

概率论与数理统计

Karry Ren

在我看来，概率论与数理统计是最重要的一门数学课，是当下金融学、人工智能的本质基础，是刻画世界的最有力的工具。因此我的后半生一定会一直和其打交道，考虑到之前的学习过程比较分散，未形成自己的体系。从 2023 年 8 月开始整理此笔记，目的就是打造一则最完善的概率论与数理统计基础宝典，再次疏通整个概率论数理统计框架，掌握最基础最深层次的含义，并用自己的话加以描述，为日后的使用奠定坚实的基础。

本笔记参考茆诗松老师的教材整理而来，常见题目出自考研课程。

概率论与数理统计

第一章 随机事件与概率

1.1 随机事件及其运算

1.2 概率的定义及其确定方法

1.2.1 概率的公理化定义

1.2.2 概率的确定方法

a. 确定概率的古典方法

b. 确定概率的几何方法

1.3 概率的性质

1.3.1 概率的基本性质

1.3.2 条件概率及其性质

a. 条件概率的定义

b. 全概率公式

c. 贝叶斯公式

1.3.3 独立性

a. 定义

b. 性质

1.4 考点与典例

1.4.1 考点（量化笔试就曾出现过考研原题）

1.4.2 典例

第二章 随机变量及其分布

2.1 随机变量及其分布的概念

2.1.1 随机变量

2.1.2 分布函数

a. 定义

b. 性质（分布函数的三点性质）

2.1.3 分布函数的形式

a. 离散变量的分布列

b. 连续变量的概率密度函数

2.2 随机变量的数字特征

2.2.1 k 阶矩

2.2.2 数学期望（1 阶原点矩）

a. 定义

b. 性质

2.2.3 方差 & 标准差 (2 阶中心矩)

a. 定义

b. 性质

c. 切比雪夫不等式

2.2.3 变异系数

2.2.4 分位数 (p 值、中位数)

2.2.5 偏度 & 峰度

a. 偏度

b. 峰度

2.3 常用分布

2.3.1 离散分布

a. 泊松分布 (Poisson D)

b. 负二项分布 (帕斯卡分布)

2.3.2 连续分布

a. 正态分布 (必须会背!)

b. 伽马分布

c. 其他分布及性质

2.4 随机变量变换的分布

2.4.1 随机变量函数的分布

2.4.2 随机变量组合的分布

2.5 考点与典例

2.5.1 考点 1 分布函数的概念与性质

2.5.2 考点 2 随机变量的性质

2.5.3 考点 3 随机变量函数的分布

第三章 多维随机变量

3.1 多维随机变量及其分布

3.1.1 多维随机变量的概念

3.1.2 联合分布函数

a. 定义

b. 性质 (联合分布的四点性质)

3.1.3 联合分布函数的形式

a. 离散变量的联合分布列

b. 连续变量联合密度函数

3.1.4 常见的多维分布

a. 多项分布

b. 二元正态分布 (必须会背, 性质必须掌握!)

3.2 边际分布与随机变量的独立性

3.2.1 边际分布函数

3.2.2 边际分布函数形式

a. 离散变量的边际分布列

b. 连续变量的边际密度函数

- c. 常见的边际分布
- 3.2.3 随机变量间的独立性
- 3.2.4 多维随机变量函数的分布
- 3.3 多维随机变量的特征数
 - 3.3.1 多维随机变量函数的数字特征（一维）
 - a. 数学期望
 - b. 一维随机变量组合的数学期望和方差运算性质
 - 3.3.2 多维随机变量之间的数字特征（多维）
 - a. 协方差
 - b. 相关系数
- 3.4 条件分布与条件期望
 - 3.4.1 条件分布
 - a. 离散随机变量的条件分布列
 - b. 连续随机变量的条件密度函数
 - 3.4.2 条件数学期望
- 3.5 考点与典例
 - 3.5.1 考点 1 联合分布函数的概念和计算
 - 3.5.2 考点 2 二维正态分布的性质和性质
 - 3.5.3 考点 3 多维随机变量函数的分布（始终从定义出发！）

第四章 大数定律与中心极限定理

- 4.1 随机变量序列的两种收敛性
 - 4.1.1 依概率收敛
 - 4.1.2 按分布收敛
- 4.2 大数定律
 - 4.2.1 伯努利大数定律
 - 4.2.2 切比雪夫大数定律
 - 4.2.3 马尔可夫大数定律
 - 4.2.4 辛钦大数定律
- 4.3 中心极限定理
 - 4.3.1 独立同分布下的中心极限定理
 - a. 林德伯格-莱维中心极限定理
 - b. 棣莫弗-拉普拉斯中心极限定理
 - 4.3.2 独立不同分布的中心极限定理
 - a. 林德伯格中心极限定理
 - b. 李雅普诺夫中心极限定理
- 4.4 考点与典例
 - 4.4.1 考点 1 切比雪夫不等式协助求解概率范围
 - 4.4.2 考点 2 大数定理的概念与性质
 - 4.4.3 考点 3 中心极限定理标准化求解

第五章 统计量及其分布

- 5.1 总体与样本
 - 5.1.1 总体与个体
 - 5.1.2 样本

5.2 统计量及其分布

5.2.1 统计量与抽样分布

5.2.2 样本均值

5.3.3 样本方差与样本标准差

5.3.4 样本原点矩、样本中心距及其函数

5.3.5 五数概括与箱线图

5.4 三大抽样分布

5.4.1 χ^2 分布（卡方分布）【正态分布随机变量的统计量】

5.4.2 F 分布【卡方分布随机变量的统计量】

5.4.3 student-t 分布【正态分布和卡方分布随机变量的统计变量】

5.5 充分统计量【理解】

第六章 参数估计 <对总体的参数估计>

6.1 点估计的概念与常见方法

6.1.1 点估计及其评判标准

- a. 点估计
- b. 相合性
- c. 无偏性
- d. 有效性
- d. 最小方差无偏估计

6.1.2 矩估计【点估计 way 1】

6.1.3 极大似然估计【点估计 way 2】

- a. 极大似然估计
- b. EM 算法

6.1.4 贝叶斯估计【点估计 way 3】

- a. 统计推断的基础究竟是什么？
- b. 贝叶斯统计的含义与其密度函数形式

6.2 区间估计

6.2.1 区间估计的概念与估计方法

- a. 区间估计的概念
- a. 枢轴量法进行区间估计

6.2.2 单个正态总体参数的区间估计

- a. 总体的方差 σ 已知，进行 μ 的区间估计
- b. 总体的方差 σ 未知，进行 μ 的区间估计
- c. 总体的均值 μ 未知，进行 σ^2 的区间估计

6.2.3 两个正态总体下的区间估计

- a. 两个均值差 $\mu_1 - \mu_2$ 的区间估计
- b. σ_1^2/σ_2^2 的置信区间

第七章 假设检验

第一章 随机事件与概率

1.1 随机事件及其运算

所有的数学都一定是从看得见摸得着的事物出发的，概率同样也是如此，概率的直观概念就是随机事件发生的可能性大小。因此我们先研究生活中可能看到的一些随机事件，并用数学语言进行刻画，并进行相应的集合运算（交、并、补、差、互斥<一分为多>、对立<一分为二>等等），得到相关的运算法则（德摩根定理）。

\$\$

$$\overline{A \cup B} = \overline{A} \cap \overline{B} \quad \overline{A \cap B} = \overline{A} \cup \overline{B}$$

\$\$

刻画随机事件（连续 or 离散）让我们对概率所依附的实体有了直观的认识。同时也要理解清楚样本空间的含义是什么。

【定义】 样本空间 and 样本点：随机现象的一切可能基本结果组成的集合成为样本空间，记为 $\Omega = \{\omega\}$ ，其中 ω 表示基本结果，又称为样本点。样本点就是概率统计中抽样、计算的基本单元，认识随机现象首先要列出他的样本空间。

1.2 概率的定义及其确定方法

1.2.1 概率的公理化定义

尽管从直观上出发，通过解释随机事件，能够给出“概率”的定义。

但是无论是哪种定义（古典、几何）方式都无法适应于一切随机现象。因此：

- 1900 年数学家 Hilbert 提出要建立概率的公理化定义，以最少的几条本质特性来刻画概率的概念。
- 1933 年 Kolmogorov 提出了如下概率的公理化定义。

设 Ω 为一个样本空间， \mathcal{F} 为 Ω 的某些子集组成的一个事件域（简单来说就是样本空间中各样本点的各种组合，因此事件域肯定包含样本空间），如果对任一事件 $A \in \mathcal{F}$ ，定义在 \mathcal{F} 上的一个实值函数 $P(A)$ 满足：

- 非负性公理：若 $A \in \mathcal{F}$ ，则 $P(A) \geq 0$ ；
- 正则性公理： $P(\Omega) = 1$ ；
- 可列可加性公理：若 $A_1, A_2, \dots, A_n, \dots$ 互不相容，有（和的概率等于概率之和）

\$\$

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i),$$

\$\$

则称 $P(A)$ 为事件 A 的概率，称三元素 (Ω, \mathcal{F}, P) 为概率空间。

可列可加性是推出一系列概率公式的出发点 $P(A + B); P(A - B) \dots$

1.2.2 概率的确定方法

确定了概率是什么，还需要知道概率如何计算。其实本质上讲概率的计算可以划分为两类：

- 离散型随机变量的概率计算（古典方法）
- 连续型随机变量的概率计算（几何方法）

a. 确定概率的古典方法

对于在离散的样本空间中的有限个样本点且等可能而言，可以直接通过排列组合的方式直接计算（本质就是数点）。常见的模型有：抽样模型（放回、不放回）以及盒子模型等。

对于这类问题，计算最复杂但思维清晰的思路是（有些多情况可以不做唯一标识进行简化计算，但是思维的复杂度会增加）：

- 对样本空间内的每种情况做唯一标识，做分母
- 对符合条件的每种情况做唯一标识，做分子

古典方法的例子

- 盒子模型

设有 n 个球，每个球都等可能地被放到 N 个不同的盒子中的任意一个，每个盒子所放球数不限，求：

- (1) 指定的 n ($n \leq N$) 个盒子中各有一球的概率 p_1
- (2) 恰好有 n ($n \leq N$) 个盒子中各有一球的概率 p_2

// 思路：对于这个题目而言，我们既可以把所有的球看作相通的，也可以给每个球编号看作不同的。

// 为了简化思维，最好的方式就是给每个球编号进而对样本空间的每种情况做唯一标识。

(1)

- 分母（全样本空间）： n 个球，每个球都有 N 中放法 $\Rightarrow n^N$ ；
- 子分子（满足情况的点）：把 n 个球放到 n 个盒子中，而且是互斥的 $\Rightarrow n!$
- 结果： $n! / n^N$

(2)

- 分母（全样本空间）不变。
- 子分子（满足情况的点）：多了一步在 N 中选 n 个盒子的步骤，所以可能性就多了 $* C(Nn)$
- 结果：上述结果 $* C(Nn)$

b. 确定概率的几何方法

对于在连续的样本空间中的无限个样本点而言，无法通过直接数点的方式进行计算。常见的例子有：会面问题、比丰投针问题（算 π ）、算函数的积分（蒙特卡洛模拟）等。

对于这类问题，思路和上述本质是一样的，不过分子分母从之前的点的数量变成了一维（线的长度），二维（面积），三维（体积）之比。

几何方法的例子

- [贝特朗奇论](#)

在一圆内任取一条弦，问其长度超过该圆内接等边三角形的边长的概率是多少。

针对这一个问题，有三种不同的解题思路得到三种完全不同的答案。（所占的整体立场不同，结果就会截然不同）

1.3 概率的性质

1.3.1 概率的基本性质

由概率的公理化定义可以轻松得到概率的如下性质：有限可加性、单调性。条件概率满足一切的基本概率性质。在这我们强调两个公式，因为经常用到

- 加法公式（三个事件）

\$\$

$$P(A + B + C) = (P(A) + P(B) + P(C)) - (P(AB) + P(AC) + P(BC)) + P(ABC)$$

\$\$

- 互斥公式

\$\$

$$P(AB \setminus C) = P(AB) - P(ABC)$$

\$\$

1.3.2 条件概率及其性质

a. 条件概率的定义

\$\$

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

\$\$

b. 全概率公式

\$\$

$$P(B) = \sum P(A_i)P(B | A_i)$$

\$\$

全概率公式是多个分散情况到集中情况的汇总：是在已经知道多个简单事件（构成完备事件组）的概率，以及在每个简单事件发生情况下复杂事件发生的条件概率的情况下，求解复杂事件所发生实际概率（积沙成塔，条条大路通罗马，不放回未知抽签）。

c. 贝叶斯公式

\$\$

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

\$\$

贝叶斯公式是打开真实世界的一把钥匙。首先我们思考一个问题：在真实世界中真的存在非条件概率吗？我个人觉得不存在，也就是说世界上所有概率本质都是条件概率，但是由于条件的错综复杂性，很难直接获得准确、真实的条件概率值，那我们就需要一步步地细化认知，怎么做呢？按照如下的方式（最经典的例子莫过于狼来了的故事，以及检查疾病的例子）。

Step 1：依靠过去的经验，对事件 A 的发生有着基本的判断 => 通过过往认知，能够得到事件 A 发生的先验概率为 $P(A)$

Step 2：但是这个先验概率可能并不是十分准确的，因为真实世界中可能根本不存在非条件概率。

A 的发生总会和其他一些事情相牵连，所以我们总会充分获取一些事情（比如 B）的相关信息来对 A 发生的概率进行调整

Step 3：计算后验概率 $P(A|B)$ ，之所以叫做后验概率就是因为实在获取 B 事件之后才调整出的概率。

Step 1 to 3 在真实世界中是不断循环往复，以获取尽可能准确的后验概率。

// ----- //

教材上讲了狼来了的故事，我们在这再补充一个检查肝癌的例子。

- 已经知道一个村子的肝癌发生概率（先验概率 $P(L)$ ），这个概率是通过过往的经验来的（比如过去 100 年一个共有多少人得了）

- 但是只有这个先验概率，我们只能得出一个结论：每个人都有 $P(L)$ 的几率罹患肝癌，这个概率是没有意义的。

因此，我们要找和肝癌相关联的事件，科学家们历经千辛万苦找到了和肝癌具有一定“因果关系”（或许只是相关关系）的一种试剂检测方式

- 这样我们就可以计算后验概率 $P(L|A)$ 以及 $P(L|\sim A)$ 帮助甄别肝癌患病概率。当然，这要求我们有着精确的 $P(A)$ ；

$P(A|L)$; $P(A|\sim L)$;

• $P(A)$ 就是我们之前已经获取的不那么精准的先验概率；

• $P(A|L)$ 和 $P(A|\sim L)$ 都可以通过实验得出。

1.3.3 独立性

a. 定义

通过条件概率公式我们可以推得如果一个事件 A 的先验概率 $P(A)$ 等于在某事件 B 下的后验概率 $P(A|B)$ ，那么我们就说 A、B 二者独立，从直观上理解就是事件 A 的发生于事件 B 没有任何关系。更进一步地，如果满足下式子，即说明两个事件相互独立。注意，这个公式是判断事件是否独立的唯一方法，千万不要依靠直觉去判断！

\$\$

$$P(AB) = P(A)P(B)$$

\$\$

拓展来看（多个事件的相互独立性）：设有 n 个事件 A_1, A_2, \dots, A_n ，对任意的 $1 \leq i < j < k \dots \leq n$ ，如果以下等式均成立

\$\$

$$\left. \begin{aligned} & P(A_i A_j) = P(A_i) P(A_j), \\ & P(A_i A_j A_k) = P(A_i) P(A_j) P(A_k), \\ & \vdots \\ & P(A_1 A_2 \cdots A_n) = P(A_1) P(A_2) \cdots P(A_n), \end{aligned} \right.$$

\$\$

则称此 n 个事件 A_1, A_2, \dots, A_n 相互独立。拿 3 个事件举例子：设 A, B, C 是三个随机事件，如果有

\$\$

$$\left. \begin{aligned} & P(AB) = P(A) P(B), \\ & P(AC) = P(A) P(C), \\ & P(BC) = P(B) P(C), \end{aligned} \right.$$

\$\$

则只能得到 A, B, C 两两独立，若还有

\$\$

$$P(ABC) = P(A) P(B) P(C)$$

\$\$

才能得到 A, B, C 相互独立。

b. 性质

- 连锁反应：如果 A 和 B 独立，那么 A 反、 B 反、 A, B 的各种组合都独立不要怀疑！
- 必定独立：概率为 0 的时间以及概率为 1 的事件与任意一个事件均相互独立；
- A 和 B 相互独立，有： $P(A|B) = P(A|\bar{B}) = P(A)$ ，反之也成立；
- 互斥不一定独立，独立不一定互斥。二者本质上没有任何联系，根据公式就可以看出，只要 A 和 B 的概率同时都大于零那独立和互斥不可能同时出现。

1.4 考点与典例

1.4.1 考点（量化笔试就曾出现过考研原题）

- 随机事件的运算（本质是集合的计算以及集合关系的判断，利用各种集合运算定律<尤其是德摩根>，最简单的技巧就是画出集合的 VN 图）
- 古典概型排列组合，几何概型画图计算
- 有关概率的各种证明（三个概率的加法公式），尤其是条件概率的证明，是对乘法公式、贝叶斯公式的深入考察
- 全概率公式和贝叶斯公式
- 根据独立计算概率 => 伯努利古典概型

1.4.2 典例

- 三个事件的加法公式（千万不要嫌麻烦，硬算就是正确的道路）

设 A, B, C 为三个随机事件，且 $P(A) = P(B) = P(C) = \frac{1}{4}$, $P(AB) = 0$, $P(AC) = P(BC) = \frac{1}{12}$, 则 A, B, C 中恰有一个事件发生的概率为 【 】

(A) $\frac{3}{4}$. (B) $\frac{2}{3}$. (C) $\frac{1}{2}$. (D) $\frac{5}{12}$.

- 小题要学会举特例

假设必然事件或不可能事件，下面这道题目就可以先举 B 是不可能事件，再举 B 是必然事件

【例 1.4】 A, B 为任意两事件，则与事件 $(A - B) \cup (B - C)$ 相等的事件为 【 】

(A) $A \cap B \cap C$. (B) $A \cup (B - C)$. (C) $(A \cup B) - C$. (D) $(A \cup B) - BC$.

【解析】选(D).

第二章 随机变量及其分布

第一章我们充分认识了随机事件的含义，但为了更好的进行数学处理，只用随机事件这一个工具来处理随机现象是不够的，因为事件始终是一种定性的表示。为了进行定量的数学处理，必须把随机现象的结果数量化，这就是引入随机变量的原因。

2.1 随机变量及其分布的概念

2.1.1 随机变量

定义在样本空间 Ω 的实值函数 $X = X(\omega)$ 称为随机变量，常用大写字母 X, Y, Z 等表示随机变量，其取值用小写字母 x, y, z 等表示。这表明：随机变量 X 是样本点 ω 的一个函数

- 函数既可以是不同样本对应不同的实数
- 函数也可以是多个样本点对应同一个实数

函数的自变量（样本点）可以是数，也可以不是数，但因变量一定是实数。随机变量定义的本质，是将随机事件的样本空间映射到实数空间上（随机事件的数量化）。更进一步地，随机变量就是对现实世界特征的抽象化，一个随机变量就可以表征样本点一个维度的特征。。根据所映射的实数空间的特点，可以分为离散随机变量和连续随机变量。

注意，与微积分的变量不同，概率论中的随机变量 X 是一种“随机取值的变量且伴随着一个分布”，也就是说我们不仅要知道 X 可能取哪些值，而且还要知道取这些值的概率各是多少，更重要的我们知道取这些值的概率和肯定为 1。因此有没有分布是区分一般变量与随机变量的主要标志。

2.1.2 分布函数

a. 定义

分布函数是将随机事件的概率与随机变量相连接的最直接的工具。设 X 是一个随机变量，对任意实数 x ，称

$$\begin{aligned} & \text{\$ \$} \\ & F(x) = P(X \leq x) \\ & \text{\$ \$} \end{aligned}$$

为随机变量 X 的分布函数，且称 X 服从 $F(x)$ ，记为 $X \sim F(x)$ ，有时也可用 $F_X(x)$ 以表明是 X 的分布函数（把 X 作为 F 的下标）。**分布函数的概念核心是“累积”。**

b. 性质（分布函数的三点性质）

任一分布函数 $F(x)$ 都具有如下三条基本性质：

- **单调性**： $F(x)$ 是定义在整个实数轴 $(-\infty, +\infty)$ 上的单调非减函数，即对任意的 $x_1 < x_2$ ，有 $F(x_1) \leq F(x_2)$ 。
- **有界性（规范性）**：对任意的 x ，有 $0 \leq F(x) \leq 1$ ，且

$$\begin{aligned} & \text{\$ \$} \\ & \begin{array}{l} \lim_{x \rightarrow -\infty} F(x) = 0, \\ \lim_{x \rightarrow +\infty} F(x) = 1. \end{array} \\ & \text{\$ \$} \end{aligned}$$

- **右连续性**： $F(x)$ 是 x 的右连续函数，即对任意的 x_0 ，有

$$\begin{aligned} & \text{\$ \$} \\ & \lim_{x \rightarrow x_0^+} F(x) = F(x_0) \\ & \text{\$ \$} \end{aligned}$$

即

$$\begin{aligned} & \text{\$ \$} \\ & F(x_0^+) = F(x_0) \\ & \text{\$ \$} \end{aligned}$$

这也就意味着**分布函数的定义域一定是左闭右开**，左边的边界一定能包含进来，右边的边界一定无法包含进来（如果包含进来就说明右边界点的+极限小也可以包含进来，就矛盾了）。尽管对于连续函数而言计算上都是相等的，但是概念上千万不可以错。

以上三条性质都可以从定义直接推出（因为定义中包含概率，所以可以借助概率的三条公理）。需要注意的是，**这三条基本性质是判断某个函数是否能成为分布函数的充要条件**。下面我们就具体来看一下，随机变量的分布函数的形式。

2.1.3 分布函数的形式

a. 离散变量的分布列

对于离散变量而言，其所有可能的取值都是可列的，因此通过直接进行穷举就可以得到之前固有的概率表示形式，称之为分布列。其能够直观地表现出概率的大小关系，直接显示出哪些地方概率大，哪些地方概率小。

\$\$

```
\begin{array}{cccc} X & x_1 & \cdots & x_n \\ \hline P & p(x_1) & \cdots & p(x_n) \end{array}
```

\$\$

根据定义，我们也可以直接得到离散随机变量的分布函数

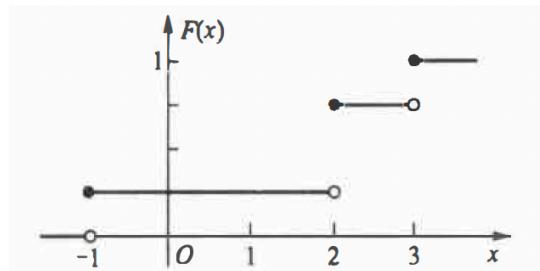


图 2.1.3 离散随机变量的分布函数

从这个地方我们就可以看出来，离散随机变量的分布列和分布函数都可以直接从概率推出来，二者之间并没有衍生关系。

b. 连续变量的概率密度函数

对连续变量而言，其所有可能的取值不可列，只能在数轴上进行连续地呈现，不存在分布列的概念，因此给出如下定义：设随机变量 X 的分布函数为 $F_X(x)$ ，如果存在实数轴上的一个非负可积函数 $p(x)$ ，使得对任意实数 x 有

\$\$

$$F(x) = \int_{-\infty}^x p(t) dt,$$

\$\$

则称 $p(x)$ 为 X 的概率密度函数，进而我们可以得到，在 $F(x)$ 存在导数的点上：

\$\$

$$p(x) = F'(x)$$

\$\$

因此概率密度函数是分布函数的一种衍生概念。概率密度函数的两大性质：

- 非负性：始终 ≥ 0
- 正则性：积分为 1

通过引入概率密度函数的概念，采用图形的面积直观展现概率大小，能够达到类似于分布列的效果，但在其本质含义上天差地别：离散随机变量 X 在其可能取值的点 $x_1, x_2, \dots, x_n, \dots$ 上的概率不为 0，而连续随机变量 X 在 $(-\infty, +\infty)$ 上任一点 a 的概率恒为 0

$$\begin{aligned} & \text{\$\$} \\ & P(X=a) = \int_{-\infty}^a p(x) dx = 0 \\ & \text{\$\$} \end{aligned}$$

这表明：

- 不可能事件的概率为 0，但概率为 0 的事件不一定是不可能事件（一个点）
- 必然事件的概率为 1，但概率为 1 的事件不一定是必然事件（连续样本空间挖掉一个点）

这也进一步呈现出了概率论和微积分的不同之处，在有限个点上（概率为 0 的部分）做操作对函数的性质没有任何的影响，也就是说在概率论中可以剔除概率为 0 的事件后讨论两个函数相等以及其他随机问题。

2.2 随机变量的数字特征

2.2.1 k 阶矩

设 X 为随机变量， k 为正整数。如果以下的数学期望都存在，则称

- $\mu_k = E(X^k)$ 为 X 的 **k 阶原点矩**；
- $\nu_k = E[X - E(X)]^k$ 为 X 的 **k 阶中心矩**。

可以看出，这是对随机变量基本数据特征的汇总，后续所有的数字特征都是通过这个概念衍生出来的，所谓矩就是距离。

2.2.2 数学期望（1 阶原点矩）

a. 定义

数学期望的本质是随机变量的值通过与概率的大小（发生的可能性）进行加权平均，表示分布所处位置的特征数，刻画了 X 的取值始终在 $E(x)$ 的周围波动，进而一定程度上消除随机变量的随机性。

针对离散和连续随机变量有不同的表述形式，但是其本质含义是相通的。我们在此给出连续随机变量的期望定义，离散随机变量只需将积分变为级数求和。设连续随机变量 X 的密度函数为 $p(x)$ ，如果

$$\begin{aligned} & \text{\$\$} \\ & \int_{-\infty}^{+\infty} |x| p(x) dx < +\infty \\ & \text{\$\$} \end{aligned}$$

则称

$$\begin{aligned} & \text{\$\$} \\ & E(X) = \int_{-\infty}^{+\infty} x p(x) dx \\ & \text{\$\$} \end{aligned}$$

