

$$\prod_{i=1}^n \frac{2y_i}{\theta^2} = \frac{2^n \pi y_i}{\theta^{2n}}$$

对数似然：

$$\ell_Y = \ln(2^n \pi y_i) - \ln(\theta^{2n}) = n \ln(2) + \sum \ln(y_i) - 2n \ln(\theta) = -2n \ln(\theta) + C$$

这是一条单调减少经过 $(1, 0)$ 点的曲线，想让对数似然越大就是要让参数 θ 越小，但 $\theta \geq y_i \quad \forall i \in \{1, 2, \dots, n\}$ ，因此 $\hat{\theta} = Y_{max}$ 。

2. 设 X_1, \dots, X_n 是来自密度函数

$$f_X(x) = \begin{cases} \theta^2 x e^{-\theta x} & x > 0, \\ 0 & x \leq 0, \end{cases}$$

的随机样本，其中 $\theta > 0$ 是未知参数。

(a) 求 θ 的最大似然估计量。

(b) 考虑仅知道最后 $n-m$ 个观测值 ($m < n$) 的具体数值，而对于前 m 个观测值，仅知道它们的和 $Z = X_1 + \dots + X_m$ 的情况。求 Z 的概率密度函数，并基于观测值 Z, X_{m+1}, \dots, X_n 确定 θ 的最大似然估计。

(c) 假设现在前 m 个观测值是删失的，即对于 $j = 1, \dots, m$ ，我们只知道 $X_j \in [u_j, v_j]$ (其中 $v_j > u_j > 0$ 是给定的)，而 X_j 本身未知。描述一个用于估计参数 θ 的 EM 算法。

解答：

(a) 似然函数为：

$$L(\theta) = \prod_{i=1}^n \theta^2 x_i e^{-\theta x_i} = \theta^{2n} \left(\prod_{i=1}^n x_i \right) e^{-\theta \sum_{i=1}^n x_i}.$$

对数似然函数为：

$$\ell(\theta) = 2n \ln \theta + \sum_{i=1}^n \ln x_i - \theta \sum_{i=1}^n x_i.$$

对 θ 求导并令导数为零：

$$\frac{\partial \ell}{\partial \theta} = \frac{2n}{\theta} - \sum_{i=1}^n x_i = 0.$$

解得最大似然估计量为：

$$\hat{\theta} = \frac{2n}{\sum_{i=1}^n X_i}.$$

(b) 注意到 $X_i \sim \text{Gamma}(2, \theta)$, 因此 $Z = \sum_{i=1}^m X_i \sim \text{Gamma}(2m, \theta)$, 其密度函数为:

$$f_Z(z) = \frac{\theta^{2m}}{\Gamma(2m)} z^{2m-1} e^{-\theta z}, \quad z > 0.$$

基于观测值 Z, X_{m+1}, \dots, X_n 的似然函数为:

$$L(\theta) = f_Z(z) \prod_{i=m+1}^n \theta^2 x_i e^{-\theta x_i} = \frac{\theta^{2m}}{\Gamma(2m)} z^{2m-1} e^{-\theta z} \cdot \theta^{2(n-m)} \left(\prod_{i=m+1}^n x_i \right) e^{-\theta \sum_{i=m+1}^n x_i}.$$

对数似然函数为:

$$\ell(\theta) = 2n \ln \theta - \theta \left(z + \sum_{i=m+1}^n x_i \right) + \text{常数}.$$

对 θ 求导并令导数为零:

$$\frac{\partial \ell}{\partial \theta} = \frac{2n}{\theta} - \left(z + \sum_{i=m+1}^n x_i \right) = 0.$$

解得最大似然估计量为:

$$\hat{\theta} = \frac{2n}{Z + \sum_{i=m+1}^n X_i}.$$

(c) EM 算法描述如下:

- 步骤 1: 初始话选择一个初始值 $\theta^{(0)}$ 。
- 步骤 2: E 步在第 t 次迭代, 计算完整数据的对数似然函数关于条件期望的表达式:

$$Q(\theta | \theta^{(t)}) = \mathbb{E} \left[\sum_{i=1}^n \ln f_X(X_i | \theta) \mid \text{观测数据}, \theta^{(t)} \right].$$

对于已知的观测值 X_{m+1}, \dots, X_n , 直接使用其值。对于删失的观测值 $X_j \in [u_j, v_j]$, 计算条件期望:

$$\begin{aligned} \mathbb{E}[X_j | X_j \in [u_j, v_j], \theta^{(t)}] &= \frac{\int_{u_j}^{v_j} x \cdot \theta^{(t)2} x e^{-\theta^{(t)}x} dx}{\int_{u_j}^{v_j} \theta^{(t)2} x e^{-\theta^{(t)}x} dx}, \\ \mathbb{E}[\ln X_j | X_j \in [u_j, v_j], \theta^{(t)}] &= \frac{\int_{u_j}^{v_j} \ln x \cdot \theta^{(t)2} x e^{-\theta^{(t)}x} dx}{\int_{u_j}^{v_j} \theta^{(t)2} x e^{-\theta^{(t)}x} dx}. \end{aligned}$$

- 步骤 3: M 步最大化 $Q(\theta | \theta^{(t)})$:

$$\theta^{(t+1)} = \frac{2n}{\sum_{j=1}^m \mathbb{E}[X_j | X_j \in [u_j, v_j], \theta^{(t)}] + \sum_{i=m+1}^n X_i}.$$

- 步骤 4: 迭代重复 E 步和 M 步直至收敛。

3. 考虑简单线性回归模型

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1), \quad i = 1, \dots, n.$$

构造原假设 $H_0: \beta_1 = 2\beta_0$ 对备择假设 $H_1: \beta_1 \neq 2\beta_0$ 的 Wald 检验统计量。

解答:

令 $\theta = (\beta_0, \beta_1)^T$, 原假设可写为:

$$H_0: \beta_1 - 2\beta_0 = 0.$$

定义约束函数 $g(\theta) = \beta_1 - 2\beta_0$, 则 $H_0 : g(\theta) = 0$ 。

在正态误差且方差已知的情况下, $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1)^T$ 的方差协方差矩阵为:

$$\text{Var}(\hat{\theta}) = I_n^{-1} = (X^T X)^{-1},$$

其中 X 是设计矩阵, 第一列为全 1, 第二列为 $(X_1, \dots, X_n)^T$ 。

$g(\theta)$ 的梯度向量为:

$$\nabla g(\theta) = \frac{\partial g}{\partial \theta} = (-2, 1)^T.$$

Wald 检验统计量为:

$$W = g(\hat{\theta})^T \left[\nabla g(\hat{\theta})^T \cdot \text{Var}(\hat{\theta}) \cdot \nabla g(\hat{\theta}) \right]^{-1} g(\hat{\theta}).$$

代入 $g(\hat{\theta}) = \hat{\beta}_1 - 2\hat{\beta}_0$ 和 $\nabla g(\hat{\theta}) = (-2, 1)^T$, 得:

$$W = (\hat{\beta}_1 - 2\hat{\beta}_0)^2 \left[(-2, 1)(X^T X)^{-1} \begin{pmatrix} -2 \\ 1 \end{pmatrix} \right]^{-1}.$$

记 $(X^T X)^{-1} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$, 则:

$$(-2, 1) \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} -2 \\ 1 \end{pmatrix} = 4a - 4b + c.$$

因此 Wald 检验统计量为:

$$W = \frac{(\hat{\beta}_1 - 2\hat{\beta}_0)^2}{4a - 4b + c},$$

其中 a, b, c 是 $(X^T X)^{-1}$ 的元素。在原假设下, W 演近服从 χ_1^2 分布。

4. 设 $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, 其中 σ^2 已知。考虑检验问题:

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$$

(a) 写出似然比检验统计量的表达式

(b) 证明该统计量服从 χ_1^2 分布

解答:

步骤 1: 写出似然函数

$$L(\mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2\right)$$

步骤 2: 求最大似然估计在 H_1 下, $\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

在 H_0 下, $\mu = \mu_0$

步骤 3: 计算似然比

$$\begin{aligned} \Lambda &= \frac{L(\mu_0)}{L(\hat{\mu})} = \frac{\exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu_0)^2\right)}{\exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2\right)} \\ &= \exp\left(-\frac{1}{2\sigma^2} \left[\sum_{i=1}^n (X_i - \mu_0)^2 - \sum_{i=1}^n (X_i - \bar{X})^2 \right]\right) \end{aligned}$$

步骤 4: 简化表达式注意到:

$$\begin{aligned}
\sum_{i=1}^n (X_i - \mu_0)^2 &= \sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu_0)^2 \\
&= \sum_{i=1}^n (X_i - \bar{X})^2 + 2(\bar{X} - \mu_0) \sum_{i=1}^n (X_i - \bar{X}) + n(\bar{X} - \mu_0)^2 \\
&= \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2
\end{aligned}$$

因为 $\sum_{i=1}^n (X_i - \bar{X}) = 0$

步骤 5: 得到检验统计量

$$\begin{aligned}
-2 \ln \Lambda &= -2 \ln \left[\exp \left(-\frac{n(\bar{X} - \mu_0)^2}{2\sigma^2} \right) \right] = \frac{n(\bar{X} - \mu_0)^2}{\sigma^2} \\
&= \left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right)^2 \sim \chi_1^2
\end{aligned}$$

5. 设 $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$, 考虑检验:

$$H_0 : \lambda = \lambda_0 \quad \text{vs} \quad H_1 : \lambda \neq \lambda_0$$

(a) 写出得分函数和对数似然函数

(b) 构造 Score 检验统计量

(c) 说明其渐近分布

解答:

步骤 1: 写出似然函数和对数似然函数

$$L(\lambda) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{X_i}}{X_i!}, \quad \ell(\lambda) = \ln L(\lambda) = -n\lambda + \left(\sum_{i=1}^n X_i \right) \ln \lambda - \sum_{i=1}^n \ln X_i!$$

步骤 2: 求得分函数 (一阶导数)

$$s(\lambda) = \frac{\partial \ell}{\partial \lambda} = -n + \frac{\sum_{i=1}^n X_i}{\lambda}$$

步骤 3: 求 Fisher 信息量

$$I(\lambda) = -E \left[\frac{\partial^2 \ell}{\partial \lambda^2} \right] = -E \left[-\frac{\sum X_i}{\lambda^2} \right] = \frac{nE[X_i]}{\lambda^2} = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}$$

步骤 4: 构造 Score 检验统计量在 $H_0 : \lambda = \lambda_0$ 下:

$$S = \frac{[s(\lambda_0)]^2}{I_n(\lambda_0)} = \frac{\left(-n + \frac{\sum X_i}{\lambda_0} \right)^2}{n/\lambda_0} = \frac{(\sum X_i - n\lambda_0)^2}{n\lambda_0}$$

步骤 5: 渐近分布在 H_0 下, $S \xrightarrow{d} \chi_1^2$

6. 考虑线性回归模型:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

检验假设:

$$H_0 : \beta_1 + \beta_2 = 1 \quad \text{vs} \quad H_1 : \beta_1 + \beta_2 \neq 1$$

(a) 用矩阵形式写出 Wald 统计量

(b) 说明需要估计哪些量

解答:

步骤 1: 矩阵表示令 $\beta = (\beta_0, \beta_1, \beta_2)^T$, $X = \begin{bmatrix} 1 & X_{11} & X_{21} \\ \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} \end{bmatrix}$, $Y = (Y_1, \dots, Y_n)^T$

步骤 2: 约束的矩阵表示约束 $\beta_1 + \beta_2 = 1$ 可写为 $R\beta = r$, 其中:

$$R = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}, \quad r = 1$$

步骤 3: OLS 估计量

$$\hat{\beta} = (X^T X)^{-1} X^T Y, \quad \hat{\sigma}^2 = \frac{1}{n-3} (Y - X\hat{\beta})^T (Y - X\hat{\beta})$$

步骤 4: 协方差矩阵估计

$$\widehat{\text{Var}}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1}$$

步骤 5: Wald 统计量

$$\begin{aligned} W &= (R\hat{\beta} - r)^T [R\widehat{\text{Var}}(\hat{\beta}) R^T]^{-1} (R\hat{\beta} - r) \\ &= (\hat{\beta}_1 + \hat{\beta}_2 - 1)^T [\hat{\sigma}^2 R(X^T X)^{-1} R^T]^{-1} (\hat{\beta}_1 + \hat{\beta}_2 - 1) \end{aligned}$$

