概率论与统计小书

luisleee

2025 年 12 月 29 日

1 概率空间

一切概率论的问题都应该被放在概率空间里进行研究。所谓概率空间，就是一个三元组，通常记作 $(\Omega, \mathcal{F}, P)$，其中：

Ω 是**样本空间**，代表所有可能的实验结果的集合。

$\mathcal{F}$ 是 Ω 上的 σ-代数，它是一个集合族，满足以下条件：

- 包含样本空间 Ω，即 $\Omega \in \mathcal{F}$。
- 若事件 $A \in \mathcal{F}$，则其补集 $A^c \in \mathcal{F}$。(这隐含 $\emptyset \in \mathcal{F}$)
- 若事件序列 $A_1, A_2, \ldots \in \mathcal{F}$，则它们的可数并集 $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$。

这是所有**事件**的集合。一个事件就是包含一些基本结果的集合。

P 是**概率测度**，它是定义在 $\mathcal{F}$ 上的函数，满足：

- 对于任意事件 $A \in \mathcal{F}$，$P(A) \geq 0$。
- $P(\Omega) = 1$，即整个样本空间的概率为 1。这一性质也叫归一性。
- 若 $A_1, A_2, \ldots$ 是一列可数的两两不相交的事件，则 $P\left(\bigsqcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$。

概率 $P(A)$ 就是衡量事件 A 发生的可能性的值，它在 $[0, 1]$ 范围内。

这样的定义，保证了我们对的概率定义是良好的，同时我们也有集合论和 Lebesgue 积分等丰富的工具来研究概率。

事件的本性就是一个集合，于是我们可以用集合论的手法来研究它。事件的运算就是集合的运算。对于两个事件 A, B：

- 和事件：$A + B = A \cup B$ 表示事件 A ，B 至少有一个事件发生。
- 积事件：$AB = A \cap B$ 表示事件 A 和事件 B 同时发生。
- 逆事件：$\overline{A}$ 表示事件 A 不发生。
- 差事件：$A - B$ 表示事件 A 发生但事件 B 不发生。

这些常见的构造也自然满足各种布尔代数性质，这里略过。

事件之间的关系另有一套术语。$A \cap B = \emptyset$ 称为 A，B **互斥**。A 和 $\overline{A}$ 称为一对**对立**事件。$A \subset B$ 称为 B **包含** A，也就是 A 发生一定 B 发生。

根据定义，我们就可以得出事件运算和概率的关系。几个简单的的结论：

$$P(\emptyset) = 0$$

若一族有限的事件 A_i 两两互斥（不交），

$$P(\bigsqcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} A_i$$

若 $A \subset B$，

$$P(B - A) = P(B) - P(A)$$
$$P(B) \geq P(A)$$

$$P(A) \leq P(\Omega) = 1$$

$$P(\overline{A}) = 1 - P(A)$$

概率另有一种连续性的存在。对于一族可数的无穷递增事件 $A_1 \subset A_2 \subset \cdots$，令 $A = \bigcup_{i=1}^{\infty} A_i = \lim_{n\to\infty} A_n$，有：

$$P(A) = \lim_{n\to\infty} P(A_n)$$

这一事实基于可数可加性。令 $A_i^* = A_{i+1} - A_i, A_0^* = A_1$，那么我可以把 A 表示成 $\bigsqcup_{i=0}^{\infty} A_i^*$，证明也就基本结束了。

对于不一定互斥的事件，他们和事件的概率是这样的：

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

这一经典结论可以推广到任意有限多个事件的和。

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{k=1}^{n}(-1)^{k-1} \sum_{1\leq i_1<i_2<\cdots<i_k\leq n} P(A_{i_1}A_{i_2}\cdots A_{i_k})$$

这个又叫容斥原理。证明方法很多样，一种简单的方式是使用归纳法。我在这里展示一种莫比乌斯反演的解释。

证明. 为了方便叙述，我们令集合 $N = \{1, \ldots, n\}$，我们后面的集合下标就会从 N 中取。N 的幂集 2^N 中的一个元素就是一组下标的集合 $i_1, \ldots, i_k$。

我们想要求的 $P(\bigcup_{i=1}^{n} A_i)$ 实际不是一个很方便处理的形式。比较方便处理的是

$$P(\bigcap_{i=1}^{n} \overline{A_i}) = 1 - P(\bigcup_{i=1}^{n} A_i)$$

接下来，只要能够处理这种积事件的概率，我们就赢了。

我们为了方便计数，我们一定选择不交的事件的积来计算概率。给出一族下标 $S \in 2^N$，我可以给出集合 $X_S = (\bigcap_{i \in S} A_i) \cap (\bigcap_{i \notin S} \overline{A_i})$。对于任意选定的不同下标组合，对应的这个集合都不交。我们通过这些集合的并，就能轻松组合出任意 $Y_S = \bigcap_{i \in S} A_i$。(注意对于 $S = \emptyset$，我们可以认为这个交为 Ω 而不产生任何问题。对偶地，我们认为空并为 $\emptyset$。)