为 X 的数学期望，或称为该分布 $p(x)$ 的数学期望，简称期望或均值。若 $\int_{-\infty}^{+\infty} |x| p(x) dx$ 不收敛，则称 X 的数学期望不存在。

b. 性质

数学期望的核心性质为如下，这个性质能够让我们在不求出随机变量 $g(X)$ 的分布的情况下求解出其均值：（该定理较难证明）

$$\begin{aligned} \mathbb{E}[g(X)] &= \begin{cases} \sum_i g(x_i) p(x_i), & \text{在离散场合;} \\ \int_{-\infty}^{+\infty} g(x) p(x) dx, & \text{在连续场合.} \end{cases} \end{aligned}$$

当然，这并不意味着所有随机变量函数的均值都要这么求，因为如果相乘后求积分很困难的话，我们还是需要用两步骤法：

- **Step 1** 求解出随机变量 $g(X)$ 的分布
- **Step 2** 根据分布直接计算均值

依据该重要性质，我们可以推导出：

- $E(c) = c$
- $E(aX) = aE(X)$

2.2.3 方差 & 标准差 (2 阶中心矩)

a. 定义

方差的本质是随机变量的值与期望的差的平方（离散程度）的期望，表示随机变量的波动程度。若随机变量 X^2 的数学期望 $E(X^2)$ 存在，则称偏差平方 $(X - E(X))^2$ 的数学期望 $E(X - E(X))^2$ 为随机变量 X （或相应分布）的方差，记为：

$$\begin{aligned} \text{Var}(X) &= E(X - E(X))^2 = \begin{cases} \sum_i [x_i - E(X)]^2 p(x_i), & \text{在离散场合;} \\ \int_{-\infty}^{+\infty} [x - E(X)]^2 p(x) dx, & \text{在连续场合.} \end{cases} \end{aligned}$$

称方差的正平方根 $\sqrt{\text{Var}(X)}$ 为随机变量 X （或相应分布）的标准差，记为 $\sigma(X)$ ，或 σ_X 。需要注意的是：如果随机变量 X 的数学期望存在，其方差不一定存在；而当 X 的方差存在时，则 $E(X)$ 必定存在，其原因在于 $|x| \leq x^2 + 1$ 总是成立的。

b. 性质

- $\text{Var}(X) = E(X^2) - [E(X)]^2$ 求解方差的常用方法。
- $\text{Var}(c) = E(c - E(c))^2 = E(c - c)^2 = 0$
- $\text{Var}(aX + b) = a^2 \text{Var}(X)$

c. 切比雪夫不等式

世间万物均有回归现象，随机变量其本质也是趋近于均值的回归，与均值偏差较大的事件发生的概率可能性要更小。因此随机变量的可能取值、均值以及方差之间存在一个很显然的不等式来表示这种回归约束，这就是切比雪夫不等式的现实含义（偏离期望越多，概率越小）。

设随机变量 X 的数学期望和方差都存在，则对任意常数 $\varepsilon > 0$ ，有：

$$\begin{aligned} & \text{P}(|X - E(X)| \geq \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2} \\ \end{aligned}$$

更进一步的解释如下：在概率论中，事件“ $|X - E(X)| \geq \varepsilon$ ”称为**大偏差**，其概率 $P(|X - E(X)| \geq \varepsilon)$ 称为**大偏差发生概率**，切比雪夫不等式给出**大偏差发生概率的上界**，这个上界与方差成正比，方差愈大上界也愈大，方差愈小随机变量的取值更加集中在均值附近。

【证明】

设 X 是一个连续随机变量，其密度函数为 $p(x)$ 。记 $E(X) = a$ ，我们有

$$\begin{aligned} & \text{P}(|X-a| \geq \varepsilon) = \int_{-\infty}^{+\infty} p(x) dx \int_{x-a \geq \varepsilon}^{\infty} p(x) dx \\ & \leq \frac{1}{\varepsilon^2} \int_{-\infty}^{+\infty} (x-a)^2 p(x) dx \\ \end{aligned}$$

2.2.3 变异系数

方差（或标准差）反映了随机变量取值的波动程度，但在比较两个随机变量的波动大小时，如果仅看方差（或标准差）的大小有时会产生不合理的现象。这有两个原因（说白了就是因为绝对的大小进行比较没有任何实际意义）

- 随机变量的取值有量纲，不同量纲的随机变量用其方差（或标准差）去比较它们的波动大小不太合理
- 在取值的量纲相同的情况下，取值的大小有一个**相对性问题**，取值较大的随机变量的方差（或标准差）也允许大一些

所以要比较两个随机变量的波动大小时，在有些场合使用以下定义的变异系数来进行比较，更具可比性。设随机变量 X 的二阶矩存在，则称比值

$$\begin{aligned} & C_v(X) = \frac{\sqrt{\text{Var}(X)}}{E(X)} = \frac{\sigma(X)}{E(X)} \\ \end{aligned}$$

为 X 的**变异系数**。因为变异系数是以其数学期望为单位去度量随机变量取值波动程度的特征数，标准差的量纲与数学期望的量纲是一致的，所以**变异系数是一个无量纲的量**。

2.2.4 分位数 (p 值、中位数)

设连续随机变量 X 的分布函数为 $F(x)$, 密度函数为 $p(x)$, 对任意 $p \in (0, 1)$, 称满足条件

$$\begin{aligned} & \text{\$\$} \\ F(x_p) = \int_{-\infty}^{x_p} p(x) dx &= p \\ & \text{\$\$} \end{aligned}$$

的 x_p 为此分布的 p 分位数, 又称下侧 p 分位数。

2.2.5 偏度 & 峰度

这两个特征数是描述分布形状的特征数, 本质是相对特征, 都是标准正态分布为基准。标准正态分布的偏度和峰度都是 0。在实际中, 一个分布标准化后的偏度和峰度皆为 0 或近似为 0 时, 常认为该分布为正态分布或近似为正态分布。

a. 偏度

设随机变量 X 的三阶矩存在, 则称比值

$$\begin{aligned} & \text{\$\$} \\ \beta_s = \frac{E[X-E(X)]^3}{[E(X-E(X))^2]^{3/2}} &= \frac{\nu_3}{(\nu_2)^{3/2}} \\ & \text{\$\$} \end{aligned}$$

为 X 的分布的偏度系数 (偏度), 是描述分布偏离对称程度的一个特征数。

- 当 $\beta_s > 0$ 时, 分布为正偏或右偏
- 当 $\beta_s = 0$ 时, 分布关于其均值 $E(X)$ 对称
- 当 $\beta_s < 0$ 时, 分布为负偏或左偏

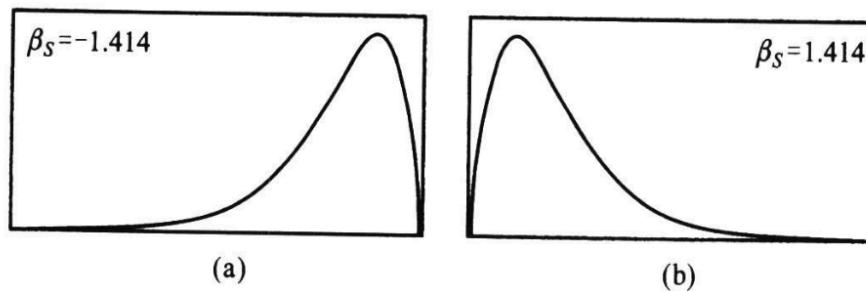


图 2.7.3 两个密度函数,(a)为左偏,(b)为右偏

b. 峰度

设随机变量 X 的四阶矩存在, 则称比值

$$\begin{aligned} & \text{\$\$} \\ \beta_k = \frac{E[X-E(X)]^4}{[E(X-E(X))^2]^2} - 3 &= \frac{\nu_4}{(\nu_2)^2} - 3 \\ & \text{\$\$} \end{aligned}$$

为 X 的分布的峰度系数，是描述分布尖锐程度和尾部粗细的一个特征数，注意不是分布的峰值高低（稍加计算就会发现正态分布的峰度和其峰值完全无关）。

- 当 $\beta_k < 0$ 时，则标准化后的分布尖锐程度比标准正态分布更平坦，称为低峰度
- 当 $\beta_2 = 0$ 时，则标准化后的分布尖锐程度与标准正态分布相当
- 当 $\beta_2 > 0$ 时，则标准化后的分布尖锐程度比标准正态分布更尖峭，称为高峰度

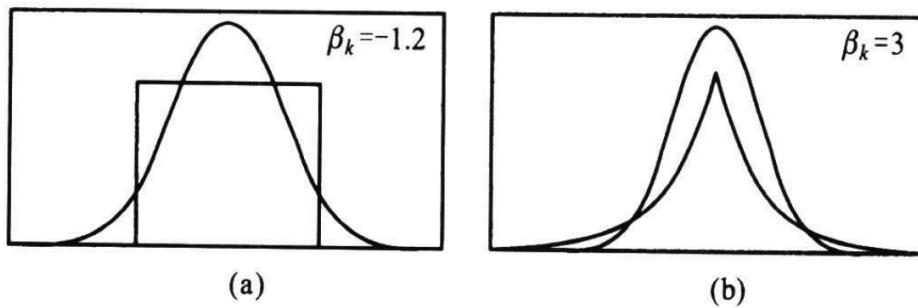


图 2.7.4 两个密度函数与标准正态分布密度函数的比较
它们的均值相等、方差相等、偏度皆为 0(对称分布)，而峰度有很大差别

2.3 常用分布

通过对现实世界的观察，可以总结发现很多常见的分布，在此对概念和常用性质做一个总结，需要经常进行记忆。后面则细化补充说明了一些常见分布的特点，并辅之以相关题目。

表 2.5.1 常用概率分布及其数学期望和方差

分 布	分布列 p_k 或分布密度 $p(x)$	期 望	方 差
0-1 分布	$p_k = p^k (1-p)^{1-k}, \quad k=0,1$	p	$p(1-p)$
二项分布 $b(n,p)$	$p_k = \binom{n}{k} p^k (1-p)^{n-k}, \quad k=0,1,\dots,n$	np	$np(1-p)$
泊松分布 $P(\lambda)$	$p_k = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k=0,1,\dots$	λ	λ
超几何分布 $h(n,N,M)$	$p_k = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}, \quad k=0,1,\dots,r, \quad r=\min\{M,n\}$	$n \frac{M}{N}$	$\frac{nM(N-M)(N-n)}{N^2(N-1)}$
几何分布 $Ge(p)$	$p_k = (1-p)^{k-1} p, \quad k=1,2,\dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
负二项分布 $Nb(r,p)$	$p_k = \binom{k-1}{r-1} (1-p)^{k-r} p^r, \quad k=r,r+1,\dots$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$
正态分布 $N(\mu, \sigma^2)$	$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < \infty$	μ	σ^2
均匀分布 $U(a,b)$	$p(x) = \frac{1}{b-a}, \quad a < x < b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$

续表

分 布	分布列 p_k 或分布密度 $p(x)$	期 望	方 差
指数分布 $Exp(\lambda)$	$p(x) = \lambda e^{-\lambda x}, \quad x \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
伽马分布 $Ga(\alpha, \lambda)$	$p(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x \geq 0$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
$\chi^2(n)$ 分布	$p(x) = \frac{x^{n/2-1} e^{-x/2}}{\Gamma(n/2) 2^{n/2}}, \quad x \geq 0$	n	$2n$
贝塔分布 $Be(a, b)$	$p(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}, \quad 0 < x < 1$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$
对数正态分布 $LN(\mu, \sigma^2)$	$p(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\right\}, \quad x > 0$	$e^{\mu + \sigma^2/2}$	$e^{2\mu + \sigma^2}(e^{\sigma^2} - 1)$
柯西分布 $Cau(\mu, \lambda)$	$p(x) = \frac{1}{\pi} \frac{\lambda}{\lambda^2 + (x-\mu)^2}, \quad -\infty < x < \infty$	不存在	不存在
韦布尔分布	$p(x) = F'(x), \quad F(x) = 1 - \exp\left\{-\left(\frac{x}{\eta}\right)^m\right\}, \quad x > 0$	$\eta \Gamma\left(1 + \frac{1}{m}\right)$	$\eta^2 \left[\Gamma\left(1 + \frac{2}{m}\right) - \Gamma^2\left(1 + \frac{1}{m}\right) \right]$

注: 表中仅列出各分布密度函数的非零区域.

2.3.1 离散分布

a. 泊松分布 (Poisson D)

泊松分布能对现实中的许多现象进行准确描述: 常与单位时间 (或单位面积、单位产品等) 上的计数过程相联系

- 在一天内来到某商场的顾客数 (可用于排队论的假设)
- 一平方米内, 玻璃上的气泡数
- 一定时期内, 放射性物质放出的粒子数

泊松定理 (二项分布的近似) : 在 n 重伯努利试验中, 记事件 A 在一次试验中发生的概率为 p_n (与试验次数 n 有关), 如果当 $n \rightarrow +\infty$ 时, 有 $np_n \rightarrow \lambda$, 则

\$\$

$$\lim_{n \rightarrow \infty} C_n^k p_n^k (1-p_n)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}.$$

\$\$

这样在试验次数 n 很大，而发生概率 p 很小的情况下，就可以将二项分布近似为泊松分布实现简化计算。

b. 负二项分布（帕斯卡分布）

定义：在伯努利试验序列中，记每次试验中事件 A 发生的概率为 p ，如果 X 为事件 A 第 r 次出现时的试验次数，则 X 的可能取值为 $r, r+1, \dots, r+m, \dots$ 。称 X 服从负二项分布，其分布列为

$$\begin{aligned} & \text{\$\$} \\ P(X = k) &= C_{k-1}^{r-1} p^r (1-p)^{k-r}; \quad k=r, r+1, \dots \\ & \text{\$\$} \end{aligned}$$

记为 $X \sim Nb(r, p)$ ，当 $r = 1$ 时，即为几何分布（几何分布具有无记忆性）。

2.3.2 连续分布

a. 正态分布（必须会背！）

$X \sim N(\mu, \sigma)$ 概率密度函数（配方、除系数、加因子）如下。积分得到的分布函数是算不出来的，只有趋近于正无穷 = 1 才有意义。

$$\begin{aligned} & \text{\$\$} \\ f(x) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ & \text{\$\$} \end{aligned}$$

大多数时间都是只分析标准正态分布，标准正态分布的均值和方差是可以积分出来的，进而得到普通正态分布的均值方差，借助这一特殊积分，我们就可以计算如下的积分形式了

$$\begin{aligned} & \text{\$\$} \\ \int_{-\infty}^{+\infty} e^{-ax^2} dx & \\ & \text{\$\$} \end{aligned}$$

b. 伽马分布

伽马函数：

$$\begin{aligned} & \text{\$\$} \\ \Gamma(\alpha) &= \int_0^{+\infty} x^{\alpha-1} e^{-x} dx \\ & \text{\$\$} \end{aligned}$$

两大性质，辅助求解积分

- $\Gamma(1) = 1, \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi};$
- $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ (可用分部积分法证得)。当 α 为自然数 n 时，有 $\Gamma(n + 1) = n\Gamma(n) = n!$

C. 其他分布及性质

指数分布无记忆性 $P(X > a + b | a) = P(x > b)$, 例如:

学了 10 小时的情况下再学 5 个小时的概率 = 刚开始就学 5 个小时的概率。

2.4 随机变量变换的分布

2.4.1 随机变量函数的分布

设 $y = g(x)$ 是定义在直线上的一个函数, X 是一个随机变量, 那么 $Y = g(X)$ 作为 X 的函数, 同样也是一个随机变量。我们要研究的问题是: 已知随机变量 X 的分布, 如何求出另一个随机变量 $Y = g(X)$ 的分布。对于离散和连续变量, 其本质都是一样的, 无非就是按照分布函数的定义进行推导, 只不过离散变量比较简单直接求解, 连续变量在某些情况下存在特殊法则。

一般情况下直接采用下面的方式对随机变量函数的分布进行求解:

- **Step 0** 确定区间范围: 画图观察 X 和 Y 之间的关系, 找到关键点, 进而确定区间【左闭右开】!
- **Step 1** 严格按照定义进行推导

$$\begin{aligned} \text{\$ \$} \\ \begin{aligned} F_Y(y) &= P\{Y \leq y\} \\ &= P\{g(X) \leq y\} \\ &= P\{\phi(y) \leq X \leq \gamma(y)\} \quad \text{反解 } X \text{ 时需要划分好区间} \\ \end{aligned} \\ \text{\$ \$} \end{aligned}$$

- **Step 2** 根据定义完成积分求出 $F_Y(y)$

$$\begin{aligned} \text{\$ \$} \\ F_Y(y) &= \int_{-\infty}^y f_X(x) dx = \int_{-\infty}^y \phi(y)^{\gamma(y)} f_X(x) dx = F_X(\gamma(y)) - F_X(\phi(y)) \\ \text{\$ \$} \end{aligned}$$

- **Step 3** $F_Y(y)$ 对 y 求导得到 $f_Y(y)$

$$\begin{aligned} \text{\$ \$} \\ f_Y(y) &= \frac{d}{dy} F_Y(y) \\ \text{\$ \$} \end{aligned}$$

来一个例子: 正态分布标准化的过程。已知: $X \sim N(\mu, \sigma)$ 证明 $Y = \frac{X-\mu}{\sigma} \sim N(0, 1)$

- **Step 1** 严格按照定义进行推导

$$\begin{aligned} \text{\$ \$} \\ \begin{aligned} F_Y(y) &= P\{Y \leq y\} \\ &= P\left\{\frac{X-\mu}{\sigma} \leq y\right\} \\ &= P\{X \leq \sigma y + \mu\} \end{aligned} \end{aligned}$$

\end{align*}

\$\$

- **Step 2** 积分求出 $F_Y(y)$

\$\$

$$F_Y(y) = \int_{-\infty}^y f_X(x) dx$$

\$\$

- **Step 3** $F_Y(y)$ 对 y 求导得到 $f_Y(y)$

\$\$

$$f_Y(y) = \frac{d F_Y(y)}{d y} = \sigma f_X(\sigma y + \mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2}}$$

2.4.2 随机变量组合的分布

上面我们求解的是一个随机变量函数变换后的分布情况，但还有另一种情况，那就是随机变量组合的情况。例如 N 个独立的 $0 - 1$ 分布随机变量 X_1, X_2, \dots, X_N 的加和 $Y = \sum_{i=1}^N X_i$ 就服从二项分布。

这也就自然地引出了后续要探讨的多维随机变量分布的学习，以及对世界上各种随机事件组合出新事件的深入思考。

2.5 考点与典例

2.5.1 考点 1 分布函数的概念与性质

1. 判断某些函数是否为分布函数，就是考察定义：1) 单调性；2) 两端 0 和 1；3) 右连续。

(1) 假设连续函数 $F(x)$ 是分布函数且 $F(0) = 0$ ，则下列函数也为分布函数的是 【 】

$$(A) G_1(x) = \begin{cases} 1 - F\left(\frac{1}{x}\right), & x > 1, \\ 0, & x \leq 1. \end{cases} \quad (B) G_2(x) = \begin{cases} 1 + F\left(\frac{1}{x}\right), & x > 1, \\ 0, & x \leq 1. \end{cases}$$

$$(C) G_3(x) = \begin{cases} F(x) - F\left(\frac{1}{x}\right), & x > 1, \\ 0, & x \leq 1. \end{cases} \quad (D) G_4(x) = \begin{cases} F(x) + F\left(\frac{1}{x}\right), & x > 1, \\ 0, & x \leq 1. \end{cases}$$

【c】 右连续直接秒

2. 根据分布函数求解概率，或者根据概率求解分布函数（套定义，得到公式）

【例 2.4】 假设随机变量 X 的绝对值不大于 1, $P\{X = -1\} = \frac{1}{8}$, $P\{X = 1\} = \frac{1}{4}$; 在事件 $\{-1 < X < 1\}$ 出现的条件下, X 在 $(-1, 1)$ 内的任一子区间上取值的条件概率与该子区间长度成正比. 试求 X 的分布函数 $F(x) = P\{X \leq x\}$.

三步走即可：1. 找到关键点（ -1 , 1 ） 2. 划分区间（牢记左闭右开<否则区间没有意义>） 3. 累积求和

2.5.2 考点 2 随机变量的性质

1. 对离散性随机变量进行判断（下面这道题必须要记住）

设 $X \sim P(\lambda)$, P_1, P_2 分别为随机变量 X 取偶数和奇数的概率，则 ()

(A) $P_1 = P_2$.

(B) $P_1 < P_2$.

(C) $P_1 > P_2$.

(D) P_1, P_2 大小关系不定.

这道题的求解太 amzing 了，感觉随时都有可能考到，这也说明了高数的重要性

- Step 1 给出泊松分布 $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$

- Step 2 给出取偶数的概率 $P(X = k \text{ (} k \text{ 为偶数)}) = \sum_{i=0}^n \left(\frac{\lambda^{2i}}{2i!} e^{-\lambda} \right) = e^{-\lambda} \sum_{i=0}^n \frac{\lambda^{2i}}{2i!}$

- Step 3 想方设法求解后面的级数，想到泰勒展开

- $e^x = \sum_{i=0}^n \frac{x^i}{i!}$

- $e^{-x} = \sum_{i=0}^n \frac{(-x)^i}{i!}$

- $e^x + e^{-x} = \sum_{i=0}^n \frac{2x^{2i}}{2i!}$ 所有偶数项的二倍

- $e^x - e^{-x} = \sum_{i=0}^n \frac{2x^{(2i+1)}}{(2i+1)!}$ 所有奇数项的二倍

因此 $e^{-\lambda} \sum_{i=0}^n \frac{\lambda^{2i}}{2i!} = e^{-\lambda} \left(\frac{e^x + e^{-x}}{2} \right) > \frac{1}{2}$

【c】 为偶数的概率大于 0.5，所以为偶数的概率大于为奇数的概率

2. (连续随机变量的概率密度函数) 和判断分布函数一样的意思。直接套定义，1) 始终大于等于 0；2) 积分为 1

设随机变量 X 的概率密度为 $f(x)$, 则下列可以作为概率密度的是 ()

- (A) $f(2x)$. (B) $f(2-x)$. (C) $f^2(x)$. (D) $f(x^2)$.

3. (指数分布) 最直接的考点就是：指数分布的定义

【例 2.19】 假设一大型设备在任何长为 t 的时间内发生故障的次数 $N(t)$ 服从参数为 λt 的泊松分布。

(I) 求相继两次故障之间时间间隔 T 的概率分布；

(II) 求在设备已经无故障工作 8 小时的情况下，继续无故障运行 8 小时的概率 Q .

\$\$

$F_T(t) = P\{T \leq t\} = 1 - P\{N(t) = 0\}$ (在 t 时间内不发生故障的概率)

\$\$

【定义题 + 智力题】只要弄清楚概率分布函数的定义并灵活转换即可 + 指数分布的无记忆性

4. (正态分布辅助计算积分) 实际上带 e^{x^2} 的积分直接求都是不可求的, 遇到之后都要借助标准正态分布, 永远记住: 不标准的正态分布根本没办法求解, 所有有关正态分布的计算都要标准化。

(7) 设 $f(x) = ke^{-x^2+2x-3}$ ($-\infty < x < +\infty$) 是某分布的概率密度, 则 $k = \underline{\hspace{2cm}}$.

2.5.3 考点 3 随机变量函数的分布

这一类题目数不胜数, 二维变量也是常见的, 我们在前面的知识点处已经梳理了整体的思路, 做题用下述步骤

1. 画图, 找到 x 和 $g(x)$ 的图像关系
2. 找关键点, 划分区间 (划分区间是最重要的, 左闭右开)
3. 套用定义完成计算 (step 1、2、3 走下去)

设随机变量 X 的概率密度为 $f_X(x) = \begin{cases} 3e^{-3x}, & x \geq 0, \\ 0, & x < 0, \end{cases}$ 求随机变量

$Y = \begin{cases} (X-1)^2 - 1, & X \geq 0 \\ -X, & X < 0 \end{cases}$ 的分布函数 $F_Y(y)$.

划分准区间为: $(-\infty, -1), [-1, 0), [0, +\infty)$

设随机变量 X 的概率密度为 $f(x) = \frac{e^x}{(1+e^x)^2}, -\infty < x < +\infty$, 令 $Y = e^x$,

- (1) 求 X 的分布函数;
- (2) 求 Y 的概率密度函数;
- (3) Y 的期望是否存在?