步骤 6: 计算标量形式由于 $R(X^T X)^{-1} R^T$ 是一个标量, 记:

$$V = R(X^T X)^{-1} R^T = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} v_{00} & v_{01} & v_{02} \\ v_{10} & v_{11} & v_{12} \\ v_{20} & v_{21} & v_{22} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = v_{11} + 2v_{12} + v_{22}$$

其中 v_{ij} 是 $(X^T X)^{-1}$ 的元素。

最终:

$$W = \frac{(\hat{\beta}_1 + \hat{\beta}_2 - 1)^2}{\hat{\sigma}^2 (v_{11} + 2v_{12} + v_{22})} \sim \chi_1^2$$

7. 设 $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$, 但前 m 个观测被右删失, 即我们只知道 $X_j > c_j$ ($j = 1, \dots, m$), 后 $n-m$ 个观测完全可见。

(a) 写出完整的似然函数

(b) 描述 EM 算法的 E 步和 M 步

(c) 写出参数更新的显式表达式

解答:

步骤 1: 完整数据似然函数如果所有数据都观测到:

$$L_c(\lambda) = \prod_{i=1}^n \lambda e^{-\lambda X_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^n X_i\right)$$

步骤 2: 观测数据似然函数对于删失数据, 贡献生存函数 $P(X_j > c_j) = e^{-\lambda c_j}$

$$L_o(\lambda) = \prod_{j=1}^m e^{-\lambda c_j} \cdot \prod_{i=m+1}^n \lambda e^{-\lambda X_i}$$

步骤 3: E 步 - 计算 Q 函数

$$\begin{aligned} Q(\lambda|\lambda^{(t)}) &= E[\ell_c(\lambda)|\text{观测数据}, \lambda^{(t)}] \\ &= E\left[\sum_{i=1}^n (\ln \lambda - \lambda X_i) |\text{观测数据}, \lambda^{(t)}\right] \\ &= n \ln \lambda - \lambda \left[\sum_{j=1}^m E[X_j | X_j > c_j] + \sum_{i=m+1}^n X_i \right] \end{aligned}$$

对于指数分布, $E[X_j | X_j > c_j] = c_j + \frac{1}{\lambda^{(t)}}$

步骤 4: M 步 - 最大化 Q 函数对 $Q(\lambda|\lambda^{(t)})$ 关于 λ 求导:

$$\frac{\partial Q}{\partial \lambda} = \frac{n}{\lambda} - \left[\sum_{j=1}^m E[X_j | X_j > c_j] + \sum_{i=m+1}^n X_i \right] = 0$$

解得:

$$\lambda^{(t+1)} = \frac{n}{\sum_{j=1}^m E[X_j | X_j > c_j] + \sum_{i=m+1}^n X_i} = \frac{n}{\sum_{j=1}^m (c_j + \frac{1}{\lambda^{(t)}}) + \sum_{i=m+1}^n X_i}$$

步骤 5: 迭代公式

$$\lambda^{(t+1)} = \frac{n}{\sum_{j=1}^m c_j + \sum_{i=m+1}^n X_i + \frac{m}{\lambda^{(t)}}}$$

8.6 Appendix: Proofs

9 Inferential theory 推断理论

9.1 Sufficiency

9.1.1 Sufficient statistics and the sufficiency principle

考虑一个由参数 θ 参数化的样本 \mathbf{Y} 。充分性为一个统计量提供了一种可能的定义，该统计量概括了 \mathbf{Y} 中包含的关于 θ 的所有相关信息。

定义 9.1 (充分性) 假设 $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ 是一个样本。一个统计量 $\mathbf{U} = h(\mathbf{Y})$ 是参数 θ 的充分统计量，如果给定 \mathbf{U} 时 \mathbf{Y} 的条件分布不依赖于 θ 。

充分性的定义指的是 \mathbf{Y} 的分布。因此，一个统计量有时被定义为对于一个分布族 $\{F_{\mathbf{Y}}(\cdot, \theta) : \theta \in \Theta\}$ 是充分的。那么 \mathbf{U} 是一个充分统计量的陈述，等价于说条件分布函数 $F_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u})$ 不是 θ 的函数。等价地，如果 $f_{\mathbf{Y}}(\cdot; \theta)$ 是 \mathbf{Y} 的概率质量/密度函数，那么 $f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u})$ 不是 θ 的函数。

条件质量/密度函数 $f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u})$ 有些特殊；我们是以一个统计量 \mathbf{U} 为条件，而 \mathbf{U} 是样本 \mathbf{Y} 的一个函数。利用这类条件的性质来推导充分统计量的结果。通常，这些结果在离散情形下更容易推导，因为条件质量可以解释为概率。

引理 9.1 (以样本函数为条件——离散情形) 假设 \mathbf{Y} 是来自离散分布的一个样本，且 $\mathbf{U} = h(\mathbf{Y})$ 是一个统计量。则给定 \mathbf{U} 时 \mathbf{Y} 的条件质量函数为：

$$f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u}) = \begin{cases} f_{\mathbf{Y}}(\mathbf{y})/f_{\mathbf{U}}(\mathbf{u}) & \text{当 } h(\mathbf{y}) = \mathbf{u} \text{ 且 } f_{\mathbf{U}}(\mathbf{u}) \neq 0 \text{ 时,} \\ 0 & \text{否则.} \end{cases}$$

如果我们相信一个关于 θ 的充分统计量包含了样本所能提供的所有相关信息，那么坚持要求我们对 θ 所做的任何推断仅使用一个充分统计量是合理的。这被称为**充分性原则**。它在下面被更正式地陈述。

充分性原则

如果 $\mathbf{U} = h(\mathbf{Y})$ 是 θ 的一个充分统计量，那么关于 θ 的推断应该仅通过 $\mathbf{u} = h(\mathbf{y})$ 依赖于样本 \mathbf{Y} 。

假设 $\mathbf{U} = h(\mathbf{Y})$ 是 θ 的一个充分统计量。考虑两个观测样本 \mathbf{x} 和 \mathbf{y} 。如果对于这两个样本，充分统计量的观测值相同，即如果 $h(\mathbf{x}) = h(\mathbf{y})$ ，那么充分性原则告诉我们，基于观测 \mathbf{x} 对 θ 的推断应该与基于观测 \mathbf{y} 的推断相同。

例题 9.1 设 X_1, X_2, \dots, X_n 是来自伯努利分布 $B(1, p)$ 的一个简单随机样本，其中 $0 < p < 1$ 。证明统计量 $T(\mathbf{X}) = \sum_{i=1}^n X_i$ 是参数 p 的充分统计量。

我们需要证明给定 $T = t$ 时， (X_1, \dots, X_n) 的条件分布与参数 p 无关。

首先， $T = \sum_{i=1}^n X_i \sim B(n, p)$ ，其概率函数为：

$$P(T = t) = \binom{n}{t} p^t (1-p)^{n-t}, \quad t = 0, 1, \dots, n$$

对于任意满足 $\sum_{i=1}^n x_i = t$ 的样本观测值 $\mathbf{x} = (x_1, \dots, x_n)$ ，其联合概率为：

$$P(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^t (1-p)^{n-t}$$

因此, 给定 $T = t$ 时, \mathbf{X} 的条件概率为:

$$P(\mathbf{X} = \mathbf{x} \mid T = t) = \frac{P(\mathbf{X} = \mathbf{x})}{P(T = t)} = \frac{p^t(1-p)^{n-t}}{\binom{n}{t} p^t(1-p)^{n-t}} = \frac{1}{\binom{n}{t}}$$

该条件概率不依赖于参数 p , 因此 $T(\mathbf{X}) = \sum_{i=1}^n X_i$ 是 p 的充分统计量。

例题 9.2 设 X_1, X_2, \dots, X_n 是来自泊松分布 $P(\lambda)$ 的一个简单随机样本, 其中 $\lambda > 0$ 。证明统计量 $T(\mathbf{X}) = \sum_{i=1}^n X_i$ 是参数 λ 的充分统计量。

我们需要证明给定 $T = t$ 时, (X_1, \dots, X_n) 的条件分布与参数 λ 无关。

由于 $X_i \sim P(\lambda)$ 且相互独立, 有 $T = \sum_{i=1}^n X_i \sim P(n\lambda)$, 其概率函数为:

$$P(T = t) = \frac{e^{-n\lambda}(n\lambda)^t}{t!}, \quad t = 0, 1, 2, \dots$$

对于任意满足 $\sum_{i=1}^n x_i = t$ 的样本观测值 $\mathbf{x} = (x_1, \dots, x_n)$, 其联合概率为:

$$P(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^t}{\prod_{i=1}^n x_i!}$$

因此, 给定 $T = t$ 时, \mathbf{X} 的条件概率为:

$$P(\mathbf{X} = \mathbf{x} \mid T = t) = \frac{P(\mathbf{X} = \mathbf{x})}{P(T = t)} = \frac{e^{-n\lambda}\lambda^t / \prod_{i=1}^n x_i!}{e^{-n\lambda}(n\lambda)^t / t!} = \frac{t!}{n^t \prod_{i=1}^n x_i!}$$

该条件概率不依赖于参数 λ , 因此 $T(\mathbf{X}) = \sum_{i=1}^n X_i$ 是 λ 的充分统计量。

例题 9.3 设 X_1, X_2, \dots, X_n 是来自正态分布 $N(\mu, \sigma_0^2)$ 的一个简单随机样本, 其中 μ 未知, σ_0^2 已知。证明样本均值 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 是参数 μ 的充分统计量。

我们需要证明给定 $\bar{X} = \bar{x}$ 时, (X_1, \dots, X_n) 的条件分布与参数 μ 无关。

考虑联合密度函数:

$$f(\mathbf{x} \mid \mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma_0^2}\right) = (2\pi\sigma_0^2)^{-n/2} \exp\left(-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu)^2\right)$$

将 $\sum_{i=1}^n (x_i - \mu)^2$ 展开:

$$\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$$

代入得:

$$f(\mathbf{x} \mid \mu) = (2\pi\sigma_0^2)^{-n/2} \exp\left(-\frac{1}{2\sigma_0^2} \left[\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \right] \right)$$

由于 $\bar{X} \sim N(\mu, \sigma_0^2/n)$, 其密度函数为:

$$f_{\bar{X}}(\bar{x} \mid \mu) = \frac{\sqrt{n}}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{n(\bar{x} - \mu)^2}{2\sigma_0^2}\right)$$

因此, 给定 $\bar{X} = \bar{x}$ 时, \mathbf{X} 的条件密度为:

$$f(\mathbf{x} \mid \bar{X} = \bar{x}, \mu) = \frac{f(\mathbf{x} \mid \mu)}{f_{\bar{X}}(\bar{x} \mid \mu)} = \frac{(2\pi\sigma_0^2)^{-n/2} \exp\left(-\frac{1}{2\sigma_0^2} [\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2]\right)}{\frac{\sqrt{n}}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{n(\bar{x} - \mu)^2}{2\sigma_0^2}\right)}$$

化简得:

$$f(\mathbf{x} \mid \bar{X} = \bar{x}, \mu) = \frac{1}{(2\pi\sigma_0^2)^{(n-1)/2} \sqrt{n}} \exp\left(-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \bar{x})^2\right)$$