第三章 多维随机变量

如在一维随机变量中所提的那样，随机变量是对某个样本空间下的样本点特征的描述。在实际随机现象中，对样本空间 Ω 里的每个样本点 ω 而言，只用一个随机变量取描述他的特征是完全不够的。比如要研究西瓜的分类，对西瓜群这一样本空间下的每个西瓜而言，只研究样本点的体积 $X(\omega)$ 或者样本点的质量 $Y(\omega)$ 又或者样本点的密度 $Z(\omega)$ 这些个单一变量，从某个局部特征刻画西瓜，往往是不全面的。实际上，我们应该把这些随机变量所代表的特征作为一个整体联合考虑：

- 一方面，讨论这些特征同时变化的统计规律性，进而从整体上更好地表征样本特点
- 另一方面，也可以讨论多个变量之间的关系，发现特征之间相互的的统计规律性

因此，可以看出研究现实世界更多依靠的是多维随机变量，需要从多个维度全面综合地考虑某一样本点的整体特征。

所以这一章我们先引入了多维随机变量的概念，诠释了什么是联合分布函数（3.1节）。进而站在数理化的角度上，通过发掘二维联合分布函数以下三个方面的信息，达到全面理解样本特征的目的

- 每个分量的分布（每个分量单独表达的所有信息），即边际分布（3.2）
- 分量之间的关联程度，即协方差和相关系数；以及分量对整体的表征能力，即对多维变量进行函数替换后的均值（3.3）
- 大多数多维随机变量之间并非独立的，需要研究这种相依性。给定一个分量时，另一个分量的分布，即条件分布（3.4）

3.1 多维随机变量及其分布

这一部分的相关概念和定义，完全对照一维变量的进行对称理解。

3.1.1 多维随机变量的概念

多维随机变量的概念是从一维衍生出来的：如果 $X_1(\omega), X_2(\omega), \dots, X_n(\omega)$ 是定义在同一样本空间 $\Omega = \{\omega\}$ 上的 n 个随机变量，则称

```
$$
X(\omega)=(X_1(\omega),X_2(\omega),\dots,X_n(\omega))
$$
```

为 n 维（或 n 元）随机变量。注意，多维随机变量的关键是定义在同一样本空间上（都研究人或骰子的特征），对于不同样本空间上的两个随机变量（同时研究人和骰子），我们只能在乘积空间 $\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2); \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$ 上讨论。而在现实实际问题中，我们往往都是对同一样本空间进行研究，比如做线性回归时，所有的解释变量基本都来自于一个样本空间：

- 在研究四岁至六岁儿童的生长发育情况时，我们感兴趣于每个儿童（样本点 ω ）的身高 $X_1(\omega)$ 和体重 $X_2(\omega)$ 。这里 (X_1, X_2) 是一个二维随机变量。
- 在研究每个人的教育回报率的，我们感兴趣每个人（样本点 ω ）的教育时间 $X_1(\omega)$ 、工作时间 $X_2(\omega)$ 、性别 $X_3(\omega)$ 、年龄 $X_4(\omega)$ ，则 (X_1, X_2, X_3, X_4) 就是一个四维随机变量。

3.1.2 联合分布函数

a. 定义

不可置否，多维随机变量中的每一个随机变量都可以使用一维随机变量的理论进行研究。但是我们更需要探索出多维随机变量联合起来的特点，就必须依靠联合分布函数完成刻画。对任意的 n 个实数 x_1, x_2, \dots, x_n ，则 n 个事件 $\{X_1 \leq x_1\}, \{X_2 \leq x_2\}, \dots, \{X_n \leq x_n\}$ 同时发生的概率：

$$\begin{aligned} & \text{P}\left(x_1, x_2, \dots, x_n\right)=\text{P}\left(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n\right) \\ & \end{aligned}$$

称为 n 维随机变量 (X_1, X_2, \dots, X_n) 的联合分布函数。

在二维随机变量 (X, Y) 场合，联合分布函数 $F(x, y) = P(X \leq x, Y \leq y)$ 是事件 $\{X \leq x\}$ 与 $\{Y \leq y\}$ 同时发生（交）的概率。如果将二维随机变量 (X, Y) 看成是平面上随机点的坐标，那么联合分布函数 $F(x, y)$ 在 (x, y) 处的函数值就是随机点 (X, Y) 落在以 (x, y) 为右上角的无穷矩形内的概率。有时为了更加直观我们将 $F(x, y)$ 记作 $F_{X,Y}(x, y)$ 。

b. 性质（联合分布的四点性质）

任一二维联合分布函数 $F(x, y)$ 必具有如下四条基本性质（和分析多变量函数一样，该函数的特征往往是通过每个变量来体现的）：

- **单调性**： $F(x, y)$ 分别对 x 或 y 是单调不减的，即
 - 当 $x_1 < x_2$ 时，有 $F(x_1, y) \leq F(x_2, y)$
 - 当 $y_1 < y_2$ 时，有 $F(x, y_1) \leq F(x, y_2)$
- **有界性（规范性）**：对任意的 x 和 y ，有 $0 \leq F(x, y) \leq 1$ ，且

$$\begin{aligned} & \text{F}(-\infty, y)=\lim _{x \rightarrow-\infty} \text{F}(x, y)=0, \quad \& \\ & \text{F}(x,-\infty)=\lim _{y \rightarrow-\infty} \text{F}(x, y)=0, \quad \backslash \\ & \text{F}(+\infty,+\infty)=\lim _{x, y \rightarrow+\infty} \text{F}(x, y)=1 . \end{aligned}$$

$\end{aligned}$
\$

- **右连续性**：对每个变量都是右连续的，即

$$\begin{aligned} & \text{F}(x+0, y) \&=\text{F}(x, y), \quad \backslash \\ & \text{F}(x, y+0) \&=\text{F}(x, y) . \end{aligned}$$

$\end{aligned}$
\$

- **非负性**：对任意的 $a < b, c < d$ 有（夹缝概率大于等于 0）

$$\begin{aligned} & \text{P}(a < X \leq b, c < Y \leq d) = F(b, d) - F(a, d) - F(b, c) + F(a, c) \geq 0. \\ & \end{aligned}$$

同一维变量这四条性质也是充分必要条件。二维变量的前三点性质可以从概率的定义出发推出，而第四条是二维变量特有的（当然也是显然的）。注意前三条并不能推出第四条，存在满足前三条但不满足第四条的联合分布函数：

$$\begin{aligned} & \text{G}(x, y) = \begin{cases} 0, & x+y < 0; \\ 1, & x+y \geq 0 \end{cases} \\ & \end{aligned}$$

3.1.3 联合分布函数的形式

a. 离散变量的联合分布列

对于二维离散变量而言，每个点仍然是可列的，因此我们可以直观地对离散变量的每个点进行概率计算得到联合分布列，如果二维随机变量 (X, Y) 只取有限个或可列个数对 (x_i, y_j) ，则称 (X, Y) 为二维离散随机变量，称

$$\begin{aligned} & p_{ij} = P(X=x_i, Y=y_j), \quad i, j=1, 2, \dots \\ & \end{aligned}$$

为 (X, Y) 的联合分布列。分析联合分布列一般都是画一个二维的表格，进行直接列举即可。

b. 连续变量联合密度函数

根据分布函数的概念，可以衍生出直观描述连续变量概率分布情况的联合密度函数。如果存在二元非负函数 $p(x, y)$ ，使得二维随机变量 (X, Y) 的分布函数 $F(x, y)$ 可表示为

$$\begin{aligned} & F(x, y) = \int_{-\infty}^x \int_{-\infty}^y p(u, v) \, dv \, du \\ & \end{aligned}$$

称 $p(u, v)$ 为 (X, Y) 的联合密度函数。反过来，在 $F(x, y)$ 偏导数存在的点上有

$$\begin{aligned} & p(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y} \\ & \end{aligned}$$

联合密度函数两条基本性质不变（非负性 + 正则性）。联合密度函数的直观理解是使用体积衡量概率大小，因此在计算某一个区域的概率时，就是求出这一部分函数所成的体积大小，也就是求出联合密度函数在这一区域内的二重积分。

3.1.4 常见的多维分布

a. 多项分布

进行 n 次独立重复试验，如果每次试验有 r 个可能结果： A_1, A_2, \dots, A_r ，且每次试验中 A_i 发生的概率为 $p_i = P(A_i), i = 1, 2, \dots, r$ 且 $p_1 + p_2 + \dots + p_r = 1$ ，记 X_i 为 n 次独立重复试验中 A_i 出现的次数，则 (X_1, X_2, \dots, X_r) 取值 (n_1, n_2, \dots, n_r) 的概率，即 A_1 出现 n_1 次， A_2 出现 n_2 次…… A_r 出现 n_r 次的概率为

$$\begin{aligned} & \text{P}\left(X_1=n_1, X_2=n_2, \dots, X_r=n_r\right) = C_n^{n_1} C_{n-n_1}^{n_2} C_{n-n_1-n_2}^{n_3} \cdots C_{n-n_1-n_2-\dots-n_{r-1}}^{n_r} \\ & = \frac{n!}{n_1! n_2! \cdots n_r!} p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r} \end{aligned}$$

其中 $n = n_1 + n_2 + \dots + n_r$ ，这个联合分布列被称为 r 项分布，又称为多项分布。显然， r 项分布中只有 $r - 1$ 个随机变量！

b. 二元正态分布（必须会背，性质必须掌握！）

如果二维随机变量 (X, Y) 的联合密度函数为：

$$\begin{aligned} & p(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left[-\frac{1}{2(1-\rho^2)} \left(\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right) \right] \end{aligned}$$

其中 $-\infty < x, y < +\infty$ ，则称 (X, Y) 服从二元正态分布，记为 $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ 。其中五个参数的取值范围分别是：

$$\begin{aligned} & -\infty < \mu_1, \mu_2 < +\infty; \quad \sigma_1, \sigma_2 > 0; \quad -1 \leq \rho \leq 1. \end{aligned}$$

μ_1, μ_2 分别是 X 与 Y 的均值， σ_1^2, σ_2^2 分别是 X 与 Y 的方差， ρ 是 X 与 Y 的相关系数。

3.2 边际分布与随机变量的独立性

正如之前提到的那样，要想诠释真实世界往往需要直接引入多维随机变量展开研究。但是多维随机变量中的每个分量可能也存在某些性质需要发掘，这个时候我们就需要对多维随机变量进行降维，或者说进行维度压缩，这时候就需要用到边际分布了。

3.2.1 边际分布函数

边际分布函数的本质是进行维度压缩（消除一个或多个维度），方式是让其中一个或多个分量的组合概率为 1。

如果在二维随机变量 (X, Y) 的联合分布函数 $F(X, Y)$ 中令 $y \rightarrow +\infty$ ，由于 $\{Y < +\infty\}$ 为必然事件，故可得

$$\begin{aligned} & \lim_{y \rightarrow +\infty} F(x, y) = P(X \leq x, Y < +\infty) = P(X \leq x), \\ & \end{aligned}$$

这是一个一维分布函数，被称为 X 的边际分布，记为

$$\begin{aligned} & \lim_{y \rightarrow +\infty} F(x, y) = F(x, +\infty) \\ & F_X(x) = F(x, +\infty) \\ & \end{aligned}$$

类似地，在 $F(x, y)$ 中令 $x \rightarrow +\infty$ ，可得 Y 的边际分布

$$\begin{aligned} & \lim_{x \rightarrow +\infty} F(x, y) = F(+\infty, y) \\ & F_Y(y) = F(+\infty, y) \\ & \end{aligned}$$

类似地，可以对一个五维联合分布进行维度压缩，可以得到 5 个一维边际分布、10 个二维边际分布、10 个三维边际分布和 5 个四维边际分布。但明显这种降维方式是直接将其中一个或多个分量通过概率加和消除掉，这种方式应用到实际里面对应离散变量就是加和，对于连续变量进行积分（进行投影），这种简单的降维方式其实连带着把和其他分量之间的关系也消除掉了，而如果想要保留原有的相关关系就要借助条件分布的思维。

举个例子，对于身高体重 (x, y) 这个二维变量，求解边际分布是指：不管身高怎么变化，体重整体的变化情况。求解条件分布则是指，给定身高为某一值的条件下，这一局部的体重变化情况。

3.2.2 边际分布函数形式

a. 离散变量的边际分布列

从边际分布函数的概念出发，可以采用边际分布列对离散变量的边际分布进行直观表示。在二维离散随机变量 (X, Y) 的联合分布列 $\{P(X = x_i, Y = y_j)\}$ 中，对 j 求和所得的分布列

$$\begin{aligned} & \sum_{j=1}^{+\infty} P(X=x_i, Y=y_j) = P(X=x_i), i=1, 2, \dots \\ & \end{aligned}$$

被称为 X 的边际分布列。类似地，对 i 求和所得的分布列

$$\begin{aligned} & \sum_{i=1}^{+\infty} P(X=x_i, Y=y_j) = P(Y=y_j), j=1, 2, \dots \\ & \end{aligned}$$

\$\$

被称为 Y 的边际分布列。在实际操作中，直接给出 (X, Y) 的联合分布列矩阵，然后进行逐行或逐列的加和即可。

b. 连续变量的边际密度函数

直接从边际分布函数的概念出发，再结合概率密度函数的概念，可以得到边际密度函数的概念，这种定义是完全对比得到的，是计算出来的定义，而非设定性的定义。如果二维连续随机变量 (X, Y) 的联合密度函数为 $p_{X,Y}(x, y)$ ，因为

\$\$

$$\begin{aligned} F_X(x) &= F_{X,Y}(x, +\infty) = \int_{-\infty}^x \left(\int_{-\infty}^{+\infty} p_{X,Y}(u, v) du \right) dv \\ u &= \int_{-\infty}^x p_X(u) du, \quad F_Y(y) = \int_{-\infty}^y \left(\int_{-\infty}^{+\infty} p_{X,Y}(u, v) du \right) dv \end{aligned}$$

\$\$

其中 $p_X(x)$ 和 $p_Y(y)$ 分别为

\$\$

$$\begin{aligned} p_X(x) &= \int_{-\infty}^{+\infty} p(x, y) dy \\ p_Y(y) &= \int_{-\infty}^{+\infty} p(x, y) dx. \end{aligned}$$

对比后发现，二者分别正好都位于对应变量的密度函数的位置上，所以称这二者为边际密度函数。借助此定义就可以很好地完成计算了，无非就是画图 => 找上下界 => 求积分。

c. 常见的边际分布

二维正态分布的边际分布是（显然的）一维正态分布，以下求解过程必须牢记在心：

设 $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ 。先把二维正态密度函数 $p(x, y)$ 的指数部分

\$\$

$$-\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x-\mu_1)(y-\mu_2)}{\sigma_1 \sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right]$$

\$\$

改写成

这一步配方改写尤为重要，基本上有关正态分布的求解分都需要这一步改写！之所以这么改写是因为想要尽可能将 x 和 y 分离开。其实思路和一维是一样的：指数处配方 + 提系数，整体添因子。

\$\$

$$-\frac{1}{2} \left[\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1 \sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right]$$

\$\$

再对积分

\$\$

$$\int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2}\left[\rho\frac{x-\mu_1}{\sigma_1\sqrt{1-\rho^2}} - \frac{y-\mu_2}{\sigma_2\sqrt{1-\rho^2}}\right]^2\right) dy$$

\$\$

作变换（注意，因为是对 y 进行积分，所以可以把 x 看作常量）

\$\$

$$t = \rho\frac{x-\mu_1}{\sigma_1\sqrt{1-\rho^2}} - \frac{y-\mu_2}{\sigma_2\sqrt{1-\rho^2}},$$

\$\$

则

\$\$

$$\begin{aligned} p_X(x) &= \int_{-\infty}^{+\infty} p(x,y) dy \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(t-\rho\frac{x-\mu_1}{\sigma_1})^2}{2\sigma_2^2}\right) dt. \end{aligned}$$

\$\$

注意到上式中的积分恰好等于 $\sqrt{2\pi}$ （通过正态分布求解）所以有

\$\$

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right).$$

\$\$

这正是一维正态分布 $N(\mu_1, \sigma_1^2)$ 的密度函数，即 $X \sim N(\mu_1, \sigma_1^2)$ 。同理可证 $Y \sim N(\mu_2, \sigma_2^2)$ 。由此可见

- 二维正态分布的边际分布中不含参数 ρ （在求解边际分布进行降维的时候失去了变量间的相关信息）：这说明二维正态分布 $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, 0.1)$ 与 $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, 0.2)$ 的边际分布是相同的。
- 具有相同边际分布的多维联合分布可以是不同的，也就是说由联合分布一定能求解出边际分布，但由边际分布却不一定能反解联合分布（下面我们就认识到，只有当变量之间独立的时候，才可以推出）。

3.2.3 随机变量间的独立性

随机事件的独立性我们之前已经探讨过了，其是由随机事件发生的概率特性定义而来，此处随机变量之间的独立性也是从这儿衍生出来的定义。先来从随机变量的角度上直观理解独立性的概念：在多维随机变量中，各分量的取值有时会相互影响，但有时毫无影响。譬如一个人的身高 X 和体重 Y 就会相互影响，但与收入 Z 一般无影响。当两个随机变量取值的规律互不影响时，就称它们是相互独立的。

数学上的严格定义如下：设 n 维随机变量 (X_1, X_2, \dots, X_n) 的联合分布函数为 $F(x_1, x_2, \dots, x_n)$ ， $F_i(x_i)$ 为 X_i 的边际分布函数，如果对任意 n 个实数 x_1, x_2, \dots, x_n ，有

\$\$

$$F(x_1, x_2, \dots, x_n) = \prod_{i=1}^n F_i(x_i),$$

\$\$

则称 X_1, X_2, \dots, X_n 相互独立。随机变量相互独立的条件似乎比随机事件相互独立要弱？独立的定义一定是由事件的独立概率计算出发的，因此随机变量的定义式一定是从联合分布函数出发，这一点毋庸置疑。但直接依靠定义判断独立性往往是不方便的，通常借助如下推论：

- 在离散随机变量的场合下直接借助分布列：如果对其任意 n 个取值 x_1, x_2, \dots, x_n ，有

$$\begin{aligned} & \text{\$\$} \\ & P(X_1=x_1, X_2=x_2, \dots, X_n=x_n) = \prod_{i=1}^n P(X_i=x_i), \\ & \text{\$\$} \end{aligned}$$

则称 X_1, X_2, \dots, X_n 相互独立。

- 在连续随机变量的场合下直接借助概率密度函数：如果对任意 n 个实数 x_1, x_2, \dots, x_n ，有

$$\begin{aligned} & \text{\$\$} \\ & p_{\{X_1, X_2, \dots, X_n\}}(x_1, x_2, \dots, x_n) = \prod_{i=1}^n p_{\{X_i\}}(x_i), \\ & \text{\$\$} \end{aligned}$$

则称 X_1, X_2, \dots, X_n 相互独立。

因此我们判断独立性的方法如下：

- **Step 1** 由联合概率密度函数/分布列求解出边际概率密度函数/分布列（积分 or 加和）
- **Step 2** 判断边际概率密度函数/分布列连乘结果是否等于联合概率密度函数/分布列

对于常见的二维随机变量来说，只需要看两步：

1. 联合分布的定义域是不是一个矩形 $x \in [a, b]; y \in [c, d] \Rightarrow (x, y)$ 一定是矩形定义
2. 联合分布的函数形式是不是可以写成变量函数想乘的形式 $g(x)h(y)$

因此我们在判断时，一般都是直接看定义域就判断了。

3.2.4 多维随机变量函数的分布

设 (X_1, X_2, \dots, X_n) 为 n 维随机变量，则 $Y = g(X_1, X_2, \dots, X_n)$ 是 (X_1, X_2, \dots, X_n) 的函数， Y 是一维随机变量。现在的问题是如何由 (X_1, X_2, \dots, X_n) 的分布，求出 Y 的分布。可以看出 $g(\cdot)$ 是一个多维随机变量向一维随机变量的映射关系，可以将其堪称是一个极致维度压缩与特征变换方式，现实中的线性回归或者深度学习就可以理解为这样的变换！

我们所说服从二项分布的随机变量 Y 也可以看 n 维随机变量 (X_1, X_2, \dots, X_n) 其中 $X_i \sim 0-1(p)$ 的函数 $Y = \sum_{i=1}^n X_i$ 。（一个很有意思的事：之前是从时间上前后进行 n 次实验的维度为二项分布下的定义，但由于每次实验的独立性，和时间并没有关系，因此我们就可以转变为空间上同时进行 n 次实验的维度。）

在求解此类问题的过程中，我们绝不使用任何的中间结论（比如卷积公式），直接从定义出发，可以分成三类（连续-连续，离散-离散[直接画表列举]，连续-离散）在此以二维连续随机变量 (X, Y) 为例，定义 $Z = g(X, Y)$ ：

- **Step 0** 求解 Z 的范围，找到关键点，划分区间（左闭右开）
- **Step 1** 严格按照定义进行初始推导

\$\$

```
\begin{aligned}
F_Z(z) &= P\{Z \leq z\} \\
&= P\{g(X, Y) \leq z\} \\
\end{aligned}
```

\$\$

- **Step 2** 根据定义进行二元积分完成 $F_Z(z)$ 的求解

\$\$

$$F_Z(z) = \iint_{\{g(X, Y) \leq z\}} p_{X, Y}(x, y) dx dy$$

\$\$

- **Step 3** $F_Z(z)$ 对 z 进行求导得到概率密度函数 $f_Z(z)$

\$\$

$$f_Z(z) = \frac{\partial F_Z(z)}{\partial z}$$

\$\$

泊松分布的独立可加性: 设 $X \sim P(\lambda_1), Y \sim P(\lambda_2)$, 且 X 与 Y 独立, 则 $Z = X + Y \sim P(\lambda_1 + \lambda_2)$ 。

因为泊松分布是离散分布, 所以直接给出分布列进行计算即可。关键是下面的式子:

\$\$

$$P(Z=k) = \sum_{i=0}^k P(X=i)P(Y=k-i)$$

\$\$

正态分布的独立可加性: 设 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$, 且 X 与 Y 独立, 则 $Z = X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ 。

通过独立性简化计算, 这个也一样能通过计算的性质得到。**但是一定要记住: 如果两个正态分布的联合分布根本就不满足联合正态分布, 那么这两者的线性组合也不是正态分布!** 也就是说, 任意两个正态分布的线性组合不一定都是正态分布, 尽管我们能得到这个分布的均值方差, 但其不一定是正态分布, 必须要再补上一个联合分布的条件。[反例](#)

$Max()$ 、 $Min()$ 的特性:

- 设 $Y = Max(X_1, X_2, \dots, X_n)$, 且 X_1, X_2, \dots, X_n 相互独立, 则

\$\$

$$\begin{aligned}
F_Y(y) &= P\{Max(X_1, X_2, \dots, X_n) \leq y\} \\
&= P\{X_1 \leq y, X_2 \leq y, \dots, X_n \leq y\} \\
&= \prod_{i=1}^n P\{X_i \leq y\} \\
&= \prod_{i=1}^n F_{X_i}(y)
\end{aligned}$$

\$\$

- 设 $Y = Min(X_1, X_2, \dots, X_n)$, 且 X_1, X_2, \dots, X_n 相互独立, 则

\$\$

```

\begin{aligned}
F_Y(y) &= P\{\min(X_1, X_2, \dots, X_n) \leq y\} \\
&= 1 - P\{\min(X_1, X_2, \dots, X_n) > y\} \\
&= 1 - P\{X_1 > y, X_2 > y, \dots, X_n > y\} \\
&= 1 - \prod_{i=1}^n P\{X_i > y\} \\
&= 1 - \prod_{i=1}^n (1 - F_{X_i}(y))
\end{aligned}
$$

```

3.3 多维随机变量的特征数

想要探究多维随机变量之间的关系，就必须对变量之间的特征数进行细致地梳理。

拿到一个多维随机变量，我们完全可以考虑以下的特征数：

- 各个分量的期望、方差、标准差以及相关运算后的关系。
- 两个随机变量间的关联程度，即协方差与相关系数（反映两个随机变量相依关系的特征数）

3.3.1 多维随机变量函数的数字特征（一维）

a. 数学期望

当进行了 $Z = g(X_1, X_2, \dots, X_n)$ 这样的特征变换后，我们当然希望再深入研究一下变量 Z 的性质，比如说数学期望、方差等。针对 Z 这么一个一维变量（这也是首先介绍这一特征数的原因，因为多维随机变量函数本质是一个一维变量），一个很显然的思路就是两步法

- Step 1** 求解 Z 的分布函数
- Step 2** 借助分布函数对相应的特征数进行求解

与此同时我们还能通过复杂的推导得到如下定理（其本质是和一维随机变量数学期望的定理是一致的）：若二维随机变量 (X, Y) 的分布用联合分布列 $P(X = x_i, Y = y_j)$ 或用联合密度函数 $p(x, y)$ 表示，则 $Z = g(X, Y)$ 的数学期望为

```

$$
E(Z)=\begin{cases} \sum_i \sum_j g(x_i, y_j) P(X=x_i, Y=y_j), & \text{在离散场合}, \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) p(x, y) dx dy, & \text{在连续场合}. \end{cases}
$$

```

这种方法通常来说较为简单：不需要先求出 Z 的分布，直接计算即可。但需要注意的是，运算中存在的乘积可能导致函数求积分难以计算，此时必须要转换到定义法，老老实实地按照两步走计算。**这个定理的核心功用是进行证明，比如接下来我们将得到的相关性质。**

特别地，如果假设 $X = g(X, Y)$ 那么刚好求解的就是 X 的均值，也可以换个角度对这个结果进行理解，无非就是把 X 存在定义的地方发生的概率都拿出来，进行二元积分加权平均即可。

b. 一维随机变量组合的数学期望和方差运算性质

接下来我们重点看几个特殊的 Z ，基于上述定理，进而得到一些中间结论，辅助我们对一维随机变量的组合结果有更深的认识。对于多个一维随机变量之间的运算，我们总是下意识的就将其认为是多维随机变量，这当然是错误的！所以此处我们换一个更清晰的角度：将其看作进行了一个 $g(X, Y)$ 变换。这种思维在后续的统计中也是贯穿始终的！

1. (对任意的随机变量：和的均值等于均值的和) 设 (X, Y) 是二维随机变量，则有

$$\begin{aligned} & \text{E}(X+Y) = \text{E}(X) + \text{E}(Y) \\ & \end{aligned}$$

【证明】从上述定理出发，不妨设 (X, Y) 为连续随机变量，其联合密度函数为 $p(x, y)$ ，若令 $g(X, Y) = X + Y$

$$\begin{aligned} & \text{E}(X+Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x+y) p(x, y) dx dy \\ & = \int_{-\infty}^{+\infty} x \left(\int_{-\infty}^{+\infty} p(x, y) dy \right) dx + \int_{-\infty}^{+\infty} y \left(\int_{-\infty}^{+\infty} p(x, y) dx \right) dy \\ & = \text{E}(X) + \text{E}(Y). \end{aligned}$$

2. (对独立的随机变量：积的均值等于均值的积) 若随机变量 X 与 Y 相互独立，则有

$$\begin{aligned} & \text{E}(XY) = \text{E}(X)\text{E}(Y) \\ & \end{aligned}$$

【证明】同样是从上述定理出发，只不过此时令 $g(X, Y) = XY$ ，自然发现必须要补上一个条件 $p(x, y) = p_X(x)p_Y(y)$ ，即两个随机变量相互独立，才能得到如下推导：

$$\begin{aligned} & \text{E}(XY) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy p_X(x)p_Y(y) dx dy \\ & = \int_{-\infty}^{+\infty} x \left(\int_{-\infty}^{+\infty} y p_Y(y) dy \right) p_X(x) dx \\ & = \text{E}(X)\text{E}(Y). \end{aligned}$$

3. (对独立的随机变量：和的方差等于方差的和) 若随机变量 X 与 Y 相互独立，则有

$$\begin{aligned} & \text{Var}(X \pm Y) = \text{Var}(X) \pm \text{Var}(Y) \\ & \end{aligned}$$

【证明】由于方差是从均值来的，所以在这就不直接从定理出发了，而是借助已经得到的两条性质进行推导

$$\begin{aligned} & \text{Var}(X+Y) = \text{E}((X+E(X))+(Y+E(Y)))^2 \\ & = \text{E}(X^2 + 2XE(Y) + E(X)^2 + 2YE(X) + 2E(X)E(Y) + E(Y)^2) \\ & = \text{E}(X^2) + 2\text{E}(X)\text{E}(Y) + \text{E}(Y^2) \\ & = \text{Var}(X) + \text{Var}(Y) + 2\text{E}(X-E(X))\text{E}(Y-E(Y)). \end{aligned}$$

当随机变量 X 与 Y 相互独立时，最后一项为 0。

4. (对于独立的随机变量：积的方差与方差的积没有确切关系) 不同于均值，我们对方差进行推导时会发现

$$\begin{aligned} & \text{Var}(XY) = E((XY)^2) - E(XY)^2 \\ & = E(X^2)E(Y^2) - E(X)^2E(Y)^2 \end{aligned}$$

但是：

$$\begin{aligned} & \text{Var}(X)\text{Var}(Y) = E(X^2)E(Y^2) - E(X^2)E(Y)^2 - E(Y^2)E(X)^2 + E(X)^2E(Y)^2 \\ & = E(X^2)E(Y^2) - E(X)^2E(Y)^2 \end{aligned}$$

显然二者相等是需要更多条件的！

3.3.2 多维随机变量之间的数字特征（多维）

求解某一维的随机变量边际分布后，就直接将该变量与其他变量之间的关系给抹去了，那这个关系到底该如何表示呢？

a. 协方差

设 (X, Y) 是一个二维随机变量，若 $E[(X - E(X))(Y - E(Y))]$ 存在，则称此数学期望为 X 与 Y 的协方差，或称为 X 与 Y 的相关（中心）矩，并记为

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

从协方差的定义可以看出，它是 X 的线性偏差 “ $X - E(X)$ ” 与 Y 的线性偏差 “ $Y - E(Y)$ ” 乘积的数学期望。由于该线性偏差可正可负，故协方差也可正可负，也可为零，其具体表现如下：

- 当 $\text{Cov}(X, Y) > 0$ 时，称 X 与 Y 正相关，这时两个线性偏差 $(X - E(X))$ 与 $(Y - E(Y))$ 同时增加或同时减少。由于 $E(X)$ 与 $E(Y)$ 都是常数，这也就等价于 X 与 Y 同时增加或同时减少，这就是正相关的含义。
- 当 $\text{Cov}(X, Y) < 0$ 时，称 X 与 Y 负相关，这时 X 增加而 Y 减少，或 Y 增加而 X 减少，这就是负相关的含义。
- 当 $\text{Cov}(X, Y) = 0$ 时，称 X 与 Y 不相关。

一个老生常谈的结论：相关性比独立性要弱得多， X 与 Y 不相关只能表明二者线性关系独立！这也就是说独立一定不相关，反之只能证明线性关系独立。

设随机变量 $X \sim N(0, \sigma^2)$ ，且令 $Y = X^2$ ，则 X 与 Y 显然不独立，而此时 X 与 Y 的协方差却为 0

$$\begin{aligned} & \text{Cov}(X, Y) = \text{Cov}(X, X^2) = E(X \cdot X^2) - E(X)E(X^2) \\ & = E(X^3) - E(X)E(X^2) \\ & = E(X^3) - E(X)^3 = 0 \end{aligned}$$

这说明二者不相关。但是能否找到一个线性变换 $Y = aX + b$ 使得 $Cov(X, Y) = 0$ 呢？对于正态分布貌似找不出来，我们可以很轻松地推算出如下结论：对于任意两个联合分布服从二维正态分布的随机变量，不相关与独立就是等价的！一个简单的理解就是对于这两个正态分布来说，二者都可以直接通过线性变化得到，因此线性不相关后也就不可能存在其他的相关性了。但是对于任意两个服从正态分布的随机变量而言就不一定了，服从正态分布的随机变量之间也可能有非线形的变化！

依靠该定义，我们直接推出来如下几条性质（随便动手一算即可得到结果）：

- 用协方差表示方差：

$$\begin{aligned} \text{Var}(X \pm Y) &= \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y) \\ &\quad \end{aligned}$$

这个性质表明：如果 X, Y 是负相关的那么和的方差一定小于方差的和 \Rightarrow 分散风险的本质。

- 与常数之间的协方差为 0：

$$\begin{aligned} \text{Cov}(X, c) &= 0 \\ &\quad \end{aligned}$$

- 线性变换的协方差：

$$\begin{aligned} \text{Cov}(aX, bY) &= ab \text{Cov}(X, Y) \\ &\quad \end{aligned}$$

- 多变量线性组合的协方差等于协方差的线性组合：

$$\begin{aligned} \text{Cov}(X+Y, Z) &= \text{Cov}(X, Z) + \text{Cov}(Y, Z) \\ &\quad \end{aligned}$$

Schwarz 不等式：对任意二维随机变量 (X, Y) ，若 X 与 Y 的方差都存在，且记 $\sigma_X^2 = \text{Var}(X), \sigma_Y^2 = \text{Var}(Y)$ ，则有：

$$\begin{aligned} \text{Cov}(X, Y)^2 &\leq \sigma_X^2 \sigma_Y^2 \\ &\quad \end{aligned}$$

【证明】不妨设 $\sigma_X^2 > 0$ ，因为当 $\sigma_X^2 = 0$ 时，结论显然成立。在 $\sigma_X^2 > 0$ 成立下，考虑 t 的如下二次函数：

$$\begin{aligned} g(t) &= E[t(X-E(X))+(Y-E(Y))]^2 = t^2 \sigma_X^2 + 2t \cdot \text{Cov}(X, Y) + \sigma_Y^2 \\ &\quad \end{aligned}$$

由于上述的二次三项式非负，平方项系数 σ_X^2 为正，所以其判别式小于或等于零，即

$$\begin{aligned} 2\text{Cov}(X, Y)^2 - 4\sigma_X^2 \sigma_Y^2 &\leq 0 \\ &\quad \end{aligned}$$

\$\$

移项后即得施瓦茨不等式（太优雅了）！

b. 相关系数

协方差表示是表示变量相关关系的一个绝对概念，就和方差表示随机变量离散程度一样，值是带量纲的，因而相互间不能进行比较。无法从协方差的绝对差异上的出变量相关性的差异，因此需要消除量纲的影响，对协方差除以相同量纲的量，就得到了相关系数的定义：设 (X, Y) 是一个二维随机变量，且 $Var(X) > 0, Var(Y) > 0$ ，则称

\$\$

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

\$\$

为 X 与 Y 的（线性）相关系数。当然，我们也可以换个角度来理解相关系数，无非就是消除量纲，那我完全可以先对随机变量消除量纲后再求一个“绝对”的协方差，因此相关系数的另一个解释是：它是相应标准化变量的协方差。若记 X 与 Y 的数学期望分别为 μ_X, μ_Y ，其标准化变量为

\$\$

$$X^* = \frac{X - \mu_X}{\sigma_X}, \quad Y^* = \frac{Y - \mu_Y}{\sigma_Y}$$

\$\$

则有

\$\$

$$\text{Cov}(X^*, Y^*) = \text{Cov}\left(\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \text{Corr}(X, Y)$$

\$\$

深入理解相关系数的本质是极其重要的：因为相关系数是线性回归的核心，其所表示的线性关系是从协方差的定义处所衍生的。

由施瓦茨不等式可以得到相关系数的有界性

\$\$

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

\$\$

进而我们可以推出线性相关的最大性质（二者相关系数为 ± 1 并不表示绝对的在一条直线上，而是相对的依概率在一条直线上，这是根基！） $\text{Corr}(X, Y) = \pm 1$ 的充要条件是 X 与 Y 间几乎处处有线性关系，即存在 $a (\neq 0)$ 与 b ，使得

\$\$

$$P(Y = aX + b) = 1.$$

\$\$

其中当 $\text{Corr}(X, Y) = 1$ 时，有 $a > 0$ ；当 $\text{Corr}(X, Y) = -1$ 时，有 $a < 0$ 。

【证明】

- 先证明充分性（十分显然）

若 $Y = aX + b$, 则将 $Var(Y) = a^2Var(X)$, $Cov(X, Y) = a \cdot Cov(X, X) = a \cdot Var(X)$ 代入相关系数的定义中得

\$\$

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{a \cdot Var(X)}{\sigma_X \sigma_Y} = \begin{cases} 1, & a > 0; \\ -1, & a < 0. \end{cases}$$

\end{cases}

\$\$

- 再证明必要性（十分优雅）

\$\$

$$Var\left(\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y}\right) = 2[1 \pm \text{Corr}(X, Y)]$$

\$\$

所以当 $\text{Corr}(X, Y) = 1$ 时，有

\$\$

$$Var\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = 0,$$

\$\$

这个式子是关键，因为如果 $\text{Corr}(X, Y) \neq \pm 1$ 的话，这方差就不为零了，通过切比雪夫不等式就可以继续往下推导。

由此得

\$\$

$$P\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y} = c\right) = 1$$

$$P\left(Y = \frac{\sigma_Y}{\sigma_X}X - c\sigma_Y\right) = 1.$$

\$\$

这就证明了：当 $\text{Corr}(X, Y) = 1$ 时， Y 与 X 几乎处处线性正相关。负相关是同理的。**概率才是本质**

也正是基于这一性质，我们可以归纳出相关系数的本质含义：

- 相关系数 $\text{Corr}(X, Y)$ 刻画了 X 与 Y 之间的线性关系强弱，因此也常称其为“线性相关系数”。
- 若 $\text{Corr}(X, Y) = 0$, 则称 X 与 Y 不相关, 不相关是指 X 与 Y 之间没有线性关系, 但 X 与 Y 之间可能有其他的函数关系, 譬如平方关系、对数关系等 (**那如何才能像确定线性不相关一样确定其他函数不相关呢?**)。
- 若 $\text{Corr}(X, Y) = 1$, 则称 X 与 Y 完全正相关; 若 $\text{Corr}(X, Y) = -1$, 则称 X 与 Y 完全负相关。所谓完全, 就是指 (X, Y) 点集上, 几乎每个点对应都是线性相关的 (概率 = 1)
- 若 $0 < |\text{Corr}(X, Y)| < 1$, 则称 X 与 Y 有“一定程度”的线性关系, 相关程度都是有限的, 有些点对应相关性不强 (概率 < 1)

- $|Corr(X, Y)|$ 越接近于 1，则线性相关程度越高；
- $|Corr(X, Y)|$ 越接近于 0，则线性相关程度越低。

3.4 条件分布与条件期望

现实生活中，二维随机变量 (X, Y) 之间往往不是独立的（甚至说世界上根本不存在独立的两个随机变量<参考贝叶斯公式>），而是相互依存的，为了研究随机变量之间的依存关系，把条件概率引入设计相关的数学工具是一个很好的思路。

3.4.1 条件分布

条件分布的现实含义是什么？对二维随机变量 (X, Y) 而言，**所谓随机变量 X 的条件分布，就是在给定 Y 取某个值的条件下 X 的分布 $F_{X|Y}(x|y)$** 。比如，记 X 为人的体重， Y 为人的身高，则 X 与 Y 之间一定有相依关系，现在如果限定 $Y = 1.7(\text{m})$ ，在这个条件下体重 X 的分布 $F_{X|Y}(x|1.7)$ 显然与 X 的无条件分布（无此限制下体重的分布） $F_X(x)$ 会有很大的不同，一个显然的结论可能是条件分布下的均值较比无条件分布下的均值要大。之后所有和条件分布相关的讨论都是来自于随机事件的条件概率定义：

$$\begin{aligned} \text{\$\$} \\ P(A | B) &= \frac{P(AB)}{P(B)} \\ \text{\$\$} \end{aligned}$$

a. 离散随机变量的条件分布列

定义：设二维离散随机变量 (X, Y) 的联合分布列为

$$\begin{aligned} \text{\$\$} \\ p_{ij} &= P(X=x_i, Y=y_j), \quad i=1,2, \dots, j=1,2, \dots. \\ \text{\$\$} \end{aligned}$$

仿照条件概率的定义，我们很容易地给出如下离散随机变量的**条件分布列**的定义。对一切使 $P(Y = y_j) = p_{\cdot j} = \sum_{i=1}^{+\infty} p_{ij} > 0$ 的 y_j ，

$$\begin{aligned} \text{\$\$} \\ p_{i | j} &= P(X=x_i | Y=y_j) = \frac{P(X=x_i, Y=y_j)}{P(Y=y_j)} = \frac{p_{ij}}{p_{\cdot j}}, \\ &\quad i=1,2, \dots \\ \text{\$\$} \end{aligned}$$

为给定 $Y = y_j$ 条件下 X 的条件分布列，因此条件分布列的个数可以有很多个。

例子：两个例子，辅助理解条件分布列的概念

- 设随机变量 X 与 Y 相互独立，且 $X \sim P(\lambda_1)$, $Y \sim P(\lambda_2)$ 。在已知 $X + Y = n$ 的条件下，求 X 的条件分布。

直接套定义即可，可以将 $Z = X + Y$ 理解为与 X 相关的新变量

- 设在一段时间内进入某一商店的顾客人数 X 服从泊松分布 $P(\lambda)$ ，每个顾客购买某种物品的概率为 p ，并且各个顾客是否购买该种物品相互独立，求进入商店的顾客购买这种物品的人数 Y 的分布列。

这个地方明显条件概率比较好求, $P(Y = k|X = m)$ 就是一个二项分布。然后用全概率公式求解出来 $P_Y()$ 即可。

这也告诉了我们, 如果直接求概率分布列不好求的话, 可以借助条件概率来完成, 这和事件的思想是一致的。

b. 连续随机变量的条件密度函数

定义: 不同于事件或者离散变量, 连续变量的条件概率公式分母直接算是 0, 这显然是不能继续往下推导的, 因此对连续随机变量的条件分布和条件密度函数进行定义时, 需要用到无穷小极限逼近。这个过程在此不赘述, 直接给出定义结论

对一切使 $p_Y(y) > 0$ 的 y , 给定 $Y = y$ 条件下 X 的条件分布函数和条件密度函数分别为

$$\begin{aligned} \&F_{X|Y}(x | y) = \int_{-\infty}^x \frac{p(u, y)}{p_Y(y)} du, \\ &p_{X|Y}(x | y) = \frac{p(x, y)}{p_Y(y)} \end{aligned}$$

需要注意条件分布函数和条件密度函数仍然是一个二元函数, 但是的核心变量是 x , 可以将 y 看作一个常量。 $F_{X|Y}(x | y_1)$ 和 $F_{X|Y}(x | y_2)$ 是两个不同的关于变量 X 的函数, 也就是说 $F_{X|Y}(x | y)$ 表示跟随 y 变化而变化的一簇分布函数, 和直觉相符。概率密度函数也是同理。

性质:

- 二维正态分布的边际分布和条件分布都是一维正态分布。其边际分布较为简单, 条件分布比较复杂。

3.4.2 条件数学期望

之前我们只探讨了条件分布 or 条件概率。举个例子, 基于现实世界中的观察, 我们认定人的身高和体重 (X, Y) 服从二维正态分布, 假设 $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, 进而我可以得到条件概率密度函数, 推算出一维随机变量, **注意到, 正态分布的条件分布不像边际分布那么简单! 这是很显然的, 条件分布的变化过程保留了变量相关性。**

$$\begin{aligned} \&(Y | X=x) \sim N(\mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1), \sigma_2^2(1 - \rho^2)) \\ & \end{aligned}$$

也就是说, 在总体的分布的基础上上, 我们得到对于身高 $X = x$ 的这一局部人群体重 Y 的条件分布。基于此, 就能顺势解答一个问题: **身高 $X = x$ 的这一局部人群体重均值是多少呢?** 明显对该条件分布求解均值就可以了, 这就是条件分数学期望的含义, 计算上就是条件分布的数学期望。身高 $X = x$ 的这一局部人群体重均值一定是关于 x 的函数而非 y 的函数, 因为在求均值的时候 y 已经被消掉了, 当条件 $X = x$ 确定下来后, 结果也就确定了, 当其变化时结果也会变化。基于此我们可以给出条件数学期望的如下定义

定义: 条件分布的数学期望 (若存在) 称为条件期望

$$\begin{aligned} \&E(X | Y=y) = \begin{cases} \sum_i x_i P(X=x_i | Y=y) & \text{if } (X, Y) \text{ are discrete random variables;} \\ \int_{-\infty}^{+\infty} x p(x | y) dx & \text{if } (X, Y) \text{ are continuous random variables.} \end{cases} \end{aligned}$$

\$\$

注意条件期望 $E(X|Y = y)$ 是 y 的函数，就和上面的例子所表现的那样，对于 y 的不同取值，条件期望 $E(X|Y = y)$ 也在变化。例如， X 表示中国成年人的身高，则 $E(X)$ 表示中国成年人的平均身高。若用 Y 表示中国成年人的足长，则 $E(X|y = y)$ 表示足长为 y 的中国成年人的平均身高，我国公安部门研究获得：

\$\$

$$E(X | Y=y) = 6.876 y$$

\$\$

这个公式对公安部门破案起着重要的作用，例如，测得案犯留下的足印长为 25.3cm，则由此公式可推算出此案犯身高约 174cm（如果辅之以假设检验，还可以推算出置信区间）。

性质：此处的条件期望性质可以类比全概率公公式进行理解。通过上述分析，我们可以记

\$\$

$$g(y) = E(X | Y = y)$$

\$\$

进一步还可以将条件期望看成是随机变量 Y 的函数，记为 $E(X|Y) = g(Y)$ ，而将 $E(X|Y = y)$ 看成是 $Y = y$ 时 $E(X|Y)$ 的一个取值，由此看出： $E(X|Y)$ 本身也是一个随机变量，它所代表的含义是按照 Y 将总体划分为不同的局部，这些局部中 X 的均值与 Y 的关系，始终谨记其中的 X 已经被均值操作消除了，只留下 Y 。基于此我们可以得到**重期望公式**

设 (X, Y) 是二维随机变量，且 $E(X)$ 存在，则

\$\$

$$E(X) = E(E(X | Y))$$

\$\$

这个公式是极为有用的，两次数学期望代表了加权平均的加权平均。也就是想要求总体的某个特征的均值时，可以先将总体按照某一个特征划分成多个小的局部个体，先在所有的局部内部加权平均求出该特征的均值，然后再在总体分局部这一层面对已经求出的局部均值再平均。更准确的说法如下：要求在一个取值于很大范围上的指标 x 的均值 $E(X)$ ，这时会遇到计算上的各种困难。为此，我们换一种思维方式，去找一个与 X 有关的量 Y ，用 Y 的不同取值把大范围划分成若干个小区域，先在小区域上求 X 的平均，再对此类平均求加权平均，即可得到大范围上 X 的平均 $E(X)$ 。

如要求全校学生的平均身高，可先求出每个班级学生的平均身高，然后再对各班级的平均身高作加权平均，其权重就是班级人数在全校学生中所占的比例。

其具体应用形式如下

- 如果 Y 是一个离散随机变量，则

\$\$

$$E(X) = \sum_j E(X | Y=y_j) P(Y=y_j)$$

\$\$

- 如果 Y 是一个连续随机变量，则

\$\$

$$E(X) = \int_{-\infty}^{+\infty} E(X | Y=y) p_Y(y) dy$$

\$\$

不难发现：该公式的形式和全概率公式简直如出一辙，可以理解为“全期望公式”，**核心思路都是分而治之然后加和**。通过几个典型的例子来深入理解一下：

- 口袋中有编号为 $1, 2, \dots, n$ 的 n 个球，从中任取 1 球。若取到 1 号球，则得 1 分，且停止摸球。若取到 i 号球 ($i \geq 2$)，则得 i 分。且将此球放回，重新摸球。如此下去，试求得到的平均总分数。（分成 n 种情况，求条件期望后加总）
- 随机个数的随机变量和的数学期望：设 X_1, X_2, \dots 为一列独立同分布的随机变量，随机变量 N 只取正整数值，且 N 与 $\{X_n\}$ 独立，请证明：

\$\$

$$E(\sum_{i=1}^N X_i) = E(X_1) E(N).$$

\$\$

【证明】

\$\$

$$\begin{aligned} & E(\sum_{i=1}^N X_i) \\ &= \sum_{n=1}^{+\infty} [E(\sum_{i=1}^n X_i | N=n) \cdot P(N=n)] \\ &= \sum_{n=1}^{+\infty} [n E(X_1) \cdot P(N=n)] \\ &= E(X_1) \sum_{n=1}^{+\infty} [n \cdot P(N=n)] \\ &= E(X_1) E(N) \end{aligned}$$

\$\$

也正是基于这个性质，我们才有了一些所谓的显而易见的结论（这些“显然”的结论如果细究起来真的要人命）

- 设一天内到达某商场的顾客数 N 是仅取非负整数值的随机变量，又设进入此商场的第 i 个顾客的购物金额为 X_i ，可认为诸 X_i 是独立同分布的随机变量， N 与 X_i 相互独立，则一天营业额的期望值为**来的人数的期望 * 每一个人消费的期望**
- 一只昆虫一次产卵数 N 服从参数为 λ 的泊松分布，每个卵能成活的概率是 p ，可设 X_i 服从 $0-1$ 分布，而 $\{X_i = 1\}$ 表示第 i 个卵成活，则一只昆虫一次产卵后的平均成活卵数为**产卵个数的期望 * 每个卵存活的期望**

3.5 考点与典例

3.5.1 考点 1 联合分布函数的概念和计算

1. (离散分布直接画表)

【例 3.3】 袋中有一个红球,两个黑球,三个白球. 现有放回地从袋中取两次,每次取一个,求以 X, Y, Z 分别表示两次取球所取得的红、黑与白球的个数.

- (I) 求 $P\{X = 1 | Z = 0\}$;
(II) 求二维随机变量 (X, Y) 的概率分布.

2. (连续随机变量直接套概念)

【例 3.9】 设随机变量 X 服从参数 $\lambda = 1$ 的指数分布,若 $Y = X^2$,求二维随机变量 (X, Y) 的分布函数 $F(x, y)$.

3. (综合性计算题目, 在联合分布的基础上计算边际分布 + 条件分布)

【例 3.14】 二维随机变量 (X, Y) 的联合概率密度为 $f(x, y) = \begin{cases} 6x^2, & 0 < y < 1, |x| < y, \\ 0, & \text{其他.} \end{cases}$

- (I) 求边缘概率密度 $f_Y(y)$;
(II) 求条件概率密度 $f_{X|Y}(x|y)$;
(III) 求概率 $P\left\{-1 < X < \frac{1}{3} \middle| Y = \frac{1}{2}\right\}$;
(IV) 求概率 $P\left\{-1 < X < \frac{1}{3} \middle| Y \leq \frac{1}{2}\right\}$.

第四问其实有两种解决方案:

- 顺延第三问, 通过条件概率分布的含义(即条件概率分布是关于条件和变量的函数)进行计算, 将 Y 分为两段进行积分计算。【但是这种计算的难度很大】
- 直接用定义, 得到分式, 直接积分求解即可。【这种方法就是在条件为范围的时候的标准解题思路】

4. 均值方差的计算

设二维随机变量 (X, Y) 的联合概率密度为

$$f(x, y) = \begin{cases} \frac{2}{\pi}(x^2 + y^2), & x^2 + y^2 \leq 1 \\ 0, & \text{其它} \end{cases}$$

- (1) 求 X 与 Y 的方差;
(2) 求 X 与 Y 是否相互独立;
(3) 求 $Z = X^2 + Y^2$ 的概率密度.

第一问暴力求解不可以, 但是第一步求 X 和 Y 的边际分布过于复杂。因此我们直接从多维随机变量函数的均值出发, 将 X 看作 $g(X, Y)$, 这样的话求 X 的均值就等价于直接二元积分了。当然我们也可以换个思路, 所谓 X 的均值无非就是 X 所有的取值进行加权平均, 也可以转换成二元积分的形式。

3.5.2 考点 2 二维正态分布的性质和性质

永远要谨记，同时服从正态分布的两个随机变量的组合不一定服从二维正态分布，进而无法套用以下的结论。因此一旦涉及到相关的概念辨析，只需要思考两个变量的组合是否服从二维正态分布即可！

二维随机变量中，最有可能考察的就是二维正态分布的性质，在本章各个章节处，我也都做了相关的标记，在此再进行一下集中的整理。设 $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ ，则：

- 两个边缘分布服从正态分布，即 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$ 【3.2】
- X 与 Y 的相互独立的充分必要条件为 $\rho = 0$ ，也即二者不相关等价于二者独立 【3.3】
- X 与 Y 的线性组合 $aX + bY$ 服从一维正态分布：

\$\$

$$aX + bY \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\rho\sigma_1\sigma_2)$$

\$\$

这个是很显然的，借助随机变量间均值方差的计算性质即可得出【3.3】。

- X 与 Y 的非零线性组合 $(aX + bY, cX + dY)$ 仍服从二维正态分布。
- 最后，如果 (X, Y) 不满足二维正态分布的话，上述的结论都是不一定成立的！！！

1. (有关均匀分布的) 直接用面积即可。

但是记住在高金的那道笔试题目如果 X, Y, Z 独立同分布，均服从 $(0 \sim 1)$ 之间的均匀分布，试问 $P\{X + Y < Z\}$ 答案应该是 $\frac{1}{4}$

现在回看这道题其实是很简单的：

- 先根据独立，得到 (X, Y) 的联合分布
- 然后根据定义，求解 $F_{X+Y}(t)$ 这一个分布，甚至不用求导数得到概率密度
- 之后根据条件乘法公式

\$\$

$$\begin{aligned} P(X + Y < Z) &= \int_0^1 P(X+Y < z | Z=z) P(Z=z) dz \\ &= \int_0^1 F_{X+Y}(z) f_Z(z) dz \end{aligned}$$

\$\$

2. (二维正态分布的简便计算) 当我们解决有关二维正态分布的问题时，思考问题的顺序应该是：

- 先想能不能分离，然后用性质进行配方 + 除系数 + 添因子（绝大多数都可以解决掉）
- 思考二维正态分布的概率密度函数，往上面靠，凑出五个参数
- 直接从概率密度函数出发进行 tough 的运算

【例 3.7】 设二维随机变量 (X, Y) 的概率密度为

$$f(x, y) = Ae^{-2x^2+2xy-y^2} \quad (-\infty < x < +\infty, -\infty < y < +\infty),$$

求常数 A 及条件概率密度 $f_{Y|X}(y|x)$.

对这个题来说，如果直接进行积分，那实在是太难了。凑出五个参数也并不容易，因此还是想看看能不能直接分离。而是配方 + 除系数 + 添因子

\$\$

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{+\infty} f(x, y) dy \\ &= Ae^{-x^2} \sqrt{\pi} \int_{-\infty}^{+\infty} e^{-y^2} dy \\ &= Ae^{-x^2} \sqrt{\pi} \end{aligned}$$

\$\$

这个思路是极为重要的，对上式再积分一次，继续配方 + 除系数 + 添因子：

\$\$

$$\begin{aligned} \int_{-\infty}^{+\infty} f_X(x) dx &= \int_{-\infty}^{+\infty} Ae^{-x^2} \sqrt{\pi} dx \\ &= A\sqrt{\pi} \int_{-\infty}^{+\infty} e^{-x^2} dx \\ &= 1 \end{aligned}$$

\$\$

类似地我们可以求解 Y 的边际分布

\$\$

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{+\infty} f(x, y) dx \\ &= Ae^{-y^2} \sqrt{\pi} \int_{-\infty}^{+\infty} e^{-x^2} dx \\ &= Ae^{-y^2} \sqrt{\pi} \end{aligned}$$

\$\$

3. (二维正态分布的性质)

设随机变量 X 和 Y 都服从正态分布，且它们不相关，则

【C】

- (A) X 和 Y 一定独立. \times
(C) X 和 Y 未必独立.