该条件密度不依赖于参数 μ , 因此 \bar{X} 是 μ 的充分统计量。

表 9: 常见分布的充分统计量

分布	参数	充分统计量
伯努利分布 $B(1, p)$	p	$T(X) = X$ (单个观测) $T(\mathbf{X}) = \sum_{i=1}^n X_i$ (样本)
二项分布 $B(n, p)$	p	$T(X) = X$ (单个观测) $T(\mathbf{X}) = \sum_{i=1}^n X_i$ (样本)
泊松分布 $P(\lambda)$	λ	$T(X) = X$ (单个观测) $T(\mathbf{X}) = \sum_{i=1}^n X_i$ (样本)
正态分布 $N(\mu, \sigma^2)$	μ (σ^2 已知)	$T(\mathbf{X}) = \bar{X}$
	σ^2 (μ 已知)	$T(\mathbf{X}) = \sum_{i=1}^n (X_i - \mu)^2$
	(μ, σ^2) (均未知)	$T(\mathbf{X}) = (\bar{X}, \sum_{i=1}^n (X_i - \bar{X})^2)$
指数分布 $Exp(\lambda)$	λ	$T(X) = X$ (单个观测) $T(\mathbf{X}) = \sum_{i=1}^n X_i$ (样本)
伽马分布 $Gamma(\alpha, \beta)$	α (β 已知)	$T(\mathbf{X}) = \sum_{i=1}^n \ln X_i$
	β (α 已知)	$T(\mathbf{X}) = \sum_{i=1}^n X_i$
	(α, β) (均未知)	$T(\mathbf{X}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n \ln X_i)$
均匀分布 $U(0, \theta)$	θ	$T(X) = X$ (单个观测) $T(\mathbf{X}) = X_{(n)} = \max\{X_1, \dots, X_n\}$ (样本)
均匀分布 $U(\theta_1, \theta_2)$	(θ_1, θ_2)	$T(X) = X$ (单个观测) $T(\mathbf{X}) = (X_{(1)}, X_{(n)})$ (样本)

- 对于来自同一分布的独立同分布样本，充分统计量通常是样本的对称函数。
- 指数族分布的充分统计量形式较为规则，通常为 $\sum T(X_i)$ 的形式。
- 均匀分布的充分统计量通常与次序统计量有关。
- 对于多参数情况，充分统计量通常是向量形式。
- 充分统计量的函数如果是一一对应的，则仍然是充分统计量。

9.1.2 Factorisation theorem

在应用充分性原则时，我们经常想要做两件事：

- 为一个参数找一个充分统计量
- 判断一个统计量是否是充分的

定义法只能帮助我们解答部分情况下的第二个问题，下面我们介绍一个重要定理。

定理 9.1 (因子分解定理) 设 $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ 是一个样本，其联合密度或质量函数为 $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$ 。统计量 $\mathbf{U} = h(\mathbf{Y})$ 是参数 $\boldsymbol{\theta}$ 的充分统计量，当且仅当我们可以找到函数 b 和 c ，使得对于所有 $\mathbf{y} \in \mathbb{R}^n$ 和 $\boldsymbol{\theta} \in \Theta$ ，有：

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = b(h(\mathbf{y}), \boldsymbol{\theta})c(\mathbf{y}).$$

例题 9.4 设 X_1, X_2, \dots, X_n 是来自伯努利分布 $B(1, p)$ 的一个简单随机样本，其中 $0 < p < 1$ 。用因子分解定理证明统计量 $T(\mathbf{X}) = \sum_{i=1}^n X_i$ 是参数 p 的充分统计量。

样本的联合概率函数为：

$$f(\mathbf{x}|p) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i}$$

令 $t = \sum_{i=1}^n x_i$ ，则上式可写为：

$$f(\mathbf{x}|p) = p^t (1-p)^{n-t}$$

根据因子分解定理，令：

$$g(t, p) = p^t (1-p)^{n-t}, \quad h(\mathbf{x}) = 1$$

则 $f(\mathbf{x}|p) = g(T(\mathbf{x}), p) \cdot h(\mathbf{x})$ ，其中 $T(\mathbf{x}) = \sum_{i=1}^n x_i$ 。

因此， $T(\mathbf{X}) = \sum_{i=1}^n X_i$ 是 p 的充分统计量。

例题 9.5 设 X_1, X_2, \dots, X_n 是来自泊松分布 $P(\lambda)$ 的一个简单随机样本，其中 $\lambda > 0$ 。用因子分解定理证明统计量 $T(\mathbf{X}) = \sum_{i=1}^n X_i$ 是参数 λ 的充分统计量。

样本的联合概率函数为：

$$f(\mathbf{x}|\lambda) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

令 $t = \sum_{i=1}^n x_i$ ，则：

$$f(\mathbf{x}|\lambda) = e^{-n\lambda} \lambda^t \cdot \frac{1}{\prod_{i=1}^n x_i!}$$

根据因子分解定理，令：

$$g(t, \lambda) = e^{-n\lambda} \lambda^t, \quad h(\mathbf{x}) = \frac{1}{\prod_{i=1}^n x_i!}$$

则 $f(\mathbf{x}|\lambda) = g(T(\mathbf{x}), \lambda) \cdot h(\mathbf{x})$ ，其中 $T(\mathbf{x}) = \sum_{i=1}^n x_i$ 。

因此， $T(\mathbf{X}) = \sum_{i=1}^n X_i$ 是 λ 的充分统计量。

例题 9.6 设 X_1, X_2, \dots, X_n 是来自正态分布 $N(\mu, \sigma_0^2)$ 的一个简单随机样本，其中 μ 未知， σ_0^2 已知。用因子分解定理证明样本均值 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 是参数 μ 的充分统计量。

样本的联合密度函数为：

$$f(\mathbf{x}|\mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma_0^2}\right) = (2\pi\sigma_0^2)^{-n/2} \exp\left(-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu)^2\right)$$

将 $\sum_{i=1}^n (x_i - \mu)^2$ 展开：

$$\begin{aligned} \sum_{i=1}^n (x_i - \mu)^2 &= \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \mu)^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + 2(\bar{x} - \mu) \sum_{i=1}^n (x_i - \bar{x}) + n(\bar{x} - \mu)^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \end{aligned}$$

其中 $\sum_{i=1}^n (x_i - \bar{x}) = 0$ 。

代入得：

$$f(\mathbf{x}|\mu) = (2\pi\sigma_0^2)^{-n/2} \exp\left(-\frac{1}{2\sigma_0^2} \left[\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \right]\right)$$

整理为：

$$f(\mathbf{x}|\mu) = \exp\left(-\frac{n(\bar{x} - \mu)^2}{2\sigma_0^2}\right) \cdot \left[(2\pi\sigma_0^2)^{-n/2} \exp\left(-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \bar{x})^2\right) \right]$$

根据因子分解定理，令：

$$g(\bar{x}, \mu) = \exp\left(-\frac{n(\bar{x} - \mu)^2}{2\sigma_0^2}\right), \quad h(\mathbf{x}) = (2\pi\sigma_0^2)^{-n/2} \exp\left(-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \bar{x})^2\right)$$

则 $f(\mathbf{x}|\mu) = g(T(\mathbf{x}), \mu) \cdot h(\mathbf{x})$ ，其中 $T(\mathbf{x}) = \bar{x}$ 。

因此， \bar{X} 是 μ 的充分统计量。

例题 9.7 设 X_1, X_2, \dots, X_n 是来自正态分布 $N(\mu, \sigma^2)$ 的样本，其中 μ 和 σ^2 均未知。求参数 $\theta = (\mu, \sigma^2)$ 的充分统计量。

样本的联合密度函数为：

$$\begin{aligned} f(\mathbf{x}|\mu, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right) \end{aligned}$$

展开平方和：

$$\begin{aligned} \sum_{i=1}^n (x_i - \mu)^2 &= \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \mu)^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \end{aligned}$$

代入得：

$$\begin{aligned} f(\mathbf{x}|\mu, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right]\right) \\ &= \exp\left(-\frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right) \cdot (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i - \bar{x})^2\right) \end{aligned}$$

根据因子分解定理，令：

$$T(\mathbf{X}) = \left(\bar{X}, \sum_{i=1}^n(X_i - \bar{X})^2\right), \quad g(t_1, t_2; \mu, \sigma^2) = \exp\left(-\frac{n(t_1 - \mu)^2}{2\sigma^2}\right) (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{t_2}{2\sigma^2}\right)$$

其中 $t_1 = \bar{x}$, $t_2 = \sum_{i=1}^n(x_i - \bar{x})^2$, 而 $h(\mathbf{x}) = 1$ 。

因此, $T(\mathbf{X}) = (\bar{X}, \sum_{i=1}^n(X_i - \bar{X})^2)$ 是 (μ, σ^2) 的充分统计量。

例题 9.8 设 X_1, X_2, \dots, X_n 是来自两参数指数分布 $f(x; \mu, \lambda) = \lambda e^{-\lambda(x-\mu)}$ 的样本, 其中 $x \geq \mu$, $\mu \in \mathbb{R}$, $\lambda > 0$ 。求参数 (μ, λ) 的充分统计量。

样本的联合密度函数为：

$$f(\mathbf{x}|\mu, \lambda) = \prod_{i=1}^n \lambda e^{-\lambda(x_i - \mu)} = \lambda^n \exp\left(-\lambda \sum_{i=1}^n(x_i - \mu)\right), \quad x_i \geq \mu$$

由于 $x_i \geq \mu$ 对所有 i 成立, 等价于 $\min\{x_1, \dots, x_n\} \geq \mu$, 即 $\mu \leq x_{(1)}$ 。

重写联合密度：

$$f(\mathbf{x}|\mu, \lambda) = \lambda^n \exp\left(-\lambda \sum_{i=1}^n x_i + n\lambda\mu\right) \cdot I_{[\mu, \infty)}(x_{(1)})$$

其中 $I_A(x)$ 是示性函数。

根据因子分解定理, 令：

$$T(\mathbf{X}) = \left(X_{(1)}, \sum_{i=1}^n X_i\right), \quad g(t_1, t_2; \mu, \lambda) = \lambda^n \exp(-\lambda t_2 + n\lambda\mu) I_{[\mu, \infty)}(t_1)$$

其中 $t_1 = x_{(1)}$, $t_2 = \sum_{i=1}^n x_i$, 而 $h(\mathbf{x}) = 1$ 。

因此, $T(\mathbf{X}) = (X_{(1)}, \sum_{i=1}^n X_i)$ 是 (μ, λ) 的充分统计量。

基于似然的推断与充分性之间存在根本联系。下面的命题描述了这种联系的一个方面。

定理 9.2 (最大似然估计与充分性的关系) 最大似然估计 $\hat{\theta}$ 是 θ 的每一个充分统计量的函数。

证明 9.1 由因子分解定理, 似然函数可写为：

$$L(\theta; \mathbf{x}) = g(T(\mathbf{x}); \theta) \cdot h(\mathbf{x})$$

由于 $h(\mathbf{x})$ 不依赖于 θ , 最大化 $L(\theta; \mathbf{x})$ 等价于最大化 $g(T(\mathbf{x}); \theta)$ 。因此, MLE 只通过 $T(\mathbf{x})$ 依赖于数据。

这个定理的直接推论就是如果 MLE 存在, 它不会使用比充分统计量更多的样本信息。这保证了 MLE 具有数据缩减 (data reduction) 的良好性质。

定理 9.3 (似然比与充分性) 考虑一个由参数 θ 参数化的样本 \mathbf{Y} , 并假设我们感兴趣于检验 $H_0 : \theta \in \Theta_0$ 。如果 \mathbf{U} 是 θ 的充分统计量, 那么基于 \mathbf{U} 的似然比统计量与基于 \mathbf{Y} 的似然比统计量相同。

例题 9.9 设 $X_1, \dots, X_n \sim N(\mu, \sigma_0^2)$, σ_0^2 已知。

充分统计量: $T(\mathbf{X}) = \bar{X}$

似然函数:

$$L(\mu; \mathbf{x}) = (2\pi\sigma_0^2)^{-n/2} \exp\left(-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu)^2\right)$$

由充分统计量性质, 只需考虑基于 \bar{x} 的似然函数:

$$L(\mu; \bar{x}) \propto \exp\left(-\frac{n(\bar{x} - \mu)^2}{2\sigma_0^2}\right)$$

最大化上式得 $\hat{\mu}_{MLE} = \bar{x}$, 计算大大简化。

9.1.3 Minimal sufficiency

找到一个充分统计量是容易的。以 $\mathbf{Y} = \mathbf{y}$ 为条件时 \mathbf{Y} 的分布是退化的, 其全部质量集中在 \mathbf{y} 上。因此, 它不依赖于任何参数。我们得出结论: 整个样本 \mathbf{Y} 对于任何参数都是一个充分统计量。显然, 对于一个作为点估计量有用的统计量, 它必须为我们提供某种程度的数据缩减, 即它的维数必须小于样本的维数。理想情况下, 我们希望定义一个充分统计量, 它允许在不损失关于参数信息的前提下实现最大程度的数据缩减。这就是定义**最小充分性**的基础。在我们给出正式定义之前, 澄清一些与样本函数相关的概念是很重要的。