- \times (B) (X, Y) 服从二维正态分布.
 \times (D) $X + Y$ 服从一维正态分布.

例 下列命题中说法错误的个数是

(C)

- (1) 若随机变量 X, Y 分别服从正态分布, 则 $X + Y$ 一定服从正态分布
- (2) 若随机变量 X, Y 分别服从正态分布, 且相互独立, 则 $X + Y$ 一定服从正态分布
- (3) 若随机变量 (X, Y) 服从二维正态分布, 则 $X + Y$ 一定服从正态分布
- (4) 若随机变量 (X, Y) 服从二维正态分布, 仅当 X, Y 相互独立时 $X + Y$ 服从正态分布

(A) 0

(B) 1

(C) 2

(D) 3

$$f(x,y) = f(x)f(y)$$

4. (二维正态分布的独立性特性) 一旦有了独立, 就可以从一维随机变量向二维拓展。

- (5) 设 X, Y, Z 相互独立, $X \sim N(1, 2)$, $Y \sim N(2, 2)$, $Z \sim N(3, 7)$, 记 $a = P\{X < Y\}$, $b = P\{Y < Z\}$, 则
- (A) $a > b$.
(B) $a < b$.
(C) $a = b$.
(D) a, b 大小关系不确定.

3.5.3 考点 3 多维随机变量函数的分布 (始终从定义出发!)

1. (混合分布) 从定义出发, 带入离散量, 这道题就迎刃而解了

【例 3.26】 设随机变量 X_1, X_2, X_3 相互独立, 其中 X_1 与 X_2 均服从标准正态分布, X_3 的概率分布为 $P\{X_3 = 0\} = P\{X_3 = 1\} = \frac{1}{2}$, $Y = X_3 X_1 + (1 - X_3) X_2$.

- (I) 求二维随机变量 (X_1, Y) 的分布函数, 结果用标准正态分布函数 $\Phi(x)$ 表示;
(II) 证明随机变量 Y 服从标准正态分布.

2. (混合分布) 不要把条件分布引入进来, 不然就复杂了, 因为条件分布并不简单, 反而是联合分布更好理解

设二维随机变量 (X, Y) 在区域 $D = \{(x, y) \mid 0 < x < 1, x^2 < y < \sqrt{x}\}$

上服从均匀分布, 令 $U = \begin{cases} 1, & X \leq Y, \\ 0, & X > Y. \end{cases}$

- (I) 写出 (X, Y) 的概率密度;
(II) 问 U 与 X 是否相互独立? 并说明理由;
(III) 求 $Z = U + X$ 的分布函数 $F(z)$.

第四章 大数定律与中心极限定理

在此之前，对概率的描述始终停留在理论层面，也即在进行了合理假设后，从事件的理论性质出发，推算该事件的理论概率及其相关性质（假设硬币是均匀的，那么扔一次朝上的概率应该是 $1/2$ ，扔 n 次朝上的总次数服从二项分布 $(n, 1/2)$ ）。

从这一章起，我们开始与现实世界相连接。在现实世界中，我们永远也无法观测到事件概率的大小（对于简单如抛硬币的事件而言，其实际情况也不可能如假设得那样完美；对于其他的复杂事件，概率甚至都是不可求的），我们只能通过进行大量实验的方法用频率推算实际概率（大数定律），又或者用分布来近似模拟分布（中心极限定理）

4.1 随机变量序列的两种收敛性

4.1.1 依概率收敛

独自一人对着大雨思索，依概率收敛的本质所在。

【定义】：设 $\{X_n\}$ 为一随机变量序列， X 为一随机变量，如果对于任意的 $\varepsilon > 0$ ，有

$$\begin{aligned} &\text{\$ \$} \\ &P(|X_n - X| \geq \varepsilon) \rightarrow 0 \quad (n \rightarrow \infty), \\ &\text{\$ \$} \end{aligned}$$

则称序列 $\{X_n\}$ 依概率收敛于随机变量 X ，记作 $X_n \xrightarrow{P} X$ 。其含义为： X_n 对 X 的绝对偏差不小于任一给定量的可能性将随着 n 的增大而越来越小，或者说出现大偏差的概率越来越小。而当随机变量 X 退化为一个确定的常数 c 时，上式就可以转化为

$$\begin{aligned} &\text{\$ \$} \\ &P(X_n = c) \xrightarrow{P} 1 \\ &\text{\$ \$} \end{aligned}$$

也即： $X_n \xrightarrow{P} c$ 。要理解清楚这个定义，就必须回答以下几个问题：

- 随机变量序列的含义是什么？
- $n \rightarrow \infty$ 怎么理解？
- 收敛于一个随机变量的含义是什么，为什么不是收敛于一个确定的值？（按照之前所说的大量实验的频率收敛于概率，概率应该是一个确定的值吗？）

我们用一个抛硬币的例子，对上述定义做一个全面的诠释。记随机变量 X_n 表示 n 次硬币实验中正面朝上的频率（注意这个频率不是一个数字，而是一个随机变量，其是和硬币朝上的概率相关的一个随机变量，这一点必须要理解清楚！），那么 $\{X_n\}, n = 1, 2, \dots$ 表示顺次做 n 次硬币实验其中每一次正面朝上的频率， $n \rightarrow \infty$ 表示做无穷多次实验。常识告诉我们，无穷多次扔硬币实验后，正面朝上发生的频率应该收敛于概率 X 。但是现在请大家扪心自问一下，这个频率是一个确定的值吗？现实世界中肯定不是！因为硬币不可能没有变化，所以严格意义上来说，硬币正面朝上的概率 X 也是一个随机变量！因此，更加直观的理解就是，无穷多次扔硬币实验后，正面朝上发生的频率 X_n 这一随机变量的取值与正面朝上的概率 X 这一随机变量的对应取值的差别较大这一事件发生的概率趋近于 0。【点点收敛】当然，我们最常讨论的是概率 X 为一常数 p 的情况，这种情况能够尽可能地刻画真实世界，而有简化了思考的复杂性。

【性质】：满足加减乘除四则运算规律，在此不做说明。

4.1.2 按分布收敛

这个概念之前是从来都没有接触过的，之所以在这一教材中出现，其是为了后续推导中心极限定理打下基础。

来看按分布收敛的**定义**：设随机变量 X_1, X_2, \dots 的分布函数分别为 $F_1(x), F_2(x), \dots$ ，随机变量 X 的分布函数为 $F(x)$ ，若对 $F(x)$ 的任一连续点 x 都有

$$\begin{aligned} & \text{\$} \\ & \lim_{n \rightarrow \infty} F_n(x) = F(x), \\ & \text{\$} \end{aligned}$$

则称 $\{F_n(x)\}$ 弱收敛于 $F(x)$ 【这个地方之所以叫做弱收敛，就是因为较比依概率收敛而言，依分布收敛只对连续点做了定义！】，记作

$$\begin{aligned} & \text{\$} \\ & F_n(x) \xrightarrow{W} F(x) \\ & \text{\$} \end{aligned}$$

也曾相应的随机变量序列 $\{X_n\}$ 按分布收敛与 X ，记作

$$\begin{aligned} & \text{\$} \\ & X_n \xrightarrow{L} X \\ & \text{\$} \end{aligned}$$

4.2 大数定律

大数定律的本质：讨论在什么情况下，随机变量序列均值 $\frac{1}{n} \sum_{i=1}^n X_i$ 依概率收敛为数学期望 $\frac{1}{n} \sum_{i=1}^n E(X_i)$ 。考虑的是一维随机变量（可以将算术平均看作 $Y = g(X_1, X_2, \dots, X_n)$ ）向常数的依概率（期望本质是一个确定的常数）收敛。

大数定律的定义：设有一随机变量序列 $\{X_n\}$ ，假设其对任意的 $\varepsilon > 0$ ，有：

$$\begin{aligned} & \text{\$} \\ & \lim_{n \rightarrow +\infty} P \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n E(X_i) \right| < \varepsilon \right) = 1 \\ & \text{\$} \end{aligned}$$

则称随机变量序列 $\{X_n\}$ 服从大数定律。那我们现在自然就有一个问题，什么条件下的随机变量序列 $\{X_n\}$ 才服从大数定律呢？根据所要求条件的不同，引出了后续大数定律的 4 种形式。

4.2.1 伯努利大数定律

【伯努利大数定律】若 $\{X_n\}$ 是独立的两点分布随机变量序列，则 $\{X_n\}$ 服从大数定律。伯努利大数定律给 $\{X_n\}$ 设定的条件有两个：

- 随机变量序列 $\{X_n\}$ 中每一个随机变量 X_i 相互独立且同分布
- 其共同分布为两点分布 $b(1, p)$

\$\$

$X_i = \begin{cases} 1, & \text{第 } i \text{ 次试验中事件 } A \text{ 发生}, \\ 0, & \text{第 } i \text{ 次试验中事件 } A \text{ 不发生}, \end{cases}$ $i=1, 2, \dots, n, \dots,$

\$\$

伯努利大数定律，为在现实生活中通过频率确定概率提供理论依据，因为抛硬币、合格率等问题，都可以抽象成独立同分布的两点分布。

【证明】从大数定律形式出发，借助切比雪夫不等式（往往都需要借助该不等式，因为第二项正好是第一项的均值！）

从大数定律的定义形式出发

\$\$

$\lim_{n \rightarrow +\infty} P \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n E(X_i) \right| > \varepsilon \right) = \lim_{n \rightarrow +\infty} P \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - p \right| > \varepsilon \right) < \varepsilon$

\$\$

又因为

\$\$

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

\$\$

因此，由切比雪夫不等式可以得到

\$\$

$$P \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - p \right| > \varepsilon \right) \geq 1 - \frac{\text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right)}{\varepsilon^2} = 1 - \frac{p(1-p)}{n \varepsilon^2}$$

\$\$

又因为 $n \rightarrow +\infty$ 时，右端趋于 1，因此由夹逼准则可以得到大数定律定义式

\$\$

$$\lim_{n \rightarrow +\infty} P \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n E(X_i) \right| > \varepsilon \right) = 1$$

\$\$

4.2.2 切比雪夫大数定律

【切比雪夫大数定律】若 $\{X_n\}$ 为一列两两不相关的随机变量序列，每个 X_i 的方差存在，且有共同的上界，即 $Var(X) \leq c, i = 1, 2, \dots$ ，则 $\{X_n\}$ 服从大数定律。切比雪夫大数定律给 $\{X_n\}$ 设定的条件也是两个：

- 随机变量序列 $\{X_n\}$ 中每一个随机变量 X_i 两两不相关（只需要不相关即可！不需要独立）
- 每一个随机变量 X_i 的方差存在且上界相同。

可以看到，其要求的条件比伯努利大数定律少了很多，伯努利大数定律就是切比雪夫大数定律的一个特殊形式。

【证明】思路同上

因为 $\{X_n\}$ 两两不相关，故

$$\begin{aligned} & \text{Var} \left(\frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) \leq \frac{c}{n}. \end{aligned}$$

再由切比雪夫不等式得到：对任意的 $\varepsilon > 0$ ，有

$$\begin{aligned} & \text{P} \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n E(X_i) \right| > \varepsilon \right) \leq \frac{\text{Var} \left(\frac{1}{n} \sum_{i=1}^n X_i \right)}{\varepsilon^2} \leq \frac{c}{n\varepsilon^2}. \end{aligned}$$

于是当 $n \rightarrow +\infty$ 时，有

$$\begin{aligned} & \lim_{n \rightarrow +\infty} \text{P} \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n E(X_i) \right| > \varepsilon \right) = 0. \end{aligned}$$

4.2.3 马尔可夫大数定律

【马尔可夫大数定律】若 $\{X_n\}$ 满足 $\frac{1}{n^2} \text{Var}(\sum_{i=1}^n X_i) \rightarrow 0$ ，则 $\{X_n\}$ 服从大数定律。马尔可夫大数定律只给 $\{X_n\}$ 设定了一个条件，我们也将其称为马尔可夫条件。这个条件是直接通过切比雪夫不等式得到的，因此马尔可夫大数定律直接通过切比雪夫不等式证明。

可以看到，马尔可夫大数定律进一步将条件放松，已经没有了任何对随机变量分布的要求，只有对方差的要求，切比雪夫大数定律是马尔可夫大数定律的一个特殊形式。

【例子】设 $\{X_n\}$ 为一同分布、方差存在的随机变量序列，且 X_n 仅与 X_{n-1} 和 X_{n+1} 相关，而与其他的 X_i 不相关。试问该随机变量序列 $\{X_n\}$ 是否服从大数定律？遇到这种不涉及分布条件的，一眼就知道要用马尔可夫大数定律进行证明，直接考虑马尔可夫条件。

$\{X_n\}$ 为相依随机变量序列，考虑其马尔可夫条件

\$\$

$$\frac{1}{n^2} \operatorname{Var} \left(\sum_{i=1}^n X_i \right) = \frac{1}{n^2} \left(\sum_{i=1}^n \operatorname{Var}(X_i) + 2 \sum_{i=1}^{n-1} \operatorname{Cov}(X_i, X_{i+1}) \right)$$

\$\$

记 $\operatorname{Var}(X_i) = \sigma^2$, 则 $|\operatorname{Cov}(X_i, X_j)| \leq \sigma^2$, 于是有

\$\$

$$\frac{1}{n^2} \operatorname{Var} \left(\sum_{i=1}^n X_i \right) \leq \frac{1}{n^2} \left(n \sigma^2 + 2(n-1) \sigma^2 \right) \rightarrow 0, \quad (n \rightarrow +\infty),$$

\$\$

即马尔可夫条件成立, 故 $\{X_n\}$ 服从大数定律。

4.2.4 辛钦大数定律

不再是进一步放松条件, 而是换个角度放松条件, 之前的切比雪夫大数定律和马尔可夫大数定律都是从方差角度放松条件, 而此处的辛钦大数定律则是从均值放松条件。

【辛钦大数定律】 设 $\{X_n\}$ 为一独立同分布的随机变量序列, 若 X_i 的数学期望存在, 则 $\{X_n\}$ 服从大数定律, 即对任意的 $\varepsilon > 0$ 。辛钦大数定律从均值放松条件, 设定了两个条件限制:

- $\{X_n\}$ 为一独立同分布的随机变量序列
- X_i 的数学期望存在

可以看出伯努利大数定律是辛钦大数定律的一个特殊形式, 其他二者则不存在一般特殊关系。

由于辛钦大数定律的证明需要借助特征函数, 在此不给出证明过程。

4.3 中心极限定理

中心极限定理的本质: 讨论在什么情况下, 相互独立的随机变量序列和 $Y = \sum_{i=1}^n X_i$ 这个一维随机变量的分布会收敛于正态分布。一旦能够用正态分布对一些特殊分布进行近似, 那么研究问题的难度就会大大降低, 这也是我们研究其的根本目的所在。

中心极限定理的定义: 如果按照我们之前推算多维随机变量函数分布的做法, 对于 2 维、3 维等情况推出 Y 的精确分布可能并不难, 但对于 10000 维就太复杂了。通过对小维度随机变量的规律探索, 发现序列和这一随机变量会随着所加和序列维度的增加愈发接近于正态分布。正是基于这个发现, 我们想探究 n 维随机变量序列和的分布与正态分布之间的关系, 更具体的是看标准化后的随机变量

\$\$

$$Y_n^* = \frac{Y_n - E(Y_n)}{\sqrt{\operatorname{Var}(Y_n)}}$$

\$\$

在什么条件下才逼近正态分布。根据所要求条件的不同, 引出中心极限定理的在独立同分布与独立不同分布下的 4 种形式。

4.3.1 独立同分布下的中心极限定理

a. 林德伯格-莱维中心极限定理

【定义】 设 $\{X_n\}$ 是独立同分布的随机变量序列，且每一项均值和方差都存在 $E(X_n) = \mu, Var(X_n) = \sigma^2 > 0$ ，记

$$\begin{aligned} & \$ \\ Y_n^* &= \frac{X_1 + X_2 + \dots + X_n - n\mu}{\sigma\sqrt{n}}, \\ & \$ \end{aligned}$$

则当 n 充分大时， Y_n^* 服从正态分布。也即对任意实数 y ，有

$$\begin{aligned} & \$ \\ \lim_{n \rightarrow +\infty} P(Y_n^* \leq y) &= \Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y e^{-t^2/2} dt. \\ & \$ \end{aligned}$$

可以看到林德伯格-莱维中心极限定理只对随机变量序列 $\{X_n\}$ 做了如下三点要求：

- 随机变量序列 **独立同分布**
- 随机变量序列中每一个变量的 **方差都存在**，而不 care 其分布具体是什么形式
- n **充分大**

一旦满足这三个要求，我们就可以用正态分布去逼近随机变量和的分布。

【应用】 该中心极限定理在生活中无处不在，最常见的是如下两个：

- 正态随机数的生成：计算机可以借助生成足够多的 $(0, 1)$ 随机分布，形成随机变量序列，然后将变量和作为正态分布。
- 误差分析：每一次测量的误差都服从均匀分布（因为测量精度是有限的，有的时候我们不得不进行四舍五入），多次测量后求平均值的方式可以得到误差的分布近似服从正态分布，进而可以得到误差的大小以及置信区间！

b. 棣莫弗-拉普拉斯中心极限定理

【定义】 是林德伯格-莱维中心极限定理的一种特殊情况，即将同分布这一条件 **限定为二值分布**。设 n 重伯努利试验中，事件 A 在每次试验中出现的概率为 p ($0 < p < 1$)，记 S_n 为 n 次试验中事件 A 出现的次数，且记

$$\begin{aligned} & \$ \\ Y_n^* &= \frac{S_n - np}{\sqrt{np(1-p)}} \\ & \$ \end{aligned}$$

则当 n 充分大时， Y_n^* 服从正态分布。可以看到，棣莫弗-拉普拉斯中心极限定理同样也对随机变量序列 $\{X_n\}$ 做了三点要求：

- 随机变量序列 **独立同分布**
- care 到每一个变量都服从 **二值分布**
- n **充分大**

更为直观的描述该条件下大数定理的方式是从二项分布的近似替代的角度，对于随机变量 $X \sim B(n, p)$ 如果在 n 特别大 p 又特别小的情况下随机变量 X 的概率是很难测度的，因此我们往往需要进行近似替代，有两种替代方式：

- 如果 p 较小，那么就选择使用泊松分布进行替代（离散分布替代离散分布，计算的时候精度更高）
- 如果 $np > 5$ ，选用正态分布近似较好（连续分布替代离散分布，计算的时候要做一定的修正左端点 -0.5 右端点 +0.5）

【应用】 生活中许多现象都是服从二项分布的，因此该中心极限定理的应用十分广泛，具体集中在对下式中三个超参的探讨

$$\begin{aligned} & \text{\$\$} \\ & P\left(\frac{S_n - np}{\sqrt{np(1-p)}} < y\right) \approx \Phi(y) = \beta \\ & \text{\$\$} \end{aligned}$$

- n, y 已知，求 β : 100 个灯泡，每个坏的概率是 0.06，亮 90 个以上的概率是多少？
- n, β 已知，求 y : 100 个灯泡，每个坏的概率是 0.06，有 95% 的把握灯泡最少亮多少个？
- y, β 已知，求 n : 灯泡每个坏的概率为 $p(unkonwn)$ ，问需要抽检多少个灯泡，才能保证有 95% 的把握认定抽检出的坏灯泡频率与概率相差不超过 1%？【如果 n 足够大，伯努利大数定律告诉我们有 100% 的把握确定二者之差几乎为 0，但如果 n 不足够大这个误差概率就需要采用中心极限定理进行估计了】

这些应用本质都是对超参概念式的考察，如果出现直接把超参概念式写出即可！

4.3.2 独立不同分布的中心极限定理

较比于独立同分布下的中心极限定理，独立不同分布的中心极限定理较为复杂，涉及到的概念也很多，因此在这只是给出概念。

因为每个随机变量的分布不同了，所以就要单独考虑每一个变量， Y_N^* 就被改写为如下形式，如果随机变量序列中的每一项随机变量都有有限的数学期望和方差， $E(X_i) = \mu_i$, $Var(X_i) = \sigma_i^2$, 那么我们要考虑的是以下随机变量的分布是否可以被近似为正态分布

$$\begin{aligned} & \text{\$\$} \\ & Y^*_N = \frac{\sum_{i=1}^n (X_i - \mu_i)}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \\ & \text{\$\$} \end{aligned}$$

虽然独立不同分布的现象在生活中更为常见（几乎所有的情况都是），但是由于工具底层原理的复杂性，所以我们一般不对这些概念需要有极为准确的记忆，只需要知道其所表达的含义是什么即可。我们在这给出一个经典的独立但不同分布的现实情形

一份考卷由 99 个题目组成，并按由易到难顺序排列，某学生答对第 1 题的概率为 0.99，答对第 2 题的概率为 0.98，一般地，他答对第 i 题的概率为 $1 - i/100$, $i = 1, 2, \dots$ 。假如该学生回答各题目是相互独立的，并且要正确回答其中 60 个题目以上（包括 60 个）才算通过考试。试计算该学生通过考试的可能性多大？

显然发现：每一道题答题情况 $\{X_i\}, i = 1, \dots, 99$ 都服从概率为 $1 - i/100$ 的二值分布，所以是一个典型的独立但不同分布的情况，因此我们计算答案概率

$$\begin{aligned} & \text{\$\$} \\ & P\left(\sum_{i=1}^{99} X_i \geq 60\right) \end{aligned}$$

\$\$

就需要先说明是满足相关的条件，才能进行正态分布近似（一般是证明满足李雅普诺夫条件）。

a. 林德伯格中心极限定理

设随机变量序列 $\{X_n\}$ 满足如下林德伯格条件，即对任意的 $\tau > 0$ ，有

\$\$

$$\lim_{n \rightarrow +\infty} \frac{1}{\tau^2} \sum_{i=1}^n \int_{-\infty}^{\infty} |x - \mu_i|^{2+\delta} p_i(x) dx = 0$$

\$\$

那么随机变量 Y_N^* 就服从正态分布！（这个条件十分难以验证！）

b. 李雅普诺夫中心极限定理

设随机变量序列 $\{X_n\}$ 满足如下条件，若存在 $\delta > 0$ ，满足

\$\$

$$\lim_{n \rightarrow +\infty} \frac{1}{n^{2+\delta}} \sum_{i=1}^n E[(\|X_i - \mu\|^{2+\delta})] = 0,$$

那么随机变量 Y_N^* 就服从正态分布！（还是可能被用到的）

4.4 考点与典例

4.4.1 考点 1 切比雪夫不等式协助求解概率范围

没有其他考点，记住公式后直接套公式即可

设随机变量 X_1, X_2, \dots, X_n 独立同分布，且 X_1 的 4 阶矩存在。设 $\mu_k = E(X_1^k)$ ($k = 1, 2, 3, 4$)，则

由切比雪夫不等式，对 $\forall \varepsilon > 0$ ，有 $P\left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i^2 - \mu_2 \right| \geq \varepsilon \right\} \leq (\quad)$

- A. $\frac{\mu_4 - \mu_2^2}{n\varepsilon^2}$ B. $\frac{\mu_4 - \mu_2^2}{\sqrt{n}\varepsilon^2}$ C. $\frac{\mu_2 - \mu_1^2}{n\varepsilon^2}$ D. $\frac{\mu_2 - \mu_1^2}{\sqrt{n}\varepsilon^2}$

4.4.2 考点 2 大数定理的概念与性质

1. （根据四大类大数定理规定的性质进行定义判断）一般来说如果是同分布就看辛钦（也就是说进一步判断均值是否存在），如果不同分布就看切比雪夫（也就是说进一步判断方差是否同上界）

(3) 随机变量 $X_1, X_2, \dots, X_n, \dots$ 相互独立且满足大数定律, 则 X_i 的分布可以是

- (A) $P\{X_i = m\} = \frac{c}{m^3}, m = 1, 2, \dots$
(B) X_i 服从参数为 $\frac{1}{i}$ 的指数分布.
(C) X_i 服从参数为 i 的泊松分布.
(D) X_i 的概率密度 $f(x) = \frac{1}{\pi(1+x^2)}$.

A、D 同分布, 套辛钦, 只需要看均值是否存在

B、C 不同分布, 套切比雪夫, 只需要看方差是否同上界

2. (从四个大数定律出发补充条件) 思路同上

(4) 设随机变量 $X_1, X_2, \dots, X_n, \dots$ 相互独立, 记 $Y_n = X_{2n} - X_{2n-1} (n \geq 1)$, 根据大数定律, 当

$n \rightarrow \infty$ 时, $\frac{1}{n} \sum_{i=1}^n Y_i$ 依概率收敛到零, 只要 $\{X_n, n \geq 1\}$ 满足

- (A) 数学期望存在. (B) 有相同的数学期望与方差.
(C) 服从同一离散型分布. (D) 服从同一连续型分布.

4.4.2 考点 3 中心极限定理标准化求解

(直接利用中心极限定理的概念进行标准化即可) 一定要把真正的随机变量序列单独地摘出来

(2) 设随机变量 X_1, X_2, \dots, X_{2n} 独立同分布, 且 $E(X_i) = D(X_i) = 1 (1 \leq i \leq 2n)$, 如果 $Y_n =$

$c \sum_{i=1}^n \frac{X_{2i} - X_{2i-1}}{\sqrt{n}}$, 则当常数 $c = \underline{\hspace{2cm}}$ 时, 根据独立同分布中心极限定理, 当 n 充分大时 Y_n 近似服从标准正态分布.

第五章 统计量及其分布

从这一章开始, 进入统计学。

首先我们先搞清楚, **统计所研究的问题究竟是什么?**

通过前面对概率论的学习, 我们发现如果随机变量 X 的分布函数已知, 那么其所有和概率相关的规律都可以被推算, 也就是说分布函数是对随机变量 X 最本质, 最正确的描述。但是! 人类真的能够完全准确刻画地随机变量的分布吗 (注意不是大概, 而是完全精准) ?

比如对于扔 n 次硬币朝上的次数 X 这样一个简单的随机变量，我们真的就能够 100% 把握说它一定服从二项分布 $(n, \frac{1}{2})$ ，这是不能的！因为我们缺少一个看似“显然”的前提：每扔一次朝上的概率是 $\frac{1}{2}$ 。而我们可以肯定，这个“显然”的前提并不一定 100% 出现，原因很简单，可能硬币制作的时候出了瑕疵，可能硬币本来就是三个面的。

因此，在现实世界中我们思考问题的方式应该是完全反过来的。在承认以下两个前提的基础上才能展开思考：

- 随机变量 X 的确服从一个确定的分布（其可能随着时间或者其他条件变化，但是当所有控制条件都给定时，它的形式是确定）
- 无法从根源上准确地得到这个分布的形式（无法基于全概率公式把所有的相关事件都找出来准确推导），只能从事实出发进行估计

回到扔硬币的例子，问题的描述方式就完全改变了，对于扔硬币正面朝上 X 这样一个简单的随机变量，“上帝”已然设定好每次实验正面朝上的概率是 p （这个概率是世界运作的规律，**是确定的，不随着人的意志而转移**，也就是说不管你实验多少次，如何实验，每次硬币正面朝上的概率都是 p ）理性上讲：世界上没有一个概率是独立的，都是条件概率的情况下，必须要通过全概率公式求解，但是谁又能把世间所有相关的事件都找出来呢？这样“上帝”也就间接设定 X 服从二值分布 $b(n, p)$ （这是确定的分布）。因此如果我们确定了 p ，所有的附加规律就都知晓了，就可以借助它帮助我们研究更加具体的结论。但是悲哀的事情是，我们永远也没办法从根源上准确破译它：

- 理论上：我们无法和“上帝”沟通（把所有的相关事件全部挖出来）从根源上把这个 p 给问出来，基于正面推导直接得到答案
- 实验上：这个规律是对于“总体”而言的，而总体是 ∞ 次硬币实验，我们没有办法把 ∞ 多的结果都拿到，基于现象观察间接得到答案

所以，现实中我们无论如何也得不到根源上的完全准确的答案，而总是尝试尽可能确定 p 也就是解决以下两个问题：

- p 大概（概率而非确定）落在什么范围内？
- 有多大的把握（概率而非确定）认为 p 满足我们设定的要求？（对抛硬币来说，就是有多大的把握认为“上帝”刚好设定 $p = 0.5$ ？）

这正是统计学研究的问题，和微积分那种发现确定规律的理论不同，统计旨在通过收集和分析现实生活中受到随机影响的数据（比如抛 $1 \sim 100$ 次硬币正面朝上的次数集合），以概率的思想，**以极大概率的结论逼近确定的规律**（硬币朝上的概率是 $1/2$ ）。统计的本质就是基于可观测到的数据挖掘世界的本质规律，这些规律的现实形式可能是多种多样的，但是从根源看就是**分布函数的具体形式**。

因为规律始终以分布的形式呈现在真实世界中，所以在概率的驱使下所观测到的数据肯定存在各种各样的噪声，如何透过现实现象（点在均值附近密集地散落）看到本质（这些点服从线性规律），是统计学始终在问的问题！

为了更好地理解，在此再举一个现实中的例子：某公司要采购一批产品，每件产品不是合格品就是不合格品，但该产品总有一个不合格品率 p （这个 p 可能是由于生产环境、机器各种各样的原因导致的，因此正向计算根本不可能，只能姑且认定其是“上帝”设定的）。由此，若从该批产品中随机抽取一件，用 X 表示这一件产品的不合格数，不难看出 X 服从一个二点分布 $b(1, p)$ ，但分布中的参数 p 却是不知道的。显然， p 的大小决定了该批产品的质量，它直接影响采购行为的经济效益。如果不格率大于 5% 那我就不买了，此时我必须对 p 做出判断，即有多大的把握认为 $p \leq 0.05$ ？如果把握很大，那我就买，反之不买。

但是永远不能做出这样的判断，那就是我有 100% 的把握认为 $p \leq 0.05$ ，因为我们无法从根源上准确地破译它，也不可能把所有的样本都检查完（因为做的事破坏性检查），只能尽可能大量的抽样数据进行推测，这显然始终是受到限制的！

后续几章的学习思路如下：

- 第五章讲清楚统计所研究的依附点是什么？统计研究依附的客体是什么（样本）以及理论落脚点是什么（统计量及其分布）
- 第六章（自下而上地）帮我们解答如何用现实中的数据来发掘本质规律？如何用样本估计总体参数
- 第七章（自上而下地）讲清楚了对规律进行了合理假设后，现实数据发生的概率是多少？自己尝试当一下“上帝”
- 第八章系统说明了如何将概率的思想应用到模型搭建中，帮我们认清更加复杂的现实规律

5.1 总体与样本

我们想要研究的规律一定是有载体（研究对象）的，比如

- 想研究扔硬币正面朝上的概率，这个规律的载体就是无穷多次的扔硬币实验
- 想研究一批产品的合格率，这个规律的载体就是这一批产品
- 想研究全世界人民的身高分布，这个规律的载体就是全世界人民
- 但如果只想研究全中国人民的身高分布，这个规律的载体就是全中国人民。

因此不同的规律有着不同的载体，如果连载体都找不对，规律探究就更是天方夜谭了。总体与样本就是“载体”的概念

5.1.1 总体与个体

总体和个体是载体（研究对象）的概念，所关注范围的不同，分成总体和个体。**总体**：研究对象的全体；**个体**：构成总体的每个成员。

- 想研究扔硬币正面朝上的概率，总体就是无穷多次实验（无限总体），个体就是每次实验
- 想研究一批产品的合格率，总体就是这一批产品（有限总体），个体就是每个产品
- 想研究全国人民的身高分布，总体就是全国人民，个体就是每一个人

5.1.2 样本

样本从底层来说就是随机抽取的总体的一部分，是我们研究所依附的内容。到这自然就有一个疑问了，明明是研究总体的规律，为什么不直接从总体入手，非得从总体中随机抽取一部分出来研究呢？

- 总体根本研究不了：比如抛硬币是无限总体，从理论上就不可行了；又比如全国人民的身高，理论上可行但实际上不能操作。
- 总体可以研究，但是是破坏性研究：比如合格率检验，必须要拆开，理论上可以但实际上检验完后都坏了，得到的结果没有价值。

所以我们只能从庞大的总体中抽出来一小部分进行研究，希望局部的特征和全局是一致的，进而用局部反映全局，用样本反映总体。

样本: 为了解总体随机变量 X 的分布，我们从总体中随机地抽取 n 个个体 X_1, X_2, \dots, X_n ，观测后记其指标值为 x_1, x_2, \dots, x_n ，则 X_1, X_2, \dots, X_n 称为总体的一个**样本**。在此必须要理解清楚，**样本的二重性（抽签思想）**，这是样本最重要的特性，是本质，是理解后续所有概念的基础（刚开始想的时候总有思维惯性，一个样本点就是一个值，怎么会是随机变量呢？这就完全忽略了随机抽取的过程）：

- 一方面，样本是从总体中随机抽取的，抽取后观测前无法预知它们的数值，**样本此时是一组随机变量的集合**，更确切的说，**样本是一个含义完全相同且独立同分布的随机变量的序列 $\{X_i\}$** ，本质上也是一个多维随机变量，这一点在从分统计量一节会有表现（这样就意味着大数定律、中心极限定理这些工具都可以在此应用！）
- 另一方面，每一个样品在观测后只能有一个确定的数值，**样本此时是一组数值**，每一个样本点只能有一个取值

不难看出，从总体中随机抽取 n 个个体的抽样方法是很多的。但是我们希望抽出来的样本能够和我们预想的一样 => 尽可能地反映总体性质。因此我们对抽样方法抽出的样本做以下两点要求：

- 样本具有**代表性**，即每一样品（样本中的个体） X_i 与总体 X 具有相同的分布，这就要求总体中每一个个体都有同等机会被选入样本，从这我们也更能理解样本的随机变量本质。

\$\$

$$P(X_i \leq x) = \sum_{x_j \leq x} P(X = x_j) = P(X \leq x)$$

\$\$

- 样本具有**独立性**，即要求样本中每一样品的取值不影响其他样品的取值，也就是说 X_1, X_2, \dots, X_n 相互独立

我们一般使用**简单随机抽样**来满足上述两点要求，一般地，当总体所含个体特别多的情况下，上述两点要求都能得到满足。

现在举一个例子来具象化上述概念，假设一批产品共 N 个，需要进行抽样检验（因为产品太多了，不可能逐一检验）以了解其不合格率 p 。显然的，基于“上帝”设定的不合格率，对这一批产品的总体而言样本合格情况 X （合格为 1 不合格为 0）这一随机变量服从二值分布

\$\$

$$P(X = 1) = p, \quad P(X = 0) = 1 - p$$

\$\$

从总体中等概率有放回抽取 n 个样品构成样本 A ，在**检测之前**它们的合格情况 $\{X_i\}, i = 1, 2, \dots, n$ 是一个随机变量序列，且有

\$\$

$$P(X_i = 1) = P(X = 1), \quad P(X_i = 0) = P(X = 0)$$

\$\$

又因为 X_i 之间相互独立，那么对于样本 A 而言：

\$\$

$$P(A = 1) = \sum_{i=1}^n P(A = 1 | X_i = 1) = \sum_{i=1}^n P(X_i = 1 | A = 1) = P(X = 1)$$

\$\$

也就意味着样本 A 的分布在概率理论上是等于总体的分布的，但是检测后的实际情况则可能发生偏离（因为只是按分布收敛哈！）。如果考虑等概率不放回的情况，那么 X_i 之间相互不独立了，就会十分复杂，而且现实中也基本上都是不放回抽样，但是理论上可以证明 N 足够大的情况下，不放回抽样与放回抽样基本一致。

5.2 统计量及其分布

上面，样本的本质是一个含义完全相同的随机变量的序列 $\{X_i\}$ ，且任一 X_i 的分布都与总体相同，且样本的分布也和总体相同。因此样本中含有总体各个方面信息，这些信息反映在样本中各个样品的确切取值上。

在现实中，我们的思考方式是完全反过来，我们可以对样品的确切取值进行分析，透过现象看本质，透过样本看总体，找到真正的规律！在此过程中必要借助一些合理有效的工具，比如统计量和分布。

5.2.1 统计量与抽样分布

既然样本是一个独立同分布的随机变量序列，那我们就完全可以用之前所学的大数定律、中心极限定理等工具对其进行分析，更统一的工具是使用多维随机变量的函数实现变化特征降维，产生新的一维随机变量。所谓统计量就是采用一系列不同的、具有特殊含义的函数对样本实现降维提取特征后得到的对应的、有一定含义的一维随机变量【统计量仍然是随机变量，其本质是对随机变量序列的特征提取】。

【定义】 设 X_1, X_2, \dots, X_n 为取自某总体的样本，若样本函数 $T = T(X_1, X_2, \dots, X_n)$ 中不含有任何未知参数，则称随机变量 T 为统计量，统计量的分布称为抽样分布。

我们再次直观解释一下统计量仍然是随机变量的含义，也就是说对于一总体，我们按照具体抽样方式得到的样本其本身是一个随机变量序列，尽管每抽一次随机变量序列值以及其对应的统计量都有确定的取值，但这并不代表该值是固定的！再抽一次完全就可能是另外的值和统计量（这不就是随机变量的定义吗？很好理解！）因此若 X_1, X_2, \dots, X_n 为样本， $\sum_{i=1}^n X_i$ ， $\sum_{i=1}^n X_i^2$ 都是统计量。而当 μ, σ^2 未知时， $x_1 - \mu, x_1/\sigma$ 等均不是统计量！

下面我们就依次学习一些常见的统计量及其对应的抽样分布。

5.2.2 样本均值

【定义】 设 X_1, X_2, \dots, X_n 为取自某总体的样本，其算术平均值为样本均值，一般用 \bar{X} 表示，即

$$\begin{aligned} & \text{\$} \\ & \overline{X} = \frac{X_1 + \dots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i. \\ & \text{\$} \end{aligned}$$

可以明显发现 \bar{X} 是一个随机变量，下面讨论其分布特性的时候更为清楚。其对于一确定样本取值：

$$\begin{aligned} & \text{\$} \\ & \bar{x} = \frac{x_1 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i. \\ & \text{\$} \end{aligned}$$

【性质】

- 若把样本中的数据与样本均值之差称为偏差，则样本所有偏差之和为 0，即 $\sum_{i=1}^n (x_i - \bar{x}) = 0$ 。

- 数据观察值与均值的偏差平方和最小，即在形如 $\sum(x_i - c)^2$ 的函数中， $\sum(x_i - \bar{x})^2$ 最小，其中 c 为任意给定常数。

\$\$

```

\begin{aligned}
& \sum (x_i - c)^2 = \sum (x_i - |x| + |x - c|)^2 \\
& = \sum (x_i - |x|)^2 + n(|x - c|)^2 + \\
& 2 \sum (x_i - |x|)(|x - c|) \\
& = \sum (x_i - |x|)^2 + n(|x - c|)^2 \geq \sum (x_i - |x|)^2.
\end{aligned}
$$

```

布特性】

为了进一步充分地理解统计量为随机变量的含义，在此节选了一个例子（以后一旦犯迷糊，就回来看这个）

下面考察样本均值的分布.

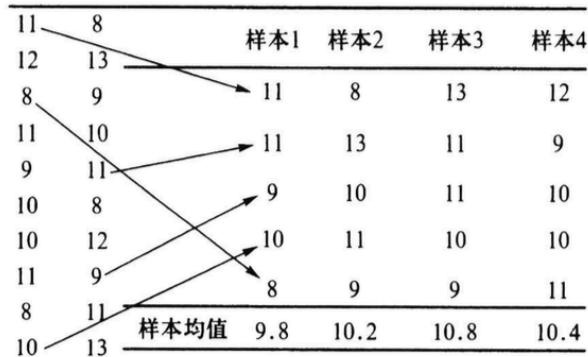


图 5.3.1 4 个样本的样本均值

例 5.3.2 设有一个由 20 个数组成的总体, 现从该总体同时取出容量为 5 的样本.

图 5.3.1 画出第一个样本的抽样过程,左侧是该总体,右侧是从总体中随机地抽出的样本,记录后,放回,再抽第二个样本.这里一共抽出 4 个样本,每个样本有 5 个观测值,我们计算了各个样本的样本均值.由抽样的随机性,每一个样本的样本均值都有差别.

设想类似抽取样本 5、样本 6……每次都计算样本均值 \bar{x} ，它们之间的差异是由于抽样的随机性引起的。假如无限制地抽下去，这样我们可以得到大量的 \bar{x} 的值，图 5.3.2 就是用这样得到的 500 个 \bar{x} 的值所形成的直方图，它反映了 \bar{x} 的抽样分布。

它的外形很像正态分布,这不是偶然的,有下面定理保证:

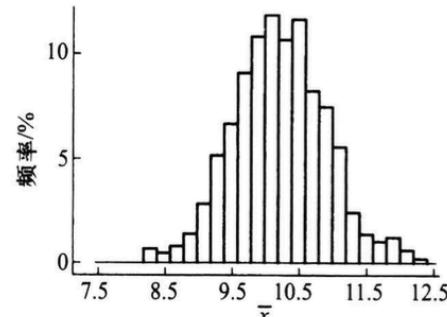


图 5.3.2 500 个样本均值形成的直方图

- 若总体 X 服从分布 $N(\mu, \sigma^2)$, 则 \bar{X} 的精确分布 (注意是精确分布, 不是渐进分布) 为 $N(\mu, \sigma^2/n)$

【证明】利用卷积公式，可得知 $\sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2)$ ，由此可知 $\bar{X} \sim N(\mu, \sigma^2/n)$

- 若总体分布未知或不是正态分布，但 $E(X) = \mu, Var(X) = \sigma^2$ ，则 n 较大时 \bar{X} 的渐近分布为 $N(\mu, \sigma^2/n)$ ，常记为 $\bar{X} \sim N(\mu, \sigma^2/n)$

【证明】由中心极限定理， $\sqrt{n}(\bar{X} - \mu)/\sigma \xrightarrow{L} N(0, 1)$ ，这表明 n 较大时 \bar{X} 的渐近分布为 $N(\mu, \sigma^2/n)$ （下面的图就能说明结论）

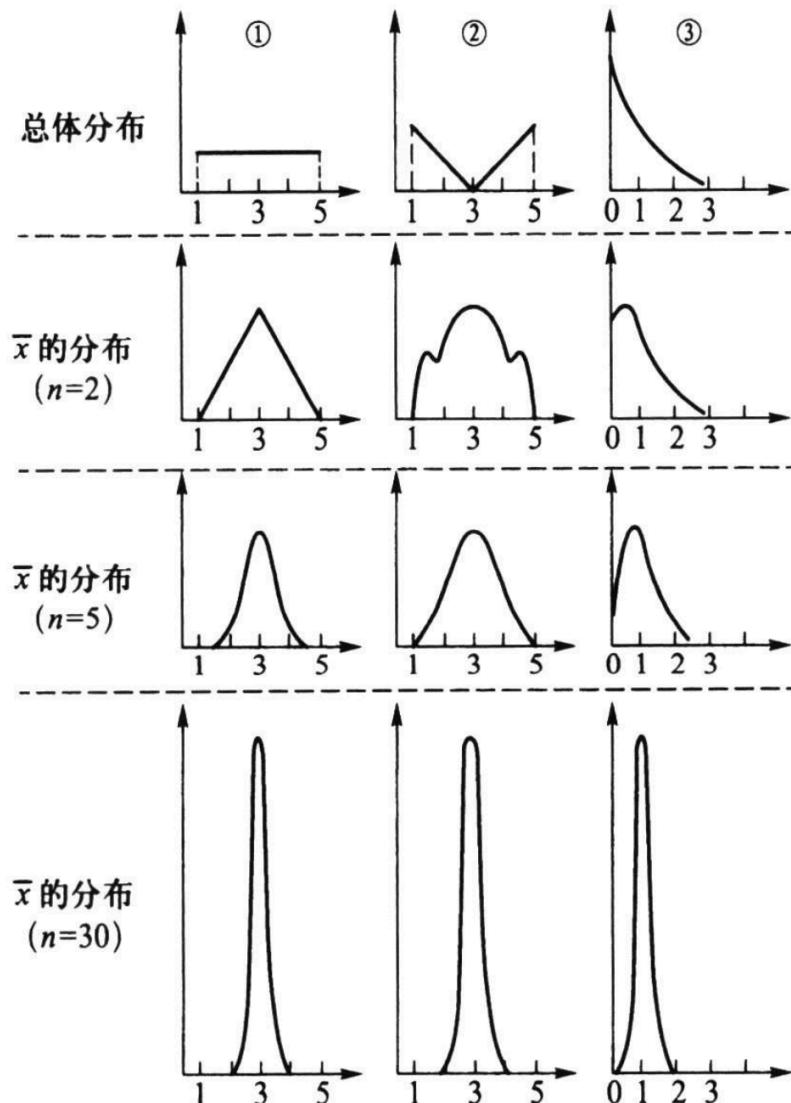


图 5.3.3 不同总体样本均值的分布

5.3.3 样本方差与样本标准差

【定义】 设 X_1, X_2, \dots, X_n 为取自某总体的样本，则它关于样本均值 \bar{X} 的平均偏差平方和

\$\$

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

\$\$

称为样本方差，其算术根 $s_n = \sqrt{s_n^2}$ 称为样本标准差，相对样本方差而言，样本标准差通常更有实际意义，因为它与样本均值具有相同的度量单位。更为精确地，使用如下统计量作为样本方差，其是对上式的有偏修正

\$\$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

\$\$

在实际中， s^2 比 s_n^2 更常用，在以后讲样本方差通常指的是 s^2 。一定要能够区分样本方差和总体方差，样本方差是个随机变量，总体方差是上帝设定的固定的数值。

在这里我们引入自由度的概念，我们已经深刻地理解到样本是一个随机变量序列，样本的均值是一个随机变量。显然样本的任一随机变量都不和样本均值独立，二者之间有着明确的相关关系（或者称之为约束关系），所以在同时用到二者进行计算的时候，就会受到一系列限制，随机的概念就会减弱，这就是自由度的来源，自由度衡量的是一个随机变量可以自由取值的程度。

$\sum_{i=1}^n (X_i - \bar{X})^2$ 的自由度为 $n - 1$ 而非 n ，其含义是在 \bar{X} 确定后， n 个偏差 $X_1 - \bar{X}, X_2 - \bar{X}, \dots, X_n - \bar{X}$ 中只有 $n - 1$ 个数据可以自由变动，而第 n 个则不能自由取值，因为存在限制条件： $\sum (X_i - \bar{X}) = 0$ 。

【性质】

- 实际拿到数据后，常见的三种计算形式：

\$\$

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{1}{n} (\sum x_i)^2$$

\$\$

- 方差的数学期望特性，对于总体 X 下的样本 X_1, X_2, \dots, X_n 有

\$\$

$$\begin{aligned} E(\sum (X_i - \bar{X})^2) &= E(\sum X_i^2 - n\bar{X}^2) = nVar(X) + nE(X)^2 - (Var(X) + nE(X)^2) \\ &= (n-1)Var(X) \end{aligned}$$

\$\$

5.3.4 样本原点矩、样本中心距及其函数

样本均值和样本方差的一般性推广为样本矩（思路和一维随机变量的特征数是类似的）

设 X_1, X_2, \dots, X_n 是样本，则统计量

\$\$

$$a_k = \frac{1}{n} \sum_{i=1}^n X_i^k$$

\$\$

称为样本 k 阶原点矩，特别地，样本一阶原点矩就是样本均值。统计量

\$\$

$$b_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$$

$$~~~\$$$

称为样本 k 阶中心矩，特别地，样本二阶中心矩就是样本方差。

样本原点原点矩和中心矩可以混合构成一些常见的描述分布形状的统计量

- 样本偏度：设 X_1, X_2, \dots, X_n 是样本，则样本偏度如下，其反映的是分位点偏离均值的程度

$$~~~\$$$

$$\beta_s = b_3/b_2^{3/2}$$

$$~~~\$$$

- 样本峰度：设 X_1, X_2, \dots, X_n 是样本，则样本峰度如下，其反映的是高点的尖锐程度

$$~~~\$$$

$$\beta_k = \frac{b_4}{b_2^2} - 3$$

$$~~~\$$$

5.3.5 五数概括与箱线图

在得到有序样本后，容易计算如下五个值：最小观测值 $x_{\min} = x_{(1)}$ ；最大观测值 $x_{\max} = x_{(n)}$ 。中位数 $m_{0.5}$ ，第一四分位数 $Q_1 = m_{0.25}$ 和第三四分位数 $Q_3 = m_{0.75}$ 。所谓五数概括就是指用这五个数：

$$~~~\$$$

$$[x_{\min}, Q_1, m_{0.5}, Q_3, x_{\max}]$$

$$~~~\$$$

来大致描述一批数据的轮廓。

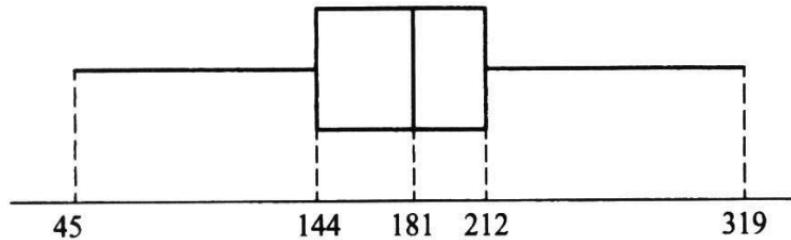


图 5.3.7 月销售量数据的箱线图

5.4 三大抽样分布

上一节我们了解了统计量，认识到统计量的分布称为抽样分布，而且深入了解了有关均值 \bar{X} 的抽样分布特性。这一节，我们在正态分布的假设下，以标准正态分布变量为基石构造三大著名的统计量，并探求其背后性质。之所以讨论这三个统计量，是因为：

- 它们都有这十分明确的含义
- 它们的密度函数都有显式表达式，这意味着可以通过积分求解概率或者分位点

- 它们在现实生活中有着无比广泛的应用（假设检验）

开门见山，直接列出三大随机变量及其对应的抽样分布：若设 X_1, X_2, \dots, X_n 和 Y_1, Y_2, \dots, Y_n 是来自标准正态分布的两个相互独立的样本，则此三个统计量的构造及其抽样分布如下表所示

统计量的构造	抽样分布密度函数	期望	方差
$\chi^2 = x_1^2 + x_2^2 + \dots + x_n^2$	$p(y) = \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{n/2}} y^{\frac{n}{2}-1} e^{-\frac{y}{2}} \quad (y>0)$	n	$2n$
$F = \frac{(y_1^2 + y_2^2 + \dots + y_m^2)/m}{(x_1^2 + x_2^2 + \dots + x_n^2)/n}$	$p(y) = \frac{\Gamma\left(\frac{m+n}{2}\right) \left(\frac{m}{n}\right)^{m/2}}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right)} y^{\frac{m}{2}-1} \cdot \left(1 + \frac{m}{n}y\right)^{-\frac{m+n}{2}} \quad (y>0)$	$\frac{n}{n-2} \quad (n>2)$	$\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \quad (n>4)$
$t = \frac{y_1}{\sqrt{(x_1^2 + x_2^2 + \dots + x_n^2)/n}}$	$p(y) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi} \Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{y^2}{n}\right)^{-\frac{n+1}{2}} \quad (-\infty < y < \infty)$	$0 \quad (n>1)$	$\frac{n}{n-2} \quad (n>2)$

对于这三个统计量，我们需要了解其密度函数分别是如何推导的以及相关性质，这些性质将成为日后构建统计量的靶子，每次想构建的时候都往这几个统计量上凑，因为只有构建出这些统计量才能用相关分布的性质。

5.4.1 χ^2 分布（卡方分布）【正态分布随机变量的统计量】

【定义】

设 X_1, X_2, \dots, X_n 独立同分布于标准正态分布 $N(0, 1)$ ，则 $\chi^2 = X_1^2 + \dots + X_n^2$ 的分布称为自由度为 n 的 χ^2 分布， $\chi^2 \sim \chi^2(n)$ 。