数据缩减涉及对样本取函数, 即计算统计量。一个函数的输出所包含的信息不可能比其输入更多。例如, 如果

$$\bar{Y} = h(\mathbf{Y}) = \frac{1}{n} \sum_{i=1}^n Y_i,$$

那么 \bar{Y} 包含的信息比 \mathbf{Y} 少。统计量可以是其他统计量的函数。例如, 如果 \mathbf{V} 是二维统计量

$$\mathbf{V} = (V_1, V_2) = \left(Y_1, \sum_{i=2}^n Y_i \right),$$

那么显然, \bar{Y} 是 \mathbf{V} 的函数。我们可以写成

$$\bar{Y} = g(\mathbf{V}) = \frac{1}{n}(V_1 + V_2).$$

因此, \bar{Y} 包含的关于样本的信息比 \mathbf{V} 少。我们现在可以证明关于统计量函数与充分性的以下结果。

引理 9.2 (统计量的函数与充分性) 假设 \mathbf{V} 是一个统计量。如果 \mathbf{U} 仅仅是 \mathbf{V} 的函数, 即对于某个 g 有 $\mathbf{U} = g(\mathbf{V})$, 那么:

- i. \mathbf{U} 是一个统计量;
- ii. 如果 \mathbf{U} 是充分的, 那么 \mathbf{V} 是充分的;
- iii. 如果 \mathbf{V} 不是充分的, 那么 \mathbf{U} 不是充分的;
- iv. 如果 g 是一一对应的并且 \mathbf{V} 是充分的, 那么 \mathbf{U} 是充分的。

定义 9.2 (最小充分统计量) 一个充分统计量 \mathbf{U} 是一个**最小充分统计量**, 如果对于任何其他充分统计量 \mathbf{V} , 都有 \mathbf{U} 是 \mathbf{V} 的函数。

然而, 直接验证最小充分性的条件是困难的, 不过以下观察有时会有所帮助。

1. 如果一个充分统计量是标量（即维度为一），那么它必定是一个最小充分统计量。
2. 最小充分统计量不是唯一的。然而，如果两个统计量都是最小充分的，那么它们必须具有相同的维度。
3. 最小充分统计量的一一对应函数也是最小充分统计量。
4. 需要谨慎，因为最小充分统计量的维度并不总是与感兴趣参数的维度相同。

在思考最小充分性时，考虑由统计量生成的划分是有用的。假设 $\mathbf{U} = h(\mathbf{Y})$ 是一个统计量，且 $h : S \rightarrow T$ ，其中 S 和 T 是某些集合。像任何函数一样， h 在其定义域 S 上定义了一个划分；如果 $h(\mathbf{x}) = h(\mathbf{y})$ ，我们说 \mathbf{x} 和 \mathbf{y} 属于该划分的同一个元素。这个划分的元素可以用值域 T 来索引，所以对于 $t \in T$,

$$A_t = \{\mathbf{y} \in S : h(\mathbf{y}) = t\}.$$

一个最小充分统计量是指其关联的划分是尽可能粗糙的充分统计量。为了理解这一点，考虑某个其他充分统计量 $\mathbf{V} = r(\mathbf{Y})$ 。如果与 \mathbf{U} 关联的划分是尽可能粗糙的，那么任何给出相同 \mathbf{V} 值的两个点也必须给出相同的 \mathbf{U} 值，即

$$r(\mathbf{y}) = r(\mathbf{x}) \Rightarrow h(\mathbf{y}) = h(\mathbf{x}).$$

因此，我们可以找到一个函数 g ，使得对于所有 $\mathbf{y} \in S$ 都有 $h(\mathbf{y}) = g(r(\mathbf{y}))$ 。我们得出结论： $\mathbf{U} = g(\mathbf{V})$ ，并且 \mathbf{U} 是最小充分的。使用充分统计量进行划分的思想构成了以下命题证明的基础。

定理 9.4 (最小充分性的刻画) 考虑一个样本 $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ ，其联合质量/密度函数为 $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$ 。如果我们能找到一个函数 h ，使得

$$h(\mathbf{y}) = h(\mathbf{x}) \Leftrightarrow f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = k(\mathbf{y}, \mathbf{x})f_{\mathbf{Y}}(\mathbf{x}; \boldsymbol{\theta}),$$

那么 $h(\mathbf{Y})$ 是 $\boldsymbol{\theta}$ 的一个最小充分统计量。这里 $k(\mathbf{y}, \mathbf{x})$ 不依赖于 $\boldsymbol{\theta}$ 。

以下是该定理的等价表述：

定理 9.5 (判别函数方法) 设似然比

$$\Lambda(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) = \frac{f(\mathbf{x}; \boldsymbol{\theta})}{f(\mathbf{y}; \boldsymbol{\theta})}$$

如果 $\Lambda(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$ 与 $\boldsymbol{\theta}$ 无关当且仅当 $T(\mathbf{x}) = T(\mathbf{y})$ ，则 T 是最小充分统计量。

例题 9.10 设 X_1, X_2, \dots, X_n 是来自正态分布 $N(\mu, \sigma^2)$ 的样本，其中 μ 和 σ^2 均未知。证明 $T(\mathbf{X}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ 是最小充分统计量。

样本的联合密度函数为：

$$f(\mathbf{x}; \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right)$$

考虑似然比：

$$\begin{aligned} \Lambda(\mathbf{x}, \mathbf{y}; \mu, \sigma^2) &= \frac{f(\mathbf{x}; \mu, \sigma^2)}{f(\mathbf{y}; \mu, \sigma^2)} \\ &= \exp\left(-\frac{1}{2\sigma^2} \left[\sum_{i=1}^n (x_i - \mu)^2 - \sum_{i=1}^n (y_i - \mu)^2 \right]\right) \end{aligned}$$

展开平方项：

$$\begin{aligned}\sum_{i=1}^n (x_i - \mu)^2 &= \sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2 \\ \sum_{i=1}^n (y_i - \mu)^2 &= \sum_{i=1}^n y_i^2 - 2\mu \sum_{i=1}^n y_i + n\mu^2\end{aligned}$$

因此：

$$\Lambda(\mathbf{x}, \mathbf{y}; \mu, \sigma^2) = \exp \left(-\frac{1}{2\sigma^2} \left[\left(\sum x_i^2 - \sum y_i^2 \right) - 2\mu \left(\sum x_i - \sum y_i \right) \right] \right)$$

如果 $\sum x_i = \sum y_i$ 且 $\sum x_i^2 = \sum y_i^2$ ，则 $\Lambda(\mathbf{x}, \mathbf{y}; \mu, \sigma^2) = 1$ 与参数无关。

反之，如果 $\Lambda(\mathbf{x}, \mathbf{y}; \mu, \sigma^2)$ 与 (μ, σ^2) 无关，则对任意 $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ 有：

$$\frac{1}{\sigma_1^2} \left[\left(\sum x_i^2 - \sum y_i^2 \right) - 2\mu_1 \left(\sum x_i - \sum y_i \right) \right] = \frac{1}{\sigma_2^2} \left[\left(\sum x_i^2 - \sum y_i^2 \right) - 2\mu_2 \left(\sum x_i - \sum y_i \right) \right]$$

这要求 $\sum x_i = \sum y_i$ 且 $\sum x_i^2 = \sum y_i^2$ 。

因此， $T(\mathbf{X}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ 是最小充分统计量。

例题 9.11 设 X_1, X_2, \dots, X_n 是来自泊松分布 $P(\lambda)$ 的样本。证明 $T(\mathbf{X}) = \sum_{i=1}^n X_i$ 是最小充分统计量。

样本的联合概率函数为：

$$f(\mathbf{x}; \lambda) = \frac{e^{-n\lambda} \lambda^{\sum x_i}}{\prod_{i=1}^n x_i!}$$

似然比为：

$$\Lambda(\mathbf{x}, \mathbf{y}; \lambda) = \frac{f(\mathbf{x}; \lambda)}{f(\mathbf{y}; \lambda)} = \lambda^{\sum x_i - \sum y_i} \cdot \frac{\prod y_i!}{\prod x_i!}$$

如果 $\sum x_i = \sum y_i$ ，则 $\Lambda(\mathbf{x}, \mathbf{y}; \lambda) = \frac{\prod y_i!}{\prod x_i!}$ 与 λ 无关。

反之，如果 $\Lambda(\mathbf{x}, \mathbf{y}; \lambda)$ 与 λ 无关，则 $\lambda^{\sum x_i - \sum y_i}$ 必须为常数，这要求 $\sum x_i = \sum y_i$ 。

因此， $T(\mathbf{X}) = \sum_{i=1}^n X_i$ 是最小充分统计量。

例题 9.12 设 X_1, X_2, \dots, X_n 是来自伯努利分布 $B(1, p)$ 的样本。证明 $T(\mathbf{X}) = \sum_{i=1}^n X_i$ 是最小充分统计量。

样本的联合概率函数为：

$$f(\mathbf{x}; p) = p^{\sum x_i} (1-p)^{n-\sum x_i}$$

似然比为：

$$\Lambda(\mathbf{x}, \mathbf{y}; p) = \frac{f(\mathbf{x}; p)}{f(\mathbf{y}; p)} = \left(\frac{p}{1-p} \right)^{\sum x_i - \sum y_i}$$

如果 $\sum x_i = \sum y_i$ ，则 $\Lambda(\mathbf{x}, \mathbf{y}; p) = 1$ 与 p 无关。

反之，如果 $\Lambda(\mathbf{x}, \mathbf{y}; p)$ 与 p 无关，则 $\left(\frac{p}{1-p} \right)^{\sum x_i - \sum y_i}$ 必须为常数。由于 $\frac{p}{1-p}$ 可以取 $(0, \infty)$ 内的任意值，这要求 $\sum x_i = \sum y_i$ 。

因此， $T(\mathbf{X}) = \sum_{i=1}^n X_i$ 是最小充分统计量。

9.1.4 Application of sufficiency in point estimation

在之前的小节中，我们讨论了均方误差作为估计量质量的度量；我们的目标是找到具有尽可能小 MSE 的估计量。充分性原则指出，我们应该基于充分统计量进行推断。在本节中，我们证明在选择估计量时，可以将充分性原则与对低 MSE 的追求结合起来。基本思想如下：假设 U 是 θ 的一个估计量。事实证明，我们可以找到一个关于 θ 的充分统计量的函数，该函数在 MSE 意义上至少与 U 一样好。这在某种意义上是一份免费的午餐；我们可以在不牺牲均方误差性能的情况下满足充分性原则。

该结果基于以下引理。这个引理相当明显，常常被视为理所当然。我们在这里包含它是为了强调充分性的作用。

引理 9.3 (以充分统计量为条件) 假设 \mathbf{Y} 是一个样本， U 是一个统计量，且 \mathbf{S} 是 θ 的充分统计量。如果我们定义 $T = \mathbb{E}(U|\mathbf{S})$ ，那么 T 是一个统计量，即 T 不依赖于 θ 。

我们现在可以证明 Rao-Blackwell 定理。该结果表明，对于任何点估计量 U ，我们都可以找到另一个点估计量，它是充分统计量的函数，并且在均方误差意义上至少与 U 一样好。

定理 9.6 (Rao-Blackwell 定理) 假设 \mathbf{Y} 是一个样本， U 是一个与 θ 维度相同的统计量，且 \mathbf{S} 是 θ 的充分统计量。如果我们定义统计量 $T = \mathbb{E}(U|\mathbf{S})$ ，那么 $\text{MSE}_\theta(T) \leq \text{MSE}_\theta(U)$ 。

Rao-Blackwell 定理的一个推论是，我们可以通过取任何点估计量关于一个充分统计量的条件期望来改进（或至少不恶化）它。注意偏差保持不变，因为根据迭代期望法则有：

$$\mathbb{E}(T) = \mathbb{E}[\mathbb{E}(U|\mathbf{S})] = \mathbb{E}(U).$$

定理 9.7 (Rao-Blackwell 定理的无偏版本) 设 $\hat{\theta}$ 是参数 θ 的一个无偏估计， T 是 θ 的充分统计量。定义 $\hat{\theta}^* = E[\hat{\theta} | T]$ ，则：