【推导概率密度函数】

若 $X \sim N(0, 1)$ ，则 $X^2 \sim Ga(1/2, 1/2)$ ，根据伽玛分布的可加性立有 $\chi^2 \sim Ga(n/2, n/2) = \chi^2(n)$ ，由此可见， $\chi^2(n)$ 分布是伽玛分布的特例。故 $\chi^2(n)$ 分布的密度函数，以及其示意图分别如下。卡方分布独立可加。

```
$$
p(y)=\frac{1}{\Gamma(n/2)}\frac{y^{n/2-1}}{2^{n/2}}e^{-y/2}, \quad y>0,
$$
```

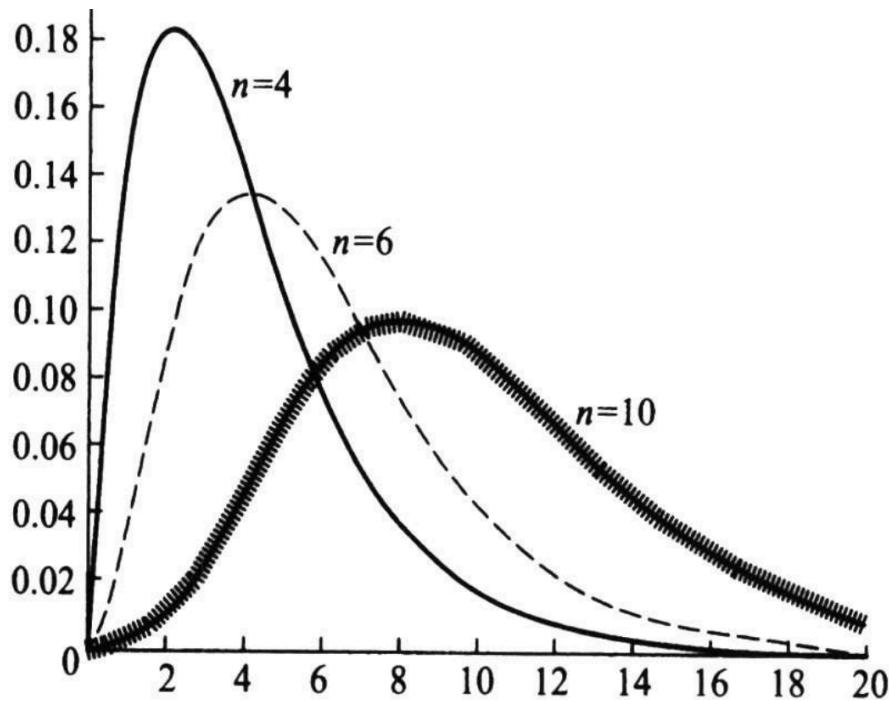


图 5.4.1 $\chi^2(n)$ 分布的密度函数

【与样本相关的重要性质】

设 X_1, X_2, \dots, X_n 是来自正态总体 $X \sim N(\mu, \sigma^2)$ 的样本，其样本均值和样本方差分别为

$$\begin{aligned} & \text{\$\$} \\ & \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{和} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \\ & \text{\$\$} \end{aligned}$$

则有 (以下三个定理的证明需要用到巧妙的矩阵运算，在此不过多说明。定理二当然是可以一眼看出来的，定理一和三则需要详细推导)

- 均值和方差相互独立： \bar{X} 与 s^2 相互独立
- 均值服从正态分布 $\bar{X} \sim N(\mu, \sigma^2/n)$
- 样本方差标准化后服从卡方分布

$$\begin{aligned} & \text{\$\$} \\ & \frac{(n-1)s^2}{\sigma^2} = \chi^2(n-1) \\ & \text{\$\$} \end{aligned}$$

5.4.2 F 分布 【卡方分布随机变量的统计量】

【定义】

设 $X_1 \sim \chi^2(m)$, $X_2 \sim \chi^2(n)$, X_1 与 X_2 独立, 则称 $F = \frac{X_1/m}{X_2/n}$ 的分布是自由度为 m 与 n 的 F 分布。记为 $F \sim F(m, n)$, 其中 m 称为分子自由度, n 称为分母自由度。可以看到, F 分布核心有三个要求:

- 两个卡方分布
- 卡方分布独立
- 卡方分布除以自由度后再相除

【推导概率密度函数】

此处的推导有些复杂, 不再赘述。观察一下 F 分布的图形即可

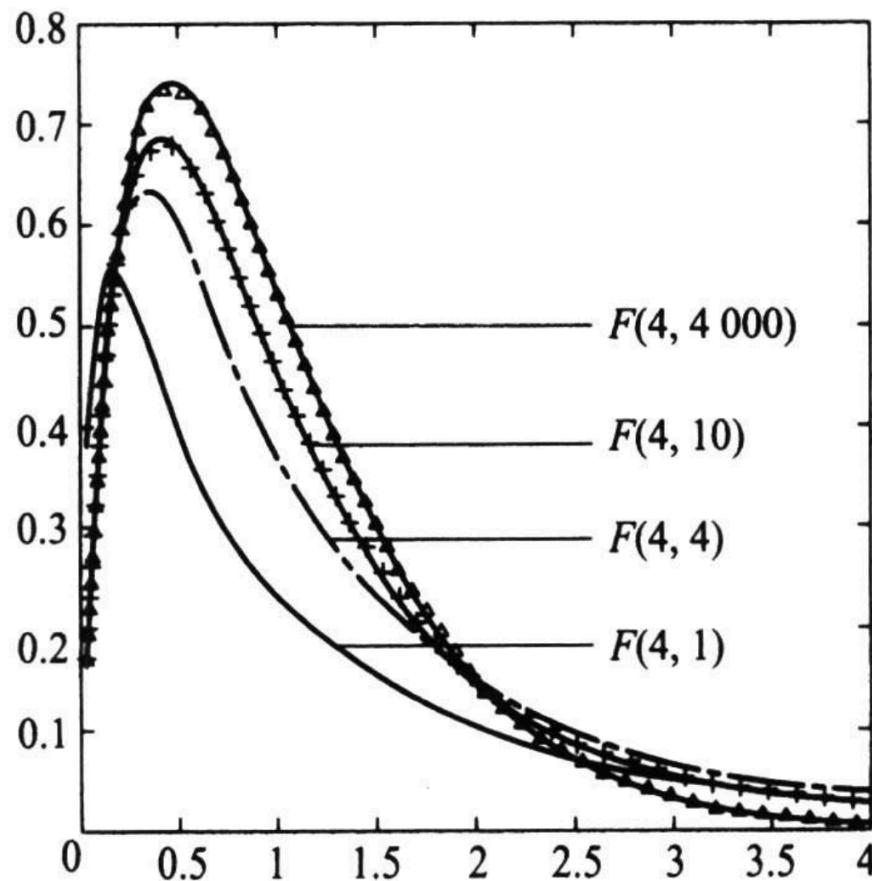


图 5.4.2 F 分布的密度函数

【与样本相关的重要性质】

设 X_1, X_2, \dots, X_m 是来自正态分布总体 $X \sim N(\mu_1, \sigma_1^2)$ 的样本, Y_1, Y_2, \dots, Y_n 是来自正态分布总体 $Y \sim N(\mu_2, \sigma_2^2)$ 的样本, 且此两样本相互独立, 记

\$\$ s_x^2 = \frac{1}{m-1} \sum_{i=1}^m (X_i - \bar{X})^2, \quad s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2, \\ \$\$

则有

\$\$ F = \frac{s_x^2 / \sigma_1^2}{s_y^2 / \sigma_2^2} \sim F(m-1, n-1) \\ \$\$

这条性质只需借助卡方分布的性质 3 即可直接证明。

5.4.3 student-t 分布【正态分布和卡方分布随机变量的统计变量】

【定义】

设随机变量 X_1 与 X_2 独立且 $X_1 \sim N(0, 1)$, $X_2 \sim \chi^2(n)$, 则称 $t = \frac{X_1}{\sqrt{X_2/n}}$ 的分布为自由度为 n 的 t 分布, 记为 $t \sim t(n)$ 。

【推导概率密度函数】

此处推导很巧妙, 利用卡方分布进行转化即可。在此给出 t 分布的性质。

t 分布的密度函数的图像是一个关于纵轴对称的分布(见图 5.4.3),与标准正态分布的密度函数形状类似,只是峰比标准正态分布低一些,尾部的概率比标准正态分布的大一些.

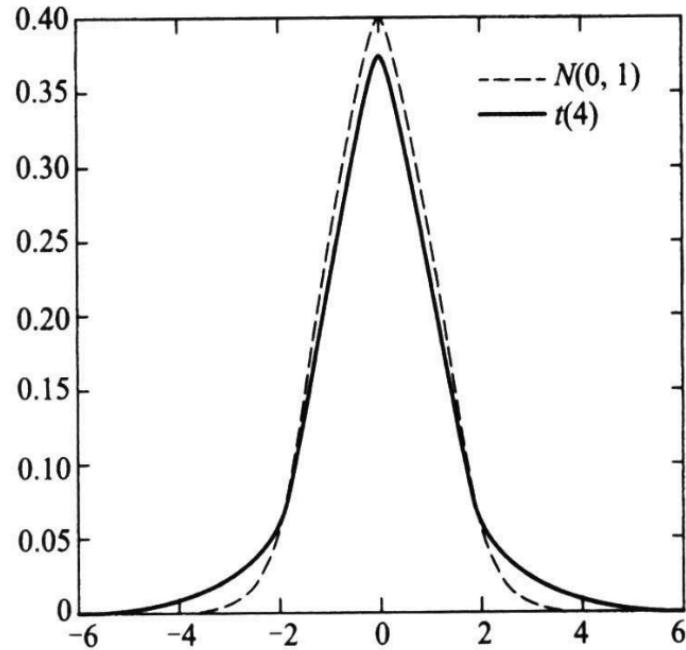


图 5.4.3 t 分布与 $N(0, 1)$ 的密度函数

注意到: 当自由度很低的时候, t 分布是肥尾的正态分布, 能够更好地刻画小样本分布情况。

- 自由度为 1 的 t 分布就是标准柯西分布, 它的均值不存在;
- $n > 1$ 时, t 分布的数学期望存在且始终为 0;
- $n > 2$ 时, t 分布的方差存在, 且为 $n/(n - 2)$;
- 当自由度较大 (如 $n \geq 30$ 时), t 分布可以用 $N(0, 1)$ 分布近似。

【与样本相关的重要性质】

- 设 X_1, X_2, \dots, X_m 是来自正态分布总体 $X \sim N(\mu, \sigma^2)$ 的样本则: (证明思路写在下面了)

\$\$

$$\begin{aligned} t &= \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim t(n-1), \\ &\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} = \frac{\frac{\bar{X} - \mu}{\sigma}}{\sqrt{\frac{1}{n}}} = \frac{\frac{\bar{X} - \mu}{\sigma}}{\sqrt{\frac{1}{n}}} \\ &\sim \sqrt{\frac{n-1}{n}} \cdot \frac{\bar{X} - \mu}{\sigma} \end{aligned}$$

\$\$

- 设 X_1, X_2, \dots, X_m 是来自正态分布总体 $X \sim N(\mu_1, \sigma_1^2)$ 的样本, Y_1, Y_2, \dots, Y_n 是来自正态分布总体 $Y \sim N(\mu_2, \sigma_2^2)$ 的样本, 且此两样本相互独立, 记

\$\$

$$s_x^2 = \frac{1}{m-1} \sum_{i=1}^m (\bar{x}_i - \bar{X})^2, \quad s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \bar{Y})^2, \\ s_w^2 = \frac{1}{m+n-2} \left\{ \sum_{i=1}^m \sum_{j=1}^n (\bar{x}_i - \bar{x}_j)^2 + \sum_{i=1}^n \sum_{j=1}^m (\bar{y}_i - \bar{y}_j)^2 \right\}$$

\$\$

则有：

\$\$

$$\frac{\left(\bar{x} - \bar{y} \right)^2 - \left(\mu_1 - \mu_2 \right)^2}{s_w^2} \sim \chi^2_{m+n-2}$$

\$\$

证明思路是同上的，无非就是对分子分母进行变换。

5.5 充分统计量【理解】

在刚接触到样本的统计量时我们就说，统计量是帮助我们理解总体的工具。更加准确地来说，构造统计量就是对样本进行加工，去粗取精，简化样本，便于统计推断发现总体规律。

但是我们不禁疑问，想要研究总体的某一特性 θ 事件，基于样本 $\{X_i\}, i = 1, 2, \dots, n$ 构建的统计量 T 真的能够充分反映出样本中中的有关总体特性 θ 的信息吗？

比如要研究某个运动员打靶命中率 θ 。一般地，设我们对该运动员进行 n 次观测，得到 X_1, X_2, \dots, X_n 每个 X_i 取值非 0 即 1，命中为 1，不命中为 0，令 $T = X_1 + X_2 + \dots + X_n$ 表示为观测到的命中次数，在研究命中率这一总体规律时，**仅仅记录使用 T 不会丢失任何与命中率 θ 有关的信息**，统计上将这种样本加工不损失信息称为充分性。相反如果设计一个统计量 $U = X_1 + X_2$ 其就不是充分的，因为它丢失了样本中与命中率有关的信息。

充分性的理论上的定义如下：设 X_1, X_2, \dots, X_n 是来自某个总体的样本，总体分布函数为 $F(x; \theta)$ ，统计量 $T = T(x_1, \dots, x_n)$ 称为 θ 的充分统计量，如果在给定 T 的取值后， X_1, X_2, \dots, X_n 的条件分布与 θ 无关。

在统计学中有一个基本原则：在充分统计量存在场合，**任何统计推断都可以基于充分统计量进行**，这可以简化统计推断的程序。通常将该原则称为**【充分性原则】**，这足以证明充分统计量的重要意义。然而直接从定义出发，证明统计量的充分性并不是那么容易的，往往需要借助**【因子分解定理】**：

引入一个在两种分布类型通用的概念—概率函数， $f(x)$ 称为随机变量 X 的概率函数：在连续型场合， $f(x)$ 表示 X 的概率密度函数；在离散型场合， $f(x)$ 表示 X 的概率分布列。设总体概率函数为 $f(x; \theta)$, X_1, \dots, X_n 为样本，则 $T = T(X_1, \dots, X_n)$ 为充分统计量的充分必要条件是：存在两个函数 $g(t, \theta)$ 和 $h(x_1, \dots, x_n)$ 使得对任意的 θ 和任一组观测值 x_1, \dots, x_n ，有

\$\$

$$f(x_1, \dots, x_n; \theta) = g(T(x_1, \dots, x_n), \theta) h(x_1, \dots, x_n)$$

\$\$

就像是完成了因子分解，把与 θ 有关的统计特征分解到一个函数中。

第六章 参数估计 <对总体的参数估计>

在上一章中，我们引入统计量的目的是明确的，就是对感兴趣的规律进行准确归纳进而对相关问题进行统计推断（统计量的本质是对随机变量序列特征的提取，也就是对整体分布某一维度特征的一种具像化表现），详细来说有两层含义：

- 如果已经知道了准确的规律，比如构建 t 统计量时就必须明确知道一个标准正态分布和卡方分布，我们可以基于这些准确的规律构建统计量进行下游任务的判断。
- 如果不知道准确的规律，我们则可以通过以下两种方式进行估计
 - 从样本出发，根据样本构造适当的统计量对总体的规律进行估计（参数估计）
 - 从总体出发，先对规律进行假设，然后将其用到样本中进行验证（假设检验）

正如我们之前所说的，规律就是总体的分布函数，其具体形式中有两大不确定的部分：

- 总体到底服从什么分布？正态分布还是 t 分布？
- 分布中所含的参数是多少？正态分布的均值、方差是多少？

这一章我们关注的点是从现实样本数据中直接估计出总体分布中的参数，正面给出参数的具体范围，破译分布函数的形式，找到总体的规律。先明确一下参数的广义概念，这里具体指的是以下三类未知参数。注意！！！参数一定是一个数，其有一个确定的值，而绝不是一个随机变量！从这也能看出：参数都是针对总体而言的，总体的参数就是规律的一种表现形式。

- 【纯正的未知参数】分布中所含有未知参数 θ 。如：两点分布 $b(1, p)$ 中的 p ；正态分布 $N(\mu, \sigma^2)$ 中的 μ 和 σ^2 ；
- 【参数的函数】分布中所含的未知参数 θ 的函数。如：服从正态分布 $N(\mu, \sigma^2)$ 的变量 X 不超过某给定值 a 的概率 $P(X \leq a) = \Phi\left(\frac{a-\mu}{\sigma}\right)$ 是未知参数 μ, σ 的函数；
- 【参数驱动的分布的特征数】分布的各种特征数也都是未知参数，如：均值 $E(X)$ ，方差 $\text{Var}(X)$ ，分布中位数等。这个地方千万不要搞混了，总体分布的特征数绝对是一个数，而不是一个随机变量，但样本分布的特征数则是一个随机变量。

常用 θ 表示参数，参数 θ 所有可能取值组成的集合称为参数空间，常用 Θ 表示。参数估计问题本质上就是自下而上的过程，是基于样本信息构造出适当的统计量对总体中的上述各种未知参数作出估计。参数估计的具体形式有两种：

- 点估计（直接估计出参数具体值）
 - 矩估计
 - 极大似然估计
 - 最小方差无偏估计
 - 贝叶斯估计
- 区间估计（估计出参数所在的区间）

6.1 点估计的概念与常见方法

6.1.1 点估计及其评判标准

a. 点估计

【点估计的定义】 设 X_1, X_2, \dots, X_n 是来自总体的一个样本，其样本值为，我们基于该样本计算统计量 $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ ，用 $\hat{\theta}$ 的取值作为对总体参数 θ 值的估计，把 $\hat{\theta}$ 称为 θ 的点估计（再次强调 $\hat{\theta}$ 是个随机变量，而 θ 是一个确切的值， θ 的定义域为 Θ ）。

此处“点”的含义就是基于总体中抽出的一个随机样本 x_1, x_2, \dots, x_n 计算出的统计量 $\hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n)$ ，用这个样本参数随机变量中的点，来估计总体的参数值。

在这里如何基于样本构造统计量 $\hat{\theta}$ 并没有明确的规定（毕竟只是做估计），只要它满足一定的合理性即可。合理性要求有很多，而最直接的一个要求就是：估计出来的 $\hat{\theta}$ 和实际值不能偏差太远，不能太远就有三层含义：

- 如果有机会把样本尽可能扩大（甚至扩大到总体），那么估计出来的 $\hat{\theta}$ 应该与实际的 θ 几乎一致（相和性）；
- 尽可能在实际值上下波动，不要系统性偏离（无偏性）；
- 波动的范围尽可能小（有效性）。

b. 相和性

【相和性的定义】 设 $\theta \in \Theta$ 为未知参数， $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \dots, X_n)$ 是 θ 的一个估计量， n 是样本容量，若对于任何一个 $\varepsilon > 0$ 都有

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

则称 $\hat{\theta}_n$ 为参数 θ 的相和估计。可以看出来相和估计是对估计的一个最基本要求，如果一个估计量，在样本量不断增大时，它都不能吧被估参数估计到任意指定京都，那么这个估计是很值得怀疑的。**因此如果一个估计量连相和性都不满足的话，根本就不用考虑它。**

b. 无偏性

【无偏性的定义】 设 $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ 是 θ 的一个估计， θ 的参数空间为 Θ ，若对任意的 $\theta \in \Theta$ ，有

$$\begin{aligned} &\text{E}(\hat{\theta}) = \theta \\ &\text{E}(\hat{\theta}) = \theta \end{aligned}$$

则称 $\hat{\theta}$ 是 θ 的无偏估计，否则称为有偏估计。可以看到，无偏性是相对于“上帝”而言的绝对概念。

无偏性所呈现的【数学本质】是估计值和实际值之间不存在系统偏差。具体而言：无偏性要求可以改写为 $E(\hat{\theta} - \theta) = 0$ ，这表示无偏估计没有系统偏差。当我们使用 $\hat{\theta}$ 估计 θ 时，由于样本的随机性，基于某一特定的样本估计出来的 $\hat{\theta}$ 与 θ 总是有偏差的，这种偏差时而（对某些样本观测值）为正，时而（对另一些样本观测值）为负，时而大，时而小。**无偏性表示，把这些偏差平均后值为 0**。而若估计不具有无偏性，则无论使用多少个样本估计多少次，其平均也会与参数真值有一定的距离，这个距离就是系统误差。

【无偏性的性质】 对于无偏性而言，有以下几点性质需要琢磨：

- 在样本上直接计算的样本方差明显是有偏的，但是随着样本量的增加就会趋近于总体参数值，称这种性质为**渐进无偏估计**。
- 无偏性不具有不变性。即若 $\hat{\theta}$ 是 θ 的无偏估计，一般而言， $g(\hat{\theta})$ 不是 $g(\theta)$ 的无偏估计，除非 $g(\theta)$ 是 θ 的线性函数。譬如， s^2 是 σ^2 的无偏估计，但 s 不是 σ 的无偏估计。
- **并不是所有的参数都存在无偏估计**，当参数存在无偏估计时，称该参数是可估的，否则称它是不可估的。

c. 有效性

【有效性的定义】 设 $\hat{\theta}_1, \hat{\theta}_2$ 是 θ 的两个无偏估计，如果对任意的 $\theta \in \Theta$ 有

$$\begin{aligned} & \text{Var}(\hat{\theta}_1) \leq \text{Var}(\hat{\theta}_2) \\ & \text{且至少有一个 } \theta \in \Theta \text{ 使得上述不等号严格成立, 则称 } \hat{\theta}_1 \text{ 比 } \hat{\theta}_2 \text{ 有效} \end{aligned}$$

且至少有一个 $\theta \in \Theta$ 使得上述不等号严格成立，则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效（注意这个地方的表述）。可以看到，有效性是对于无偏估计量而言的相对概念，只有无偏估计量才能比较有效性，如果一个估计量连无偏性都没有，那就更不用提有效性了。但是请注意，我们提到有效估计时一般指方差最小的无偏估计量，是一个绝对概念。

有效性所呈现的【数学本质】是估计值和实际值之间偏差的波动尽可能小。参数的无偏估计可以有很多，如何在无偏估计中进行选择？直观的想法是希望该估计围绕参数真值的波动越小越好，波动大小可以用方差来衡量，因此人们常用无偏估计的方差的大小作为度量无偏估计优劣的标准，这就是有效性。

无偏性和有效性的计算与比较是常见的笔面试题目。归根到底是对多维随机变量函数的分布的求解：因为根据样本得到的估计值 $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ 本质是多维随机变量 $\{X_1, X_2, \dots, X_n\}$ 经过函数变换后得到的一个一维随机变量，求解 $E(\theta)$ 和 $Var(\theta)$ 直接套用之前的理论工具即可。

d. 最小方差无偏估计

上面已经提到了多种点估计量评价的标准，最小方差无偏估计这一部分对上述做了一个汇总。这一部分的难度较高，可以先不着急掌握。

6.1.2 矩估计【点估计 way 1】

【矩估计的定义】 矩估计是最简单直接的点估计方法，其核心是替换原理（就是用样本的数字特征去替换总体的数字特征）：

- 用样本矩去替换总体矩，这里的矩可以是原点矩也可以是中心矩；
- 用样本矩的函数去替换相应的总体矩的函数。

根据这个替换原理，在总体分布形式未知场合也可对各种参数作出估计，譬如：

- 用样本均值 \bar{X} 这一随机变量在的样本上的具体值 \bar{x} 估计总体均值 $E(X)$ ，即 $\hat{E}(X) = \bar{x}$ ；（一次原点矩）
- 用样本方差 s_n^2 这一随机变量在样本上的具体值，估计总体方差 $\text{Var}(x)$ ，即 $\hat{\text{Var}}(x) = s_n^2$ ；（二次中心矩）
- 用事件 A 出现的频率估计事件 A 发生的概率。（核心仍然是均值的概念）

【矩估计的具体形式】 设总体具有已知的概率函数 $p(x; \theta_1, \dots, \theta_k)$, $(\theta_1, \dots, \theta_k) \in \Theta$ 为未知参数或参数向量。

X_1, X_2, \dots, X_n 是样本，假定总体的 k 阶原点矩 μ_k 存在，则对所有的 j , $0 < j < k$, μ_j 都存在。若假设 $\theta_1, \dots, \theta_k$ 能够表示成 u_1, \dots, u_k 的函数 $\theta_j = \theta_j(u_1, \dots, u_k)$ ，则可给出诸 θ_j 的矩法估计：

```
$$\hat{\theta}_j = \theta_j(a_1, \dots, a_k), \quad j=1, \dots, k,$$
```

其中 a_1, \dots, a_k 是样本原点矩的和： $a_j = \frac{1}{n} \sum_{i=1}^n X_i^j$ 。从经验上来说当 $k = 1$ 时，也就是说存在 1 个未知参数，我们通常可以由样本均值（一阶原点矩）出发对未知参数进行估计；如果 $k = 2$ ，也就是说存在 2 个未知参数，可以由样本均值（一阶原点矩阵）和样本方差（二阶中心矩）出发估计未知参数。

【矩估计的经典例子】 X_1, X_2, \dots, X_n 是来自 (a, b) 上的均匀分布 $U(a, b)$ 的样本， a 与 b 均是未知参数，这里 $k = 2$ ，也就是说存在 2 个未知参数，由于

```
$$E(X) = \frac{a+b}{2}, \quad \text{Var}(X) = \frac{(b-a)^2}{12},$$
```

不难推出

```
$$a = E(X) - \sqrt{3\text{Var}(X)}, \quad b = E(X) + \sqrt{3\text{Var}(X)},
```

由此即可得到 a, b 的矩估计：

```
$$\hat{a} = \bar{x} - \sqrt{3}s, \quad \hat{b} = \bar{x} + \sqrt{3}s,
```

6.1.3 极大似然估计【点估计 way 2】

a. 极大似然估计

极大似然估计是真实估计中最常用的一种参数估计方法，其核心思想很简单直接：既然存在一个看不见的规律，其操纵着整个世界有序运行，那么我们就有理由相信，现实中所看到的样本中样本点的组合应该是这一规律下最有可能发生的（发生的概率最大）。

【极大似然估计的定义】 设总体 X 的概率密度函数为 $p(X; \theta)$, $\theta \in \Theta$, 其中 θ 是一个未知参数或几个未知参数组成的参数向量, Θ 是参数 θ 可能取值的参数空间。 X_1, X_2, \dots, X_n 是来自该总体的样本, 样本值为 x_1, x_2, \dots, x_n , 将样本的联合概率函数看成 θ 的函数, 用 $L(\theta; x_1, x_2, \dots, x_n)$ 表示, 简记为 $L(\theta)$, 因为样品之间独立同分布, 所以

$$\begin{aligned} & \\ L(\theta) &= L(\theta; x_1, x_2, \dots, x_n) = p(X = x_1; \theta) \cdot p(X = x_2; \theta) \cdots p(X = x_n; \theta), \end{aligned}$$