1. $\hat{\theta}^*$ 是 θ 的无偏估计
2. $\text{Var}(\hat{\theta}^*) \leq \text{Var}(\hat{\theta})$ ，等号成立当且仅当 $P(\hat{\theta} = \hat{\theta}^*) = 1$

例题 9.13 设 X_1, X_2, \dots, X_n 是来自伯努利分布 $B(1, p)$ 的样本，其中 $0 < p < 1$ 。考虑初始估计 $\hat{p} = X_1$ ，使用 Rao-Blackwell 定理改进这个估计。

步骤 1：验证充分统计量 已知 $T = \sum_{i=1}^n X_i$ 是 p 的充分统计量，且 $T \sim B(n, p)$ 。

步骤 2：计算条件期望 我们需要计算 $E[X_1 | T = t]$ 。

由于样本的对称性，对于任意 i, j ，有：

$$E[X_i | T = t] = E[X_j | T = t]$$

因此：

$$\sum_{i=1}^n E[X_i | T = t] = E \left[\sum_{i=1}^n X_i | T = t \right] = E[T | T = t] = t$$

所以：

$$E[X_1 | T = t] = \frac{t}{n}$$

步骤 3：得到改进估计

$$\hat{p}^* = E[X_1 | T] = \frac{T}{n} = \bar{X}$$

步骤 4：验证改进效果

- 无偏性: $E[\hat{p}^*] = E[\bar{X}] = p$

- 方差比较:

$$\begin{aligned}\text{Var}(\hat{p}) &= \text{Var}(X_1) = p(1-p) \\ \text{Var}(\hat{p}^*) &= \text{Var}(\bar{X}) = \frac{p(1-p)}{n}\end{aligned}$$

显然 $\text{Var}(\hat{p}^*) < \text{Var}(\hat{p})$ 当 $n > 1$ 。

例题 9.14 设 X_1, X_2, \dots, X_n 是来自泊松分布 $P(\lambda)$ 的样本。考虑初始估计 $\hat{\lambda} = X_1$, 使用 Rao-Blackwell 定理改进这个估计。

步骤 1: 充分统计量 $T = \sum_{i=1}^n X_i$ 是 λ 的充分统计量, 且 $T \sim P(n\lambda)$ 。

步骤 2: 计算条件分布 首先求 X_1 在给定 $T = t$ 下的条件分布:

$$\begin{aligned}P(X_1 = x \mid T = t) &= \frac{P(X_1 = x, T = t)}{P(T = t)} \\ &= \frac{P(X_1 = x)P(\sum_{i=2}^n X_i = t - x)}{P(T = t)} \\ &= \frac{\frac{e^{-\lambda} \lambda^x}{x!} \cdot \frac{e^{-(n-1)\lambda} [(n-1)\lambda]^{t-x}}{(t-x)!}}{\frac{e^{-n\lambda} (n\lambda)^t}{t!}} \\ &= \frac{t!}{x!(t-x)!} \left(\frac{1}{n}\right)^x \left(1 - \frac{1}{n}\right)^{t-x}\end{aligned}$$

即 $X_1 \mid T = t \sim B(t, \frac{1}{n})$ 。

步骤 3: 计算条件期望

$$E[X_1 \mid T = t] = t \cdot \frac{1}{n} = \frac{t}{n}$$

步骤 4: 得到改进估计

$$\hat{\lambda}^* = E[X_1 \mid T] = \frac{T}{n} = \bar{X}$$

步骤 5: 验证改进效果

- 无偏性: $E[\hat{\lambda}^*] = E[\bar{X}] = \lambda$

- 方差比较:

$$\begin{aligned}\text{Var}(\hat{\lambda}) &= \text{Var}(X_1) = \lambda \\ \text{Var}(\hat{\lambda}^*) &= \text{Var}(\bar{X}) = \frac{\lambda}{n}\end{aligned}$$

改进后的估计方差更小。

例题 9.15 设 X_1, X_2, \dots, X_n 是来自正态分布 $N(\mu, 1)$ 的样本。考虑初始估计 $\hat{\mu} = \frac{X_1+X_2}{2}$, 使用 Rao-Blackwell 定理改进这个估计。

步骤 1: 充分统计量 $T = \sum_{i=1}^n X_i$ 是 μ 的充分统计量。

步骤 2: 计算条件期望 我们需要计算 $E\left[\frac{X_1+X_2}{2} \mid T = t\right]$ 。

由对称性:

$$E[X_1 \mid T = t] = E[X_2 \mid T = t] = \frac{t}{n}$$

因此：

$$E\left[\frac{X_1 + X_2}{2} \mid T = t\right] = \frac{1}{2}\left(\frac{t}{n} + \frac{t}{n}\right) = \frac{t}{n}$$

步骤 3：得到改进估计

$$\hat{\mu}^* = \frac{T}{n} = \bar{X}$$

步骤 4：验证改进效果

- 无偏性：两者都是无偏估计
- 方差比较：

$$\begin{aligned} \text{Var}(\hat{\mu}) &= \text{Var}\left(\frac{X_1 + X_2}{2}\right) = \frac{1}{4}(1+1) = \frac{1}{2} \\ \text{Var}(\hat{\mu}^*) &= \text{Var}(\bar{X}) = \frac{1}{n} \end{aligned}$$

当 $n > 2$ 时， $\frac{1}{n} < \frac{1}{2}$ ，方差减小。

9.2 Variance of unbiased estimators

无偏估计量具有一个吸引人的性质，即它们的期望值就是参数的真实值。如果我们将注意力限制在无偏估计量上，一个显而易见的目标是尝试找到具有最小方差的估计量，即 **最小方差无偏估计量** (MVUE)。我们首先证明一个结果，该结果描述了我们期望从一个无偏估计量中获得的最佳性能。在接下来的内容中，我们将假设 $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ 是一个联合密度为 $f_{\mathbf{Y}}(\mathbf{y}; \theta)$ 的样本，其中 θ 是一个标量参数； θ 是参数空间 Θ 中的某个值。

定理 9.8 (Cramér-Rao 下界) 设 $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ 是一个由标量参数 θ 参数化的样本，且 $U = h(\mathbf{Y})$ 是参数函数 $g(\theta)$ 的一个无偏估计量。在正则性条件下（允许我们在积分号下求导），对于任何 $\theta \in \Theta$ ，有：

$$\text{Var}(U) \geq \frac{\left[\frac{d}{d\theta} g(\theta)\right]^2}{I_{\mathbf{Y}}(\theta)},$$

其中 $I_{\mathbf{Y}}(\theta)$ 是样本 \mathbf{Y} 中关于 θ 的总 Fisher 信息量。

Cramér-Rao 定理的一般形式除了对联合分布的正则性有一些要求外，不对样本做任何假设。在实践中，我们通常假设样本是随机的。这导致无偏估计量方差下界的计算大为简化。回顾之前的发现，对于一个随机样本 $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ ，有：

$$I_{\mathbf{Y}}(\theta) = n I_Y(\theta),$$

其中 $I_{\mathbf{Y}}(\theta)$ 是总信息量，而 $I_Y(\theta)$ 是与单次观测相关的信息量（这里向量 \mathbf{Y} 和标量 Y 之间的区别很重要）。这直接引出了以下推论，该推论为基于随机样本的无偏估计量的方差建立了一个下界。

定理 9.9 如果 $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ 是一个随机样本，且 $U = h(\mathbf{Y})$ 是 $g(\theta)$ 的一个无偏估计量，那么：

$$\text{Var}(U) \geq \frac{\left[\frac{d}{d\theta} g(\theta)\right]^2}{n I_Y(\theta)},$$

其中 $I_Y(\theta)$ 是单个样本成员中关于 θ 的 Fisher 信息量。

当然，在许多情况下，我们感兴趣的是估计 θ 本身，而不是它的某个函数。这导致下界的进一步简化。特别地，如果 $\hat{\theta}$ 是基于随机样本 $(Y_1, \dots, Y_n)^T$ 的 θ 的一个无偏估计量，那么：

$$\text{Var}(\hat{\theta}) \geq \frac{1}{nI_Y(\theta)}.$$

以上定理为无偏估计量的方差建立了一个下界。一个明显的问题是，这个下界是否能够达到。如果一个统计量 U 达到了该下界，即如果 U 是 $g(\theta)$ 的一个无偏估计量并且：

$$\text{Var}(U) = \frac{\left[\frac{d}{d\theta}g(\theta)\right]^2}{I_Y(\theta)},$$

那么 U 必定是 $g(\theta)$ 的最小方差无偏估计量。然而，反之则不成立；存在最小方差无偏估计量未能达到 Cramér-Rao 下界的情况。需要附加条件来确保达到该下界。

定理 9.10 (达到 Cramér-Rao 下界) 如果 $U = h(\mathbf{Y})$ 是 $g(\theta)$ 的一个无偏估计量，那么 U 达到 Cramér-Rao 下界当且仅当：

$$s(\theta; \mathbf{y}) = b(\theta)[h(\mathbf{y}) - g(\theta)],$$

其中 $b(\theta)$ 是一个涉及参数 θ 但不涉及 \mathbf{y} 的函数。

到目前为止，我们考虑的是标量参数 θ 。对于向量参数 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_r)^T$ ，估计量的方差 $\text{Var}(\mathbf{U})$ 和信息矩阵 $\mathbf{I}_Y(\boldsymbol{\theta})$ 都是 $r \times r$ 矩阵。

定理 9.11 (向量参数的 Cramér-Rao 下界) 如果 $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ 是一个由参数 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_r)^T$ 参数化的随机样本，且 \mathbf{U} 是 $\boldsymbol{\theta}$ 的一个无偏估计量，那么：

$$\text{Var}(\mathbf{U}) \geq \{\mathbf{I}_Y(\boldsymbol{\theta})\}^{-1},$$

这里的含义是 $\text{Var}(\mathbf{U}) - \{\mathbf{I}_Y(\boldsymbol{\theta})\}^{-1}$ 是一个非负定矩阵。

例题 9.16 设 X_1, X_2, \dots, X_n 是来自正态分布 $N(\mu, \sigma^2)$ 的样本，其中 σ^2 已知。求 μ 的 Cramér-Rao 下界，并验证样本均值 \bar{X} 达到该下界。

步骤 1：计算对数似然函数单个观测的密度函数：

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

对数似然：

$$\ln f(x; \mu) = -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

步骤 2：计算一阶导数

$$\frac{\partial}{\partial \mu} \ln f(x; \mu) = \frac{x-\mu}{\sigma^2}$$

步骤 3：计算 Fisher 信息量

$$\begin{aligned} I(\mu) &= E\left[\left(\frac{\partial}{\partial \mu} \ln f(X; \mu)\right)^2\right] \\ &= E\left[\left(\frac{X-\mu}{\sigma^2}\right)^2\right] = \frac{1}{\sigma^4} E[(X-\mu)^2] = \frac{1}{\sigma^4} \cdot \sigma^2 = \frac{1}{\sigma^2} \end{aligned}$$

步骤 4: 计算 Cramér-Rao 下界

$$\text{Var}(\hat{\mu}) \geq \frac{1}{nI(\mu)} = \frac{\sigma^2}{n}$$

步骤 5: 验证样本均值达到下界 样本均值 \bar{X} 的方差:

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

正好等于 Cramér-Rao 下界, 因此 \bar{X} 是有效估计。

例题 9.17 设 X_1, X_2, \dots, X_n 是来自泊松分布 $P(\lambda)$ 的样本。求 λ 的 Cramér-Rao 下界, 并验证样本均值达到该下界。

步骤 1: 计算对数似然函数

$$f(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad \ln f(x; \lambda) = -\lambda + x \ln \lambda - \ln(x!)$$

步骤 2: 计算一阶导数

$$\frac{\partial}{\partial \lambda} \ln f(x; \lambda) = -1 + \frac{x}{\lambda}$$

步骤 3: 计算 Fisher 信息量

$$\begin{aligned} I(\lambda) &= E \left[\left(\frac{\partial}{\partial \lambda} \ln f(X; \lambda) \right)^2 \right] \\ &= E \left[\left(-1 + \frac{X}{\lambda} \right)^2 \right] = E \left[\left(\frac{X - \lambda}{\lambda} \right)^2 \right] \\ &= \frac{1}{\lambda^2} E[(X - \lambda)^2] = \frac{1}{\lambda^2} \cdot \lambda = \frac{1}{\lambda} \end{aligned}$$

步骤 4: 计算 Cramér-Rao 下界

$$\text{Var}(\hat{\lambda}) \geq \frac{1}{nI(\lambda)} = \frac{\lambda}{n}$$

步骤 5: 验证样本均值达到下界 样本均值 \bar{X} 的方差:

$$\text{Var}(\bar{X}) = \frac{\lambda}{n}$$

达到 Cramér-Rao 下界, 因此是有效估计。

例题 9.18 设 X_1, X_2, \dots, X_n 是来自伯努利分布 $B(1, p)$ 的样本。求 p 的 Cramér-Rao 下界, 并验证样本比例达到该下界。

步骤 1: 计算对数似然函数

$$f(x; p) = p^x (1-p)^{1-x}, \quad \ln f(x; p) = x \ln p + (1-x) \ln(1-p)$$

步骤 2: 计算一阶导数

$$\frac{\partial}{\partial p} \ln f(x; p) = \frac{x}{p} - \frac{1-x}{1-p}$$

步骤 3: 计算 Fisher 信息量

$$\begin{aligned}
I(p) &= E \left[\left(\frac{\partial}{\partial p} \ln f(X; p) \right)^2 \right] \\
&= E \left[\left(\frac{X}{p} - \frac{1-X}{1-p} \right)^2 \right] \\
&= E \left[\left(\frac{X-p}{p(1-p)} \right)^2 \right] = \frac{1}{p^2(1-p)^2} E[(X-p)^2] \\
&= \frac{1}{p^2(1-p)^2} \cdot p(1-p) = \frac{1}{p(1-p)}
\end{aligned}$$

步骤 4: 计算 Cramér-Rao 下界

$$\text{Var}(\hat{p}) \geq \frac{1}{n I(p)} = \frac{p(1-p)}{n}$$

步骤 5: 验证样本比例达到下界样本比例 $\hat{p} = \frac{1}{n} \sum X_i$ 的方差:

$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

达到 Cramér-Rao 下界, 因此是有效估计。

例题 9.19 设 X_1, X_2, \dots, X_n 是来自正态分布 $N(\mu, \sigma^2)$ 的样本, 两者均未知。求 σ^2 的 Cramér-Rao 下界, 并与样本方差比较。

步骤 1: 计算对数似然函数

$$\ln f(x; \mu, \sigma^2) = -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

步骤 2: 计算关于 σ^2 的导数

$$\frac{\partial}{\partial \sigma^2} \ln f(x; \mu, \sigma^2) = -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4}$$

步骤 3: 计算 Fisher 信息矩阵需要计算信息矩阵 $I(\mu, \sigma^2)$ 中关于 σ^2 的元素:

$$\begin{aligned}
I_{22} &= -E \left[\frac{\partial^2}{\partial (\sigma^2)^2} \ln f(X; \mu, \sigma^2) \right] \\
\frac{\partial^2}{\partial (\sigma^2)^2} \ln f(x; \mu, \sigma^2) &= \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6} \\
I_{22} &= -E \left[\frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6} \right] = - \left[\frac{1}{2\sigma^4} - \frac{\sigma^2}{\sigma^6} \right] = \frac{1}{2\sigma^4}
\end{aligned}$$

步骤 4: 计算 Cramér-Rao 下界在多参数情况下, σ^2 的 Cramér-Rao 下界是信息矩阵逆矩阵的对应元素:

$$\text{Var}(\hat{\sigma}^2) \geq [I^{-1}(\mu, \sigma^2)]_{22} = \frac{2\sigma^4}{n}$$

步骤 5: 与样本方差比较样本方差 $S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$ 是无偏估计, 其方差为:

$$\text{Var}(S^2) = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n}$$

因此 S^2 没有达到 Cramér-Rao 下界。

9.3 Most powerful tests

由于我们需要比较不同的检验，能够使用特定的标识符来指代检验是很有用的。一个检验由其决策函数或等价地由其拒绝域来表征。对于一个检验 S ，我们使用 ϕ_S 表示其决策函数， R_S 表示其拒绝域， β_S 表示其势函数。注意 $\beta_S(\theta)$ 是当 θ 为真实参数值时，检验 S 拒绝 H_0 的概率。下面的引理建立了势函数与决策函数之间的联系，它对于证明本节的主要结果非常有用。

引理 9.4 (决策函数与势函数的关系) 如果 ϕ_S 是假设检验 S 的决策函数，那么其势函数由下式给出：

$$\beta_S(\theta) = \mathbb{E}[\phi_S(\mathbf{Y})].$$

基于势函数来选择检验的理论，其出发点是考虑两个简单假设。假设我们检验：

$$H_0 : \theta = \theta_0,$$

$$H_1 : \theta = \theta_1.$$

在这种情况下，一个检验 S 的势就是其势函数在备择假设指定的值处的取值，即 $\beta_S(\theta_1)$ 。

定义 9.3 (最大势检验) 假设我们检验 $H_0 : \theta = \theta_0$ 相对于 $H_1 : \theta = \theta_1$ 。我们称 T 是一个 **最大势检验**，如果对于所有满足 $\beta_S(\theta_0) = \beta_T(\theta_0)$ 的检验 S ，都有 $\beta_T(\theta_1) \geq \beta_S(\theta_1)$ 。

正如我们所料，上述定义表明，如果一个检验 T 的势大于或等于任何其他相同显著性水平的检验的势，那么它就是最大势检验。我们通常会提到一个 **显著性水平为 α** 的最大势检验；明确写出显著性水平是为了强调其在最大势检验定义中的关键作用。

现在我们可以给出假设检验中最早期的一个结果的表述。该定理的证明是直接的，尽管在某些地方有些微妙。

定理 9.12 (Neyman-Pearson 引理) 假设 T 是检验 $H_0 : \theta = \theta_0$ 相对于 $H_1 : \theta = \theta_1$ 的一个显著性水平为 α 的检验。如果 T 的拒绝域为：

$$R_T = \{\mathbf{y} \in \mathbb{R}^n : L_{\mathbf{Y}}(\theta_1; \mathbf{y}) - k_{\alpha} L_{\mathbf{Y}}(\theta_0; \mathbf{y}) > 0\},$$

那么 T 就是显著性水平为 α 的最大势检验。注意 $L_{\mathbf{Y}}$ 是似然函数，而 k_{α} 是一个依赖于 α 的常数。

9.4 Further exercises

9.5 Appendix: Proofs

10 Bayesian inference 贝叶斯推断

10.1 Prior and posterior distributions

贝叶斯方法要求我们为参数指定一个边际分布。该分布被称为**先验分布**。围绕贝叶斯统计的许多争议都集中在先验分布上。在其早期形式中，贝叶斯主义与概率的主观解释紧密相关。对于主观主义者来说，概率反映了我们信念的程度，因此不同的人可能理性地将不同的概率赋予同一件事。“先验”这个名字反映了这一渊源；在主观贝叶斯的世界里，先验代表了我们在观测数据之前对参数分布的看法。一旦我们获得了观测数据，先验分布就会使用贝叶斯定理进行更新。由此产生的**后验分布**构成了贝叶斯推断的基础。

一些符号有助于阐明这个过程。我们放弃之前用下标标识分布所属随机变量的惯例。在本章的剩余部分，我们将注意力集中在两个随机变量向量上：

- 样本： $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ ，
- 参数： $\boldsymbol{\Theta} = (\Theta_1, \dots, \Theta_r)^T$ 。

我们使用 f 表示与样本相关的密度函数。与经典的频率学派方法一样，我们假设已知**抽样密度** $f(\mathbf{y}|\boldsymbol{\theta})$ ，即给定 $\boldsymbol{\Theta}$ 取特定值 $\boldsymbol{\theta}$ 时 \mathbf{Y} 的密度。这就是随机变量 $\mathbf{Y}|\boldsymbol{\Theta} = \boldsymbol{\theta}$ 的密度函数。当我们从参数 $\boldsymbol{\theta}$ 的角度看待抽样密度时，我们称 $f(\mathbf{y}|\boldsymbol{\theta})$ 为似然函数。

与参数相关的密度函数使用 π 表示。具体来说，我们有：

- 先验密度： $\pi(\boldsymbol{\theta})$ ，
- 后验密度： $\pi(\boldsymbol{\theta}|\mathbf{y})$ 。

注意，先验密度是 $\boldsymbol{\Theta}$ 的边际密度，而后验密度是随机变量 $\boldsymbol{\Theta}|\mathbf{Y} = \mathbf{y}$ 的密度。我们使用 $\mathbb{E}(\cdot|\mathbf{y})$ ，而不是通常的 $\mathbb{E}(\cdot|\mathbf{Y} = \mathbf{y})$ ，来表示涉及后验密度的期望。

对于 \mathbf{Y} 和 $\boldsymbol{\Theta}$ 的联合密度，我们不引入新的符号，而是使用似然函数（抽样密度）乘以先验密度，即 $f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ 。因此， \mathbf{Y} 的边际密度为：

$$f(\mathbf{y}) = \int_{-\infty}^{\infty} f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

贝叶斯定理提供了先验分布与后验分布之间的联系：

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\mathbf{y})}. \quad (10.1)$$

我们将 $\pi(\boldsymbol{\theta}|\mathbf{y})$ 视为 $\boldsymbol{\theta}$ 的函数，而将 $f(\mathbf{y})$ 视为仅与 \mathbf{y} 有关的函数，因此方程 (10.1) 常被写作：

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \quad (10.2)$$

比例关系 (10.2) 通常被记作“后验正比于似然乘以先验”。

例题 10.1 假设我们希望通过最大化后验分布 $\pi(\boldsymbol{\theta}|\mathbf{y})$ 来估计 $\boldsymbol{\theta}$ 。在什么情况下，由此得到的估计与 $\boldsymbol{\theta}$ 的最大似然估计相同？

最大后验估计是通过最大化后验分布 $\pi(\boldsymbol{\theta}|\mathbf{y})$ 得到的，根据贝叶斯公式：

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{y})\pi(\boldsymbol{\theta}).$$

最大似然估计是通过最大化似然函数 $L(\theta; \mathbf{y})$ 得到的。

当先验分布 $\pi(\theta)$ 是均匀分布时，即 $\pi(\theta) \propto$ 常数，那么后验分布满足：

$$\pi(\theta|\mathbf{y}) \propto L(\theta;\mathbf{y}) \times \text{常数} \propto L(\theta;\mathbf{y}).$$

此时，最大化后验分布等价于最大化似然函数，因此最大后验估计与最大似然估计相同。

结论：当先验分布是无信息先验（均匀分布）时，最大后验估计与最大似然估计相同。

例题 10.2 设 y_1, y_2 是来自 $f(y|\theta)$ 的两个独立随机样本，其中 θ 具有先验分布 $\pi(\theta)$ 。考虑两种可能的情况：(i) 我们先观测到 y_1 ，将先验更新为 $\pi(\theta|y_1)$ ，然后观测到 y_2 并再次更新分布；(ii) 我们同时观测到 y_1, y_2 ，并一步更新先验。证明在两种情况下我们得到相同的 θ 的后验分布。

情况 (i): 分两步更新

- 第一次更新（观测 y_1 ）：

$$\pi(\theta|y_1) \propto f(y_1|\theta)\pi(\theta).$$

- 第二次更新（观测 y_2 ）：

$$\pi(\theta|y_1, y_2) \propto f(y_2|\theta)\pi(\theta|y_1) \propto f(y_2|\theta)f(y_1|\theta)\pi(\theta).$$

情况 (ii): 一步更新 由于 y_1, y_2 独立同分布，且给定 θ 时条件独立，有：

$$f(y_1, y_2|\theta) = f(y_1|\theta)f(y_2|\theta).$$

因此，一步更新的后验为：

$$\pi(\theta|y_1, y_2) \propto f(y_1, y_2|\theta)\pi(\theta) = f(y_1|\theta)f(y_2|\theta)\pi(\theta).$$

比较两种情况：

- 情况 (i) 最终后验： $\pi(\theta|y_1, y_2) \propto f(y_1|\theta)f(y_2|\theta)\pi(\theta)$
- 情况 (ii) 最终后验： $\pi(\theta|y_1, y_2) \propto f(y_1|\theta)f(y_2|\theta)\pi(\theta)$