$L(\theta)$ 称为样本的似然函数, 如果某统计量 $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ 满足

$$\begin{aligned} & \\ L(\hat{\theta}) &= \max_{\theta \in \Theta} L(\theta) \end{aligned}$$

则称 $\hat{\theta}$ 是 θ 的极大似然估计, 简记为 **MLE (Maximum Likelihood Estimate)**。所以极大似然估计的整体思路十分简单：

- **step 1.** 得到总体 X 的概率密度函数
- **step 2.** 基于上述密度函数, 算出在特定样本下的概率似然函数 $L(\theta)$ (一般都是变换为 $\ln L(\theta)$ 连乘变连加)
- **step 3.** 解出令 $L(\theta)$ 取最大值的参数 θ (本质是一个规划问题, 常见的方式就是求导和 EM)

极大似然估计的优势在于其估计出的参数通常具有渐进正态性。

【极大似然估计的典例】 对正态总体 $N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$ 是二维参数, 设有样本 X_1, X_2, \dots, X_n , 样本值为 x_1, x_2, \dots, x_n , 则似然函数及其对数分别为

$$\begin{aligned} & \\ \ln L(\mu, \sigma^2) &= -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 + \frac{n}{2} \ln \sigma^2 - \frac{n}{2} \ln (2\pi) \end{aligned}$$

将 $\ln L(\mu, \sigma^2)$ 分别关于两个分量求偏导并令其为 0 即得到似然方程组

\$\$

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0 \\ \frac{\partial \ln L(\mu, \sigma^2)}{\partial \sigma^2} = \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2\sigma^2} = 0$$

\$\$

解此方程组，可得 μ 和 σ^2 的极大似然估计为

\$\$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = s^2$$

\$\$

利用二阶导函数矩阵的非正定性可以说明上述估计使得似然函数取极大值。

但是需要注意！尽管求导函数是求最大似然估计最常用的方法，但并不是在所有场合求导都是有用的。一个经典例子就是均匀分布的极大似然估计结果：设 x_1, \dots, x_n 是一个来自均匀总体 $U(0, \theta)$ 样本的样本值，试求 θ 的最大似然估计。当然还有更为复杂的情况，我们可能就要借助 EM 算法求解了。

【极大似然估计的性质】 最大似然估计有一个简单而有用的性质：如果是 θ 的最大似然估计，则对任一函数 $g(\theta)$ ，其最大似然估计为 $g(\hat{\theta})$ 。该性质称为最大似然估计的不变性，从而使一些复杂结构的参数的最大似然估计的获得变得容易了。这也为我们后续的区间估计打下基础！

b. EM 算法

求导只能解决一部分的规划问题，很多情况下想要根据求导得到参数的结果是较为困难的。Dempster 等人于 1977 年提出了 EM 算法，其出发点就是为了解决难以通过求导得到结果的 MLE 问题。其具体步骤分为三步：

- 引入潜变量（latent variable）将极大似然函数尽可能简化
- E 步：求期望。基于当前步骤下得到的参数以及样本观测值求解期望，以消除潜变量
- M 步：找最大。将上述消除潜变量的式子对参数求导，完成迭代更新

E 和 M 两步不断循环，迭代求得最重的收敛值。

6.1.4 贝叶斯估计 【点估计 way 3】

在统计学中有两个大的学派：频率学派（也称经典学派）和贝叶斯学派。在之前的学习中包括对概率的学习，我们基本上都是从容易理解的频率学派出发的。当然我们在对贝叶斯公式进行详细说明时已经通过一些例子了解了贝叶斯的思想，世界上没有任何两件事是完全独立的，一些已经积累的经验和已经发生的事情，都可以帮助认清新的规律，换句话说之前发生的事情对当下研究的内容有着一定的影响。本节将从贝叶斯的角度上深入探讨这一学派的核心思考方式。

a. 统计推断的基础究竟是什么？

在前面已经讲过，统计推断是根据样本信息对总体分布或总体的特征数进行推断，发现客观规律，也就是说想要获得总体的特征，我们只需要深入研究在该总体下抽取样本所包含的信息即可，不用 care 其他的信息。事实上，这是经典学派对统计推断的规定，这里的统计推断基于两种信息：总体信息和样本信息；而贝叶斯学派认为，除了上述两种信息以外，统计推断还应该基于第三种信息：先验信息。

- **总体信息**：总体信息即总体分布或总体所属分布族提供的信息（大家研究的总体可能比较类似，因此前人对总体的研究结果就可以作为我们所使用的信息），上面的两节在介绍矩估计和极大似然估计时，我们就用到了大量的总体信息。
 - 譬如，若已知“总体是正态分布”，则我们就知道很多信息
 - 譬如，总体的一切阶矩都存在；总体密度函数关于均值对称；总体的所有性质由其一、二阶矩决定；有许多成熟的统计推断方法可供我们选用等
- **样本信息**：样本信息即抽取样本所得观测值提供的信息（就是我们抽样后实实在在看到的信息）。譬如，在有了样本观测值后，我们可以根据它大概知道总体的一些特征数，如总体均值、总体方差等等在一个什么范围内。这是最“新鲜”的信息，并且越多越好，希望通过样本对总体分布或总体的某些特征作出较精确的统计推断。没有样本就没有统计学可言。
- **先验信息**：如果我们把抽取样本看作做一次试验，则样本信息就是试验中得到的信息。实际中，人们在试验之前对要做的问题在经验上和资料上总是有所了解的，而且这些信息对统计推断是有益的。先验信息即是抽样（试验）之前有关统计问题的一些信息。一般来说，**先验信息来源于经验和历史资料**。可以看到，总体信息也是先验信息的一种，但仅限于对总体的信息描述。先验信息则更加全面包含更加具体的特征信息。

为了更加清晰地认识先验信息的含义，我们在这举一个例子：在某工厂的产品中每天要抽检 n 件以确定该厂产品的质量是否满足要求。产品质量要用每天该厂产品的总体不合格品率 p 来度量，我们可以用 n 件抽查产品中的不合格品件数 X 来估计不合格品率。由于生产过程有连续性，可以认为**每天的产品质量是有关联的**（这是很自然的假设），即是说，在估计现在的 p 时，以前所积累的资料应该是可供使用的，这些积累的历史资料就是先验信息，从这可以看出我们在估计今天该工厂的产品的总体不合格率 p 时，用到了三个角度的信息：

- 总体信息：总体应该服从二项分布 (n, p)
- 样本信息：今天的 n 件抽查产品中的不合格件数
- 先验信息：过去几天抽查的产品不合格数

为了能使用这些先验信息，需要对它进行加工，譬如，在经过一段时间后，就可根据历史资料对过去 n 件产品中的不合格品件数 X 构造一个分布

\$\$

$$P(X = i) = \prod_{j=1}^n p_j^{x_j} (1-p_j)^{n-x_j}$$

\$\$

这种先验分布是对该厂过去产品的不合格品率的一个全面看法，其可以为我们今天的估计提供比较准确的信息。

b. 贝叶斯统计的含义与其密度函数形式

贝叶斯学派的基本观点是：**任一未知量 θ 都可看作随机变量，可用一个概率分布去描述，这个分布称为先验分布，这是理解贝叶斯估计的核心所在。**也就是说，“上帝”设定的规律本质上就是一个随机变量，而不想我们之前想的那样是一个固定的值。这也可以很自然地理解，可以把“上帝”正向制造规律的过程，想象成它在调节巨大的世界机器，其中每一个零件都是随机变量，有自己的分布函数，由其组合出来的各种规律自然也是随机变量，受每个零件的状态条件而改变。还是那句话世界上没有任何两件事情是绝对独立的，一个规律的产生背后有千千万万个条件，全概率公式极为复杂。所谓先验分布，就是挖掘那些已经被破译的与该规律相关的信息所得到的该规律的分布情况。

永远记住，规律是随机变量一直在变化，甚至分布也发生了改变，所以只依靠先验分布是极为粗糙的，甚至是错误的，要想挖掘出该规律，就是不断地去看世界现在的样子，用新的知识不断修正过去认知。所以在获得样本之后，总体分布、样本与先验分布通过贝叶斯公式结合起来得到一个关于未知量 θ 新的分布——后验分布；任何关于 θ 的统计推断都应该基于 θ 的后验分布进行，原因很简单，就像刚才说的那样，后验分布才是对当下规律最好的估计。

【贝叶斯公式的密度函数形式】

- 总体依赖于参数 θ 的概率函数在经典统计中记为 $p(x; \theta)$ ，它表示参数空间 Θ 中不同的 θ 对应不同的分布，也就意味着对于同一个总体而言规律是固定死的， θ 是确定的。而在贝叶斯统计中应记为 $p(x|\theta)$ ，它表示在随机变量 θ 取某个给定值时总体的条件概率函数，也就是说规律时可能发生变化的， θ 是服从某一分布的随机取值的。
- 根据参数 θ 的先验信息确定先验分布 $\pi(\theta)$ 。也就是说，可以基于过往的参数情况对参数所属的分布做一个大体判断
- 从贝叶斯观点看，样本 $\mathbf{X} = (x_1, \dots, x_n)$ 的产生要分两步进行：首先设想从先验分布 $\pi(\theta)$ 产生一个样本 θ_0 ，这一步是“上帝”做的，或者说是由世界交错纵横的条件相互影响产生的。第二步从 $p(\mathbf{X}|\theta_0)$ 中产生一组样本，这时在给定 θ_0 的条件下，样本 $\mathbf{X} = (x_1, \dots, x_n)$ 的联合条件概率函数为：

\$\$

$$p(\boldsymbol{X} | \theta) = p(x_1, \dots, x_n | \theta) = \prod_{i=1}^n p(x_i | \theta)$$

\$\$

这个分布综合了总体信息和样本信息。

- 由于 θ_0 是设想出来的，仍然是未知的，它是按先验分布 $\pi(\theta)$ 产生的。为把先验信息综合进去，不能只考虑 θ_0 这一种情况，对 θ 的其他值发生的可能性也要加以考虑，故要用 $\pi(\theta)$ 进行综合。这样一来，样本 \mathbf{X} 和参数 θ 的联合分布为（总体也服从该联合分布）

\$\$

$$h(\boldsymbol{X}, \theta) = p(\boldsymbol{X} | \theta) \pi(\theta)$$

\$\$

这个联合分布把总体信息、样本信息和先验信息三种可用信息都综合进去了。

- 我们的目的是要对总体的未知参数 θ 作统计推断。在没有样本信息时，我们只能依据先验分布对 θ 作出推断（假如我们在实验前对整个事件没有任何了解，因此对 θ 没有任何先验的信息，在这种场合下，贝叶斯本人建议采用“同等无知”的原则**使用区间 $(0, 1)$ 上的均匀分布作为 θ 的先验分布，也就是进行贝叶斯假设**）。在有了样本观察值 $\mathbf{X} = (x_1, \dots, x_n)$ 之后，我们应依据 $h(\mathbf{X}, \theta)$ 对 θ 作出推断。若把 $h(\mathbf{X}, \theta)$ 作如下分解：

\$\$

$h(\boldsymbol{X}, \theta) = \pi(\theta | \boldsymbol{X}) m(\boldsymbol{X})$
\$\$

其中 $m(\boldsymbol{X})$ 是 \boldsymbol{X} 的边际概率函数：

\$\$

$m(\boldsymbol{X}) = \int_{\theta} h(\boldsymbol{X}, \theta) d\theta$
 $\theta = \int_{\boldsymbol{X}} p(\boldsymbol{X} | \theta) \pi(\theta) d\theta$
\$\$

它与 θ 无关，或者说 $m(\boldsymbol{X})$ 中不含 θ 的任何信息。因此能用来对 θ 作出推断的仅是条件分布 $\pi(\theta | \boldsymbol{X})$ ，它的计算公式是

\$\$

$\pi(\theta | \boldsymbol{X}) = \frac{h(\boldsymbol{X}, \theta)}{m(\boldsymbol{X})} = \frac{\pi(\theta) p(\boldsymbol{X} | \theta)}{\int_{\theta} p(\boldsymbol{X} | \theta) \pi(\theta) d\theta}$
\$\$

这个条件分布称为 θ 的后验分布，它集中了总体、样本和先验中有关 θ 的一切信息。其就是用密度函数表示的贝叶斯公式，也是用样本数据对先验分布 $\pi(\theta)$ 作调整的结果，它要比 $\pi(\theta)$ 更接近 θ 的实际情况！（这就是后验分布的含义所在）。

【贝叶斯估计的典型例子】

设某事件 A 在一次试验中发生的概率为 θ 。为了估计 θ ，对试验进行了 n 次独立观测，其中事件 A 发生了 X 次，显然根据**总体信息**我们可以得到： $X | \theta \sim b(n, \theta)$ 。因为我们之前对事件 A 没有任何的了解，因此没有任何可靠的先验信息，所以只能对 θ 做出的**先验假设**为： $\theta \sim U(0, 1)$ 。还在实验中观测到事件 A 发生了 x 次这一**样本信息**。我们开始利用贝叶斯公式求出 θ 在这三重信息下的贝叶斯估计。

- **Step 1:** 先写出 X 和 θ 的联合分布

\$\$

$h(x, \theta) = C_n x^\theta (1-\theta)^{n-x}$, $x=0, 1, \dots, n$, $0 < \theta < 1$
\$\$

- **Step 2:** 然后求 X 的边际分布

\$\$

$m(x) = C_n x^\theta \int_0^1 (1-\theta)^{n-x} d\theta = C_n x^\theta \frac{\Gamma(n+1)}{\Gamma(n+2)}$
\$\$

- **Step 3:** 根据贝叶斯公式求解出参数 θ 的后验分布

\$\$

$$\begin{aligned} \pi(\theta | x) &= \frac{h(x, \theta)}{m(x)} \\ &\propto \theta^x (1-\theta)^{n-x} \end{aligned}$$

\$\$

- **Step 4:** 求解期望得到最终结果, 最后的结果说明 $\theta|x \sim Be(x + 1, n - x + 1)$, 注意, 这个以观测值为参数的后验分布本质上才是对参数(随机变量)的最准确估计, 因为我们估计的出发点就是该参数服从一个分布而非具体数值, 但是不可否认的是, 这个分布并不是那么直观的。所以, 我们利用后验分布的期望给出估计的最终结果。其后验期望估计为:
(期望后仍然是一个随机变量)

$$\begin{aligned} & \$ \\ & \hat{\theta}_{\text{B}} = E(\theta | x) = \frac{x+1}{n+2} \\ & \$ \end{aligned}$$

与之对比, 如果不用先验信息, 只用总体信息与样本信息构造矩估计或者最大似然估计, 那么事件 A 发生的概率的估计值如下, 注意该估计的结果也是一个随机变量, 但是这个随机变量的随机性仅来自于“抽样”:

$$\begin{aligned} & \$ \\ & \hat{\theta}_{\text{M}} = \frac{x}{n} \\ & \$ \end{aligned}$$

二者在 n 比较小时存在比较大的差异, 性质也有所不同。

6.2 区间估计

6.2.1 区间估计的概念与估计方法

a. 区间估计的概念

首先, 我们再次回到点估计的出发点——希望根据样本表现给出一个“尽可能”准确的参数估计值, 我们可以用无偏性等性质来说明这种准确性, 但是无法通过该随机变量的具体表现值, 说明其估计精度到底如何。因此就想构造一种更加直观的估计, 既能描述出想要估计参数的具体值是多少, 还能描述估计的精度如何。这就是区间估计的出发点。

【区间估计的定义】 设 θ 是总体的一个参数, 其参数空间为 Θ , x_1, \dots, x_n 是来自该总体的样本, 对给定的一个 $\alpha (0 < \alpha < 1)$, 若有两个统计量 $\hat{\theta}_L = \hat{\theta}_L(x_1, \dots, x_n)$ 和 $\hat{\theta}_U = \hat{\theta}_U(x_1, \dots, x_n)$, 若对任意的 $\theta \in \Theta$, 都有

$$\begin{aligned} & \$ \\ P_{\{\theta\}} \left(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U \right) & \geq 1 - \alpha \\ & \$ \end{aligned}$$

则称随机区间 $[\hat{\theta}_L, \hat{\theta}_U]$ 为 θ 的置信水平为 $1 - \alpha$ 的置信区间, 或简称 $[\hat{\theta}_L, \hat{\theta}_U]$ 是 θ 的 $1 - \alpha$ 置信区间, $\hat{\theta}_L$ 和 $\hat{\theta}_U$ 分别称为 θ 的(双侧)置信下限和置信上限。可以从概率的角度这么理解此概念: 对于上帝创造出来的参数 θ 而言, 根据此参数下的某一样本所得出的却见估计能够涵盖该参数的概率为 $1 - \alpha$ 。

【区间估计的理解】 我们也可以从频率的概念出发展开理解: 在计算大量重复使用 θ 的置信区间 $[\hat{\theta}_L, \hat{\theta}_U]$ 时, 尽管总体的参数并没有改变, 但是每次得到的样本观测值是不同的, 从而每次得到的区间估计值也是不一样的。而对一次具体的观测值而言, θ 可能在 $[\hat{\theta}_L, \hat{\theta}_U]$ 内, 也可能不在。平均而言, 在这大量的区间估计观测值中, 至少有 $100(1 - \alpha)\%$ 包含 θ , 如下图所示, 100 次估计得到 100 个置信区间, 上帝构造的参数值 15 在该置信区间中的概率大概就是 90%。注意这种理解和单样本理解有着明显的区别。

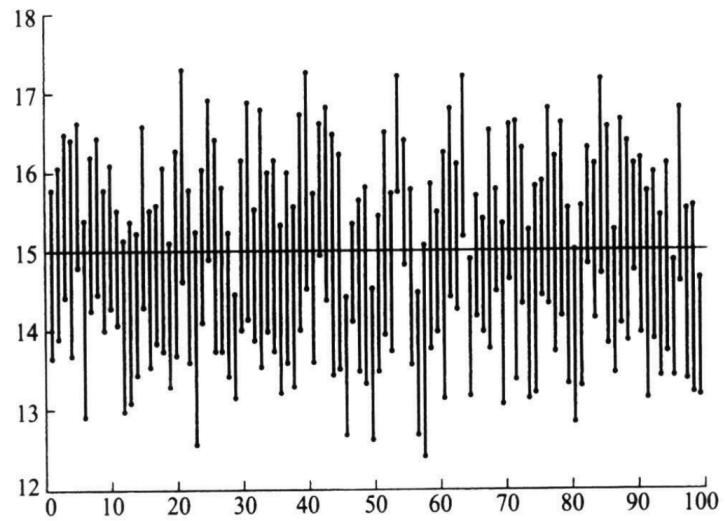


图 6.6.1 μ 的置信水平为 0.90 的置信区间

a. 枢轴量法进行区间估计

构造未知参数 θ 的置信区间的最常用的方法是枢轴量法，其步骤可以概括为如下三步：

- 设法构造一个样本和 θ 的函数 $G = G(x_1, \dots, x_n, \theta)$ 使得 G 的分布不依赖于未知参数，一般称具有这种性质的 G 为枢轴量。
- 适当地选择两个常数 c, d ，使对给定的 $\alpha (0 < \alpha < 1)$ ，有

\$\$

$$P(c \leq G \leq d) = 1 - \alpha$$

\$\$

- 假如能将 $c \leq G \leq d$ 进行不等式等价变形化为 $\hat{\theta}_L \leq \theta \leq \hat{\theta}_U$ ，则有

\$\$

$$P(\theta \in [\hat{\theta}_L, \hat{\theta}_U]) = 1 - \alpha$$

\$\$

这表明 $[\hat{\theta}_L, \hat{\theta}_U]$ 是 θ 的 $1 - \alpha$ 同等置信区间。

因此说白了，该方法的关键是要构造一个与参数无关的枢轴量 G ，一般是通过点估计来寻找。确定该枢轴量后，就可以对其进行范围规定，然后反变化为对原参数的范围规定，得到区间估计值。

6.2.2 单个正态总体参数的区间估计

对于正态总体 $N(\mu, \sigma^2)$ 的两个参数而言，进行区间估计整体来说分为以下三种情况，重点都在与枢轴量的构造。

a. 总体的方差 σ 已知, 进行 μ 的区间估计

这种情况下构造的枢轴量为:

$$\begin{aligned} & \text{\$\$} \\ & G = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1) \\ & \text{\$\$} \end{aligned}$$

得到 μ 的区间估计结果为:

$$\begin{aligned} & \text{\$\$} \\ & \left[\overline{x} - u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \overline{x} + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right] \\ & \text{\$\$} \end{aligned}$$

b. 总体的方差 σ 未知, 进行 μ 的区间估计

这种情况下构造的枢轴量为:

$$\begin{aligned} & \text{\$\$} \\ & t = \frac{\sqrt{n}(\overline{x} - \mu)}{s} \sim t(n-1) \\ & \text{\$\$} \end{aligned}$$

得到 μ 的区间估计结果为:

$$\begin{aligned} & \text{\$\$} \\ & \left[\overline{x} - t_{1-\alpha/2}(n-1) s / \sqrt{n}, \overline{x} + t_{1-\alpha/2}(n-1) s / \sqrt{n} \right] \\ & \text{\$\$} \end{aligned}$$

c. 总体的均值 μ 未知, 进行 σ^2 的区间估计

均值已知的情况可以参考 a. 但是现实中知道均值的难度远大于知道方差!

这种情况下构造的枢轴量为:

$$\begin{aligned} & \text{\$\$} \\ & G = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \\ & \text{\$\$} \end{aligned}$$

因此得到 σ^2 的区间估计结果为:

$$\begin{aligned} & \text{\$\$} \\ & \left[(n-1)s^2 / \chi_{1-\alpha/2}^2(n-1), (n-1)s^2 / \chi_{\alpha/2}^2(n-1) \right] \\ & \text{\$\$} \end{aligned}$$

6.2.3 两个正态总体下的区间估计

设 x_1, \dots, x_m 是来自 $N(\mu_1, \sigma_1^2)$ 的样本, y_1, \dots, y_n 是来自 $N(\mu_2, \sigma_2^2)$ 的样本, 且两个样本相互独立。 \bar{x} 与 \bar{y} 分别是它们的样本均值, $s_x^2 = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2$ 和 $s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$ 分别是它们的样本方差。此处讨论两个均值差和两个方差比的区间估计。

注意, 这一部分的内容在现实生活中有着极为重要的应用, 比如判断两个 model 的 performance 是否有着明显的差距, 差距在统计学上是否显著。一般而言, 我就需要采用假设检验的方法, 假设检验的反面就是区间估计, 所以在此我们先提到。所以在 coding 的时候, 我们一般都要先判断一下方差是否显著相等, 原因就在于如果方差显著相等, 那么均值差的区间估计就会更容易。

a. 两个均值差 $\mu_1 - \mu_2$ 的区间估计

这一参数的估计有着重要的意义, 可以判断两个样本所属整体分布均值是否相同, 应用于实际情况就是判断使用某一疗效后恢复速度是否明显变好; 加了某一条件后, 结果有无理想改变等。对于这一参数的估计, 总体分为以下五种情况

- σ_1^2 和 σ_2^2 均已知;
- $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 但具体值未知;
- $\sigma_1^2/\sigma_2^2 = \theta$ 已知时;
- 当 m 和 n 都很大, 但方差没有任何信息时;
- 一般情况下的近似置信区间 —— 历史上著名的 Behrens-Fisher 问题。

b. σ_1^2/σ_2^2 的置信区间

由于 $(m-1)s_x^2/\sigma_1^2 \sim \chi^2(m-1)$, $(n-1)s_y^2/\sigma_2^2 \sim \chi^2(n-1)$ 且 s_x^2 与 s_y^2 相互独立, 故可仿照 F 变量构造如下枢轴量:

$$\begin{aligned} & \text{\$ \$} \\ & F = \frac{s_x^2/\sigma_1^2}{s_y^2/\sigma_2^2} \sim F(m-1, n-1) \\ & \text{\$ \$} \end{aligned}$$

进而得到其区间估计为:

$$\begin{aligned} & \text{\$ \$} \\ & \left[\frac{s_x^2}{s_y^2} \cdot \frac{1}{F_{1-\alpha/2}(m-1, n-1)}, \frac{s_x^2}{s_y^2} \cdot \frac{1}{F_{\alpha/2}(m-1, n-1)} \right] \\ & \text{\$ \$} \end{aligned}$$

第七章 假设检验

这和上面一章参数估计是一脉相承的, 上一章要求估计出某个参数的准确值。现在我们要根据样本情况, 验证某个参数是否满足某些条件, 在假设成立的前提下, 小概率事件是否发生。

定义：从小概率事件是否发生的角度来讲这个故事：不论是参数估计还是假设检验，我们针对的始终是总体的参数。在不知道总体的某个参数时，可以假设该参数符合某种条件（比如均值为0），带着这个假设进入到实际样本中。问这么一个问题：在已有假设的基础上，实际样本产生的概率有多大，是不是一个小概率事件？如果在此假设下，样本发生的概率极小，反过来极小概率的事件发生了，那很有可能是假设错了，我们完全可以拒绝原假设。因此，关键在于判断该样本产生是否为小概率事件——基于已知统计量构建分布，同时怎么定义小概率事件？到底概率多小才算小，这就引出了置信水平或显著性水平。

两类错误：

1. 第一类错误：小概率事件不是没有可能发生，假设定义发生概率为5%的事件为小概率事件，那也意味着在原假设成立的情况下，仍有5%的可能性会发生样本所看到的小概率事件，因此如果直接把原假设拒绝了，那就有5%的可能性犯错。原假设明明为正确的，但却把他拒绝了的概率，就是显著性水平
2. 第二类错误：原假设明明错了，但是根据样本计算的统计量却接受了它的概率。这个算起来就不容易了，首先就需要正确的统计量到底是什么，然后带到犯错的区间内去求概率。

基本步骤：

1. 根据实际情况提出原假设 H_0 和备择假设 H_1 。如果是双边的 H_0 肯定是等于， H_1 是不等于。但如果是单边的话刚开始设定 H_0 为与题中相反的方向（题中说减小，这个地方就设定为大于等于），备择假设和题中所说相符，然后将 H_0 转化为等于。
2. 假设 H_0 成立，构造适当的统计量
3. 基于该统计量，给定置信水平 α 根据统计量的分布情况查表，确定拒绝域 W
4. 根据样本的观察值计算统计量的值，将其与拒绝域作比较并下结论。

统计量的构建 和上一章完全相同的四种情况，注意拒绝域的符号不要搞错。