两者完全一致。这表明在贝叶斯更新中，数据的顺序不影响最终的后验分布，只要所有数据都被用于更新。这个性质体现了贝叶斯更新的顺序一致性。

10.2 Choosing a prior

从方程 (10.1) 可以看出，后验分布的形式依赖于先验分布的选择。我们所选择的某些东西会对我们的推断产生影响，这一事实常被引述为贝叶斯方法的一个弱点。这种批评有些站不住脚；频率学派的推断同样涉及任意选择，例如假设检验中显著性水平的选择。然而，如果我们要进行贝叶斯推断，选择先验分布这个实际问题必须得到解决。如果我们认为先验分布应该反映我们对参数各种可能取值的信念程度，那么就应该通过仔细获取并结合理性专家的意见来选择它。对于存在大量领域特定知识的问题，这种方法是实用（且明智的）。在我们关于先验选择的讨论中，我们假设我们对参数没有任何成熟的看法，可能只是有一个模糊的合理取值范围。我们首先建立一些术语。

实践中使用的一些先验是所谓的非正常先验（improper prior）。这个术语让人联想到一位不太严肃对待誓言的僧侣；贝叶斯学派对该术语的使用有些类似。可以定义那些没有正常密度函数（指

其积分不为 1) 的先验, 但它们仍然能产生正常的后验密度。后验是推断的基础, 因此只要后验具有正常的密度, 我们就不必过分担心先验是否非正常。下面给出了几个非正常先验的例子。

我们可能尝试使贝叶斯分析尽可能客观的一种方法是使用**无信息先验**(non-informative prior)。“无信息”一词在多种语境中使用; 我们用它来表示任何试图将关于参数的主观信息包含降至最低的先验。另一个常用于指代包含信息量很少且被似然函数主导的先验的术语是**参考先验**(reference prior)。在其他情况下, 我们出于数学便利选择特定类型的先验; 这些被称为**共轭先验** (conjugate prior)。我们在下面讨论参考先验和共轭先验的构建。

10.2.1 Constructing reference priors

1. 平坦先验 (flat prior)

传达先验无知的一种可能机制是使用平坦先验。平坦先验在其支撑集上对所有参数值赋予相等的概率。我们可以为合理参数值区间选择端点, 然后使用在该区间上均匀的先验。在某些情况下, 参数可能取值范围是明显的。例如, 如果 $Y|\theta = \theta \sim \text{Bernoulli}(\theta)$, 我们知道 $\theta \in [0, 1]$ 。通常, 采用这种方法需要任意选择区间端点。另一种方法是指定:

$$\pi(\theta) = c \quad \text{对所有 } \theta \text{ 成立.}$$

显然, 除非 θ 的支撑集是一个有限区间, 否则这是一个非正常先验; π 不是一个有效的密度函数, 因为 $\int \pi(\theta) d\theta$ 不是有限的。然而, 我们仍然可以从似然函数与先验的乘积来评估后验密度。在这种情况下, 由于先验是一个固定常数, 后验密度与似然函数成正比:

$$\pi(\theta|y) \propto f(y|\theta).$$

2. 基于数据平移形式的先验 (prior from data-translated forms)

假设我们可以将似然函数表示为**数据平移形式** (data-translated form):

$$f(y|\theta) \propto r(g(\theta) - h(y)),$$

其中 r , g , 和 h 是某些函数。注意这里的比例关系是关于 θ 的, 因此比例常数可能涉及 y 。考虑由 $\psi = g(\theta)$ 给出的重新参数化。数据的影响是相对于 ψ 平移了似然函数。一种可能的选择是对 ψ 使用平坦先验。根据变量变换公式, 隐含的 Θ 的先验具有比例关系:

$$\pi(\theta) \propto \left| \frac{\partial \psi}{\partial \theta} \right|.$$

如果 g 是恒等函数, 即如果 $\psi = \theta$, 那么使用此准则等价于对 θ 指定一个平坦先验。如果似然函数中 θ 和 y 的关系呈现比率形式, 我们仍然可以通过设定下式将似然函数表示为数据平移形式:

$$r\left(\frac{g(\theta)}{h(y)}\right) = r\left(\exp\left\{\log\left(\frac{g(\theta)}{h(y)}\right)\right\}\right) = r\left(\exp\{\log g(\theta) - \log h(y)\}\right).$$

在这种情况下, 我们将取 $\psi = \log g(\theta)$, 而我们对 θ 的先验将是:

$$\pi(\theta) \propto \left| \frac{g'(\theta)}{g(\theta)} \right|.$$

3. 变换不变性与 Jeffreys 准则 (transformation invariance and Jeffreys' rule)

考虑一种情况，我们对原始参数的一个性质良好的单调变换感兴趣。我们将变换后的参数表示为 $\psi = g(\Theta)$ 。 ψ 的隐含先验为：

$$\pi_\psi(\psi) = \pi_\Theta(g^{-1}(\psi)) \left| \frac{\partial}{\partial \psi} g^{-1}(\psi) \right| = \pi_\Theta(\theta) \left| \frac{\partial \theta}{\partial \psi} \right|, \quad (10.3)$$

其中 $\theta = g^{-1}(\psi)$ 。我们直观上合理地希望 π_ψ 能代表与 π_Θ 相同程度的先验知识。如果这一点成立，我们称 π_Θ 为 **变换不变的**。

Jeffreys 准则建议的先验设定通过利用变量变换公式的机制来实现变换不变性。考虑关于参数 ψ 的 Fisher 信息量 $I(\psi)$ 。我们可以写出：

$$\frac{\partial}{\partial \psi} \log f(\mathbf{y}|\theta) = \frac{\partial}{\partial \theta} \log f(\mathbf{y}|\theta) \frac{\partial \theta}{\partial \psi},$$

进而有：

$$-\left(\frac{\partial}{\partial \psi} \log f(\mathbf{y}|\theta) \right)^2 = -\left(\frac{\partial}{\partial \theta} \log f(\mathbf{y}|\theta) \right)^2 \left(\frac{\partial \theta}{\partial \psi} \right)^2,$$

对两边取期望得到：

$$I(\psi) = I(\theta) \left(\frac{\partial \theta}{\partial \psi} \right)^2. \quad (10.4)$$

Θ 的 **Jeffreys 先验**则被指定为：

$$\pi_\Theta(\theta) = \sqrt{|I(\theta)|}. \quad (10.5)$$

比较方程 (10.3) 和 (10.4)，我们可以看到参数变换后版本的隐含先验也是 Jeffreys 先验，因此 π_Θ 是变换不变的。如果 Θ 是一个参数向量， $|I(\theta)|$ 是 Fisher 信息矩阵的行列式，表达式 (10.5) 给出了联合 Jeffreys 先验。

10.2.2 Conjugate priors

对于给定的似然函数 $f(\mathbf{y}|\theta)$ ，我们有时可以以一种特定的方式指定先验分布，使得先验和后验都属于同一个参数族。在这种情况下，我们称 $\pi(\theta)$ 是 $f(\mathbf{y}|\theta)$ 的一个**共轭先验 (conjugate prior)**。选择这种先验的主要动机是数学上的便利性，然而，使用共轭先验通常会产生直观上令人满意的结果。先验 $\pi(\theta)$ 可能依赖于未知参数，这些参数被称为**超参数 (hyperparameters)**。我们目前假设这些超参数是已知的。

10.3 Bayesian estimation

10.3.1 Point estimators

在贝叶斯分析中，用于推断的关键量是**后验分布** (posterior distribution)，即给定数据时参数的条件分布。在实践中，能够生成点估计量是很有用的。例如，我们可能想知道后验分布的中心在哪里。贝叶斯分析中点估计量的明显选择是**后验均值** (posterior mean)、**后验中位数** (posterior median) 和**后验众数** (posterior mode)。

例如，后验均值为：

$$\hat{\theta} = h(Y) = \mathbb{E}(\theta|Y),$$

其中

$$h(y) = \mathbb{E}(\theta|Y = y) = \int_{-\infty}^{\infty} \theta \pi(\theta|y) d\theta.$$

更一般地，贝叶斯参数估计基于**损失函数** (loss function) 的概念。损失函数记为 $L(\theta, a)$ ，其中 θ 是要估计的参数，而 $a = a(Y)$ 是一个估计量。直观上，损失代表了由于估计值 a 与真实值 θ 之间的差异而产生的惩罚；它是 θ 的函数，因此是一个随机变量。其思想是选择能够最小化 θ 的后验分布下期望损失的估计量 $\hat{\theta}$ 。这个最优估计量被称为**贝叶斯估计量** (Bayes estimator)，可以表示为 $\hat{\theta} = g(Y)$ ，其中

$$g(y) = \arg \min_a \mathbb{E}[L(\theta, a)|y] = \arg \min_a \int_{-\infty}^{\infty} L(\theta, a)\pi(\theta|y)d\theta.$$

我们现在将考虑一些最常见的损失函数及其产生的贝叶斯估计量。

10.3.2 Quadratic loss

如果选择的损失函数是 $L(\theta, a) = (\theta - a)^2$ ，那么我们需要最小化：

$$\mathbb{E}[L(\theta, a)|y] = \mathbb{E}[(\theta - a)^2|y] = \mathbb{E}(\theta^2|y) - 2a\mathbb{E}(\theta|y) + a^2.$$

对 a 求导得到：

$$\frac{\partial}{\partial a} \mathbb{E}[L(\theta, a)|y] = -2\mathbb{E}(\theta|y) + 2a.$$

当 $a = \mathbb{E}(\theta|y)$ 时，该导数为 0，并且 $\frac{\partial^2}{\partial a^2} \mathbb{E}[L(\theta, a)|y] = 2 > 0$ 。因此，后验均值 $\hat{\theta} = \mathbb{E}(\theta|Y)$ 就是贝叶斯估计量。我们可以很容易地证明，如果我们判断一个良好估计量的标准是它最小化后验均方误差，那么后验均值是最优的。这在概念上类似于最小二乘回归，在回归中我们的目标是最小化平方和（损失函数），并最终得到 $\hat{Y} = \mathbb{E}(Y|X)$ 作为我们的估计量（条件均值）。

定理 10.1 (后验均值) 设 $\hat{\theta} = \mathbb{E}(\theta|Y)$ 。那么，对于任何估计量 $\tilde{\theta}$ ，有：

$$\mathbb{E}[(\tilde{\theta} - \theta)^2] \geq \mathbb{E}[(\hat{\theta} - \theta)^2],$$

其中期望是关于 Y 和 θ 两者取的。

10.3.3 Absolute loss

设损失函数为 $L(\theta, a) = |\theta - a|$ 。我们对期望损失关于 a 求导，回顾 $\frac{d}{dx} \int_{-\infty}^x g(u)du = g(x)$ 和 $\frac{d}{dx} \int_x^{\infty} g(u)du = -g(x)$ ，得到：

$$\begin{aligned} \frac{\partial}{\partial a} \mathbb{E}[L(\theta, a)|y] &= \frac{\partial}{\partial a} \int_{-\infty}^{\infty} |\theta - a|\pi(\theta|y)d\theta \\ &= \frac{\partial}{\partial a} \left\{ \int_{-\infty}^a (a - \theta)\pi(\theta|y)d\theta + \int_a^{\infty} (\theta - a)\pi(\theta|y)d\theta \right\} \\ &= \int_{-\infty}^a \pi(\theta|y)d\theta - \int_a^{\infty} \pi(\theta|y)d\theta, \end{aligned}$$

其余项相互抵消。当这两个积分相等时，该导数为 0，即当 a 是 $\theta|y$ 的中位数时。二阶导数为 $2\pi(a|y) > 0$ ，因此这是一个最小值。所以，贝叶斯估计量为 $\hat{\theta} = \text{median}(\theta|Y)$ ，即**后验中位数** (posterior median)。

10.3.4 0-1 loss

考虑损失函数

$$L(\theta, a) = 1 - \mathbf{1}_{[a-c, a+c]}(\theta) = \begin{cases} 0 & \text{当 } |\theta - a| \leq c \text{ 时,} \\ 1 & \text{否则,} \end{cases}$$

其中 c 是一个给定的常数。期望损失为

$$\mathbb{E}[L(\theta, a)|y] = \mathbb{E} [1 - \mathbf{1}_{[a-c, a+c]}(\theta)|y] = 1 - P(a - c \leq \theta \leq a + c|y),$$

当最后表达式中的概率最大化时，期望损失最小化。我们选择 a ，使得区间 $[a - c, a + c]$ 在所有长度为 $2c$ 的区间中具有最高的后验概率；这被称为长度为 $2c$ 的**众数区间** (modal interval)。我们可以取这个众数区间的中点作为我们的贝叶斯估计量。

现在令 $c \rightarrow 0$ ，得到称为 0-1 损失的损失函数：

$$L(\theta, a) = \begin{cases} 0 & \text{如果 } \theta = a, \\ 1 & \text{否则.} \end{cases}$$

如果后验分布是连续的，则无法直接最小化期望损失。然而，如果分布是单峰的，则众数区间的端点总是位于众数的两侧，因此当 $c \rightarrow 0$ 时，众数区间收敛于众数。我们得出结论，贝叶斯估计量是**后验众数** (posterior mode) $\hat{\theta} = \text{mode}(\theta|Y)$ ，定义为

$$\text{mode}(\theta|Y) = g(Y),$$

其中

$$g(y) = \arg \max_{\theta} \pi(\theta|y) = \arg \max_{\theta} f(y|\theta)\pi(\theta).$$

如果我们选择一个平坦的（无信息的）先验，函数 $\pi(\theta)$ 是常数，表达式简化为

$$g(y) = \arg \max_{\theta} f(y|\theta),$$

这在经典的频率学派意义下，是基于样本 y 对未知（固定）参数 θ 的最大似然估计量。

10.3.5 Interval estimates

我们也可以使用后验分布来生成感兴趣参数的区间估计。如果 C 是一个满足下式的区间：

$$P(\theta \in C|Y = y) = 1 - \alpha, \quad (10.6)$$

那么我们称 C 为 θ 的一个 $100(1 - \alpha)\%$ 可信区间 (credible interval)（也称为**贝叶斯置信区间** (Bayesian confidence interval)）。注意，贝叶斯区间估计比其频率学派的对应物允许更自然的概率解释：给定数据，参数值落在 C 内的概率是 $1 - \alpha$ 。此外，请注意，对于给定的后验分布，有无限多个区间满足方程 (10.6)。如果后验是严格单峰的，那么满足 $\pi(a|y) = \pi(b|y)$ 的区间 $C = (a, b)$ 是最优的，但（为方便起见）我们通常选择一个具有相等尾部概率的区间，即：

$$\int_{-\infty}^a \pi(\theta|y)d\theta = \int_b^{\infty} \pi(\theta|y)d\theta = \alpha/2.$$

10.4 Hierarchical models and empirical Bayes

我们现在将关注先验 $\pi(\theta)$ 中的超参数，我们将其记为 η 。在 η 未知的情况下，拟合模型主要有两种方法。第一种是考虑一个完全的贝叶斯 formulation，其中 η 本身是一个具有先验密度 $h(\eta)$ 的随机变量，我们称之为**超先验** (hyperprior)；这种方法产生了**多阶段** (multistage) 或**分层模型** (hierarchical models)。第二种方法是直接从数据中估计 η ，然后将该估计值视为一个固定值。这被称为**经验贝叶斯** (empirical Bayes) 方法。

10.4.1 Hierarchical models

通常，如果参数 θ 隐式地依赖于具有超先验 (hyperprior) $h(\eta)$ 的超参数 η ，我们可以显式地表示这种依赖性，并将参数的联合后验分布表示为以下形式：

$$\pi(\theta, \eta|y) \propto f(y|\theta, \eta)\pi(\theta, \eta) = f(y|\theta)\pi(\theta|\eta)h(\eta). \quad (10.7)$$

我们写作 $f(y|\theta)$ 而不是 $f(y|\theta, \eta)$ ，因为 y 仅通过 θ 依赖于 η 。这是一个两阶段的分层贝叶斯模型。它可以很容易地扩展为一个多阶段模型，其中参数 η 依赖于另一个参数 ϕ ，而 ϕ 又可能依赖于另一个参数 ω ，依此类推。这种方法允许我们构建更精细的模型，但通常会导致数学上难以处理的积分。这些积分需要使用下一章描述的蒙特卡洛方法进行数值计算。

10.4.2 Empirical Bayes (EB)

在一个贝叶斯模型中，数据 \mathbf{Y} 依赖于一个参数 $\boldsymbol{\theta}$ ，而 $\boldsymbol{\theta}$ 又可能依赖于其他参数。分层贝叶斯方法允许我们构建这样的多阶段模型，但最终的超参数 η 必须遵循一个完全已知的分布。而使用经验贝叶斯方法，我们不做这样的假设，而是尝试从数据中估计 η 。

我们考虑一个如前所述的两阶段模型，但相同的方法可以扩展到具有更多阶段的模型。我们关注的技术被称为**参数化经验贝叶斯** (parametric empirical Bayes)，因为我们假设只有密度 $\pi(\theta|\eta)$ 的参数，即超参数向量 η ，是未知的。将函数 $\pi(\theta|\eta)$ 的形式视为未知则会导致**非参数经验贝叶斯** (nonparametric empirical Bayes) 方法，这超出了本笔记的范围。

给定 η 时 \mathbf{Y} 的条件密度可以写成以下形式：

$$\begin{aligned} f(\mathbf{y}|\eta) &= \frac{f(\mathbf{y}, \eta)}{h(\eta)} = \int \frac{f(\mathbf{y}, \theta, \eta)}{h(\eta)} d\theta \\ &= \int \frac{f(\mathbf{y}, \theta, \eta)}{f(\theta, \eta)} \frac{f(\theta, \eta)}{h(\eta)} d\theta \\ &= \int f(\mathbf{y}|\theta, \eta)\pi(\theta|\eta)d\theta. \end{aligned}$$

如果这个积分是可处理的，我们可以将此密度视为 η 的似然函数，并将其最大化以获得一个估计量 $\hat{\eta}(\mathbf{y})$ 。这种方法被称为**II型最大似然** (type-II maximum likelihood)。如果积分难以处理，或者无法直接最大化 $f(\mathbf{y}|\eta)$ (例如在 Beta-二项模型中)，一种常见的方法是通过迭代期望来计算 $\mathbf{Y}|\eta$ 的矩，然后利用这些矩来构建 η 的矩估计量。例如，条件均值为：

$$\mathbb{E}(Y|\eta) = \mathbb{E}[\mathbb{E}(Y|\eta, \theta)] = \mathbb{E}[\mathbb{E}(Y|\theta)|\eta].$$

10.4.3 Predictive inference

假设我们已经观测到 X_1, \dots, X_n ，并希望基于这些观测对某个 Y 进行预测。在贝叶斯预测推断中，我们使用参数的后验分布来生成随机变量 $Y|\mathbf{X} = \mathbf{x}$ 的密度。给定观测数据时 Y 的分布被称为**后验预测分布** (posterior predictive distribution)。这是为了将其与**先验预测分布** (prior predictive distribution) 区分开来，后者就是 \mathbf{X} 的边际分布。

使用与上节式子中相同的操作，我们可以将后验预测密度写为：

$$f(y|\mathbf{x}) = \int_{-\infty}^{\infty} f(y|\mathbf{x}, \theta)\pi(\theta|\mathbf{x})d\theta.$$

这种方法背后的动机是，密度 $f(y|\mathbf{x}, \theta)$ 往往比 $f(y|\mathbf{x})$ 简单得多。事实上，通常有 $f(y|\mathbf{x}, \theta) = f(y|\theta)$ 。

10.5 Further exercises

10.6 Appendix: Proofs

11 Simulation methods 模拟方法

在统计推断中，我们常常对确定某个特定模型下统计量的性质感兴趣。这可能是为了把握点估计量的质量，或者是为了构建检验的拒绝域。我们的首选方法是使用直接的数学推理，然而，在许多情况下，特别是对于有实际意义的模型，所涉及的数学是难以处理的。我们或许可以求助于渐近结果，但通常不清楚这种近似在有限样本下效果如何。例如，中心极限定理告诉我们样本均值的极限分布是正态分布，但并未说明正态分布对于（比如说）大小为 50 的样本近似程度如何。

解决这些问题的一个方案是使用模拟。概率模拟通常被称为蒙特卡洛方法（Monte Carlo techniques），此名源于以其赌场闻名的蒙特卡洛。蒙特卡洛方法基于生成随机数的技术，为评估难以处理的积分提供了一种简单而灵活的机制。利用快速、廉价计算机的普及，蒙特卡洛方法的应用极大地推动了贝叶斯方法在实际中的适用性。

11.1 Simulating independent values from a distribution

回忆一下，数学模型分为两类：确定性的和随机的。对于一个确定性模型，如果我们知道输入，我们就能精确确定输出将是什么。这对于随机模型则不成立，因为它的输出是一个随机变量。请记住，我们讨论的是模型的性质；关于现实生活现象是确定性还是随机性的讨论常常涉及非常混乱的思维，应该留给哲学家们。

简单的确定性模型可以产生非常复杂和美丽的行为。一个例子是逻辑斯蒂映射：

$$x_{n+1} = rx_n(1 - x_n),$$

其中 $n = 1, 2, \dots$ ，初始值 $x_0 \in (0, 1)$ 。对于参数的某些值 r ，这个映射会产生一个看起来随机的数字序列 x_1, x_2, \dots 。更准确地说，如果我们检验这些值，我们不会拒绝独立性的原假设。我们称这类序列为**伪随机**（pseudo-random）序列。

计算机程序无法产生真正的随机数；对于给定的输入，原则上我们可以精确预测程序的输出。然而，我们可以使用计算机生成一个伪随机数序列。这种区分有些人为，因为没有有限的数字序列是随机的；随机性是一种只能归属于数学模型（随机变量）的属性。从现在开始，我们将省略“伪”字，并理解当我们使用术语随机数时，我们指的是与随机变量实例无法区分的序列。

在本节中，我们描述了从我们迄今遇到的一些分布中生成随机样本的算法。所有这些方法都依赖于我们能够从区间 $[0, 1]$ 上的连续均匀分布中获得一个随机样本的前提，这是随机数生成器通常的输出。

11.1.1 Table lookup

生成离散随机变量实例的一种简单方法是采用如下方法。设 X 为一个离散随机变量，其分布函数为 $F_X(x)$ 。首先生成一个均匀随机变量 $U \sim \text{Unif}[0, 1]$ ，并令 u 为生成的值。

然后，令 $X = x_i$ ，其中 x_i 满足 $F_X(x_{i-1}) < u \leq F_X(x_i)$ 。这被称为**查表法**（table lookup），因为我们实际上是在 X 的累积概率表中查找 U 的值。

生成的变量具有所需的分布，因为

$$P(X = x_i) = P(F_X(x_{i-1}) < U \leq F_X(x_i)) = F_X(x_i) - F_X(x_{i-1}).$$

11.1.2 Probability integral

11.1.3 Box-Muller method

设 U, V 为独立的 $\text{Unif}[0, 1]$ 随机变量，并定义 $R = \sqrt{-2 \log U}$ 和 $\Theta = 2\pi V$ 。然后使用通常的公式将极坐标 (R, Θ) 转换为笛卡尔坐标 (X, Y) ，即

$$X = R \cos \Theta$$

和

$$Y = R \sin \Theta$$

随机变量 X, Y 是独立的标准正态分布。

Box-Muller 方法 (Box-Muller method) 也可用于从一般正态分布中抽样，因为

$$\mu + \sigma X \sim N(\mu, \sigma^2).$$

11.1.4 Accept/reject method

11.1.5 Composition

11.1.6 Simulating model structure and the bootstrap

11.2 Monte Carlo integration

11.2.1 Averaging over simulated instances

11.2.2 Univariate vs multivariate integrals

11.2.3 Importance sampling

11.2.4 Antithetic variates

11.3 Markov chain Monte Carlo (MCMC)

11.3.1 Discrete Metropolis

11.3.2 Continuous Metropolis

11.3.3 Metropolis-Hastings algorithm

11.3.4 Gibbs sampler

11.4 Further exercises

11.5 Appendix: Proofs