比方说对于下标集合 S，那么我这样就能从一些 X 不交并出集合 Y_S：

$$Y_S = \bigsqcup_{S \subset T} X_T$$

对应到概率，令 $f(S) = P(X_S), g(S) = P(Y_S)$，有

$$g(S) = \sum_{S \subset T} f(T)$$

这给我们一个方便的莫比乌斯反演：2^N，加上 $\subset$ 构成的偏序集上面的莫比乌斯反演函数是 $\mu(S, T) = (-1)^{|S|-|T|}$。莫比乌斯反演的玩法就是，有了上面这个式子，就可以立马得到：

$$f(S) = \sum_{S \subset T} \mu(S, T) g(T) = \sum_{S \subset T} (-1)^{|S|-|T|} g(T)$$

我们上面想要的东西就是 $f(\emptyset)$，按照反演公式，

$$
\begin{aligned}
1 - P(\bigcup_{i=1}^{n} A_i) &= P(\bigcap_{i=1}^{n} \overline{A_i}) \\
&= f(\emptyset) \\
&= \sum_{S \in 2^N} (-1)^{|S|} g(S) \\
&= \sum_{S \in 2^N} (-1)^{|S|} P(\bigcap_{i \in S} A_i) \\
&= \sum_{k=0}^{n} (-1)^k \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} P(A_{i_1} A_{i_2} \cdots A_{i_k})
\end{aligned}
$$

好嘛！证明基本就完成了，只差一个 1 和一个负号！单独把 $k = 0$ 的情况拿出来，问题就解决了，证明就完成了。

□

从容斥原理可以得到两个方便近似的不等式：Boole 不等式和 Bonferroni 不等式。

如果我们只取容斥原理的单事件概率部分，就有：

$$
P(\bigcup_{i=1}^{n} A_i) \le \sum_{1 \le i \le n} P(A_i)
$$

如果我们只取到奇数个事件之积，那会得到一个概率的上界；只取到偶数个事件之积，那就会得到一个下界：

$$
\sum_{k=1}^{m} (-1)^{2k-1} \sum_{1 \le i_1 < i_2 < \cdots < i_2 k \le n} P(A_{i_1} A_{i_2} \cdots A_{i_{2k}}) \le P(\bigcup_{i=1}^{n} A_i)
$$

$$
P(\bigcup_{i=1}^{n} A_i) \le \sum_{k=1}^{m} (-1)^{2k-2} \sum_{1 \le i_1 < i_2 < \cdots < i_{2k-1} \le n} P(A_{i_1} A_{i_2} \cdots A_{i_{2k-1}})
$$

这得益于式子的正负交错性质。

条件概率在直觉上表示的是，在某个事件 $A, P(A) \neq 0$ 已发生的条件下，另一个事件 B 发生的概率，记作 $P(B|A)$。我们可以认为，这由事件 $A, P(A) \neq 0$ 诱导的一个新的概率空间给出。这个概率空间的样本空间有 $\Omega' = \Omega$，$\mathcal{F}' = \mathcal{F}$，而条件概率 $P(B|A) = \frac{P(A \cap B)}{P(A)}$。很容易验证，它满足概率空间公理。

事件**独立性**的讨论和条件概率有关。事件 A 与 B 独立也就是 $P(A|B) = P(A)$ 或者 $P(B|A) = P(B)$。不过这里需要注意的是，刚才的条件概率不允许条件发生的概率为 0。为了不失一般性，我们可以认为概率为 0 的事件和其他任意事件都是无关的。刚才的式子可以修改成 $P(A)P(B) = P(AB)$。

对于多个事件的无关性，我们起码要求其中任意两个事件之间独立，同时其中任意三个、四个、乃至更多事件放在一起也满足一定的独立性条件。这个独立性条件可以类比为 $\prod_{i=1}^{n} P(A_i) = P(\bigcap_{i=1}^{n} A_i)$。这一概念推广到可数多事件也是不费力的：可数多事件 $A_1, A_2, \ldots$ 独立当且仅当

$$\forall I \in 2^{\mathbb{N}}, \prod_{i \in I} P(A_i) = P(\bigcap_{i \in I} A_i)$$

关于条件概率，几个重要的结论是全概率公式和 Bayes 公式。全概率公式告诉我们如何通过一些条件概率求出事件的概率。令 B 为某个事件，$A_k, k = 1, 2, \cdots$ 为一族可数的事件，且 $\bigsqcup_{i=1}^{\infty} A_i = \Omega, P(A_i) \neq 0$，则有：

$$P(B) = \sum_{i=1}^{\infty} P(B|A_i)P(A_i)$$

这是比较容易理解的：我们无一遗漏无一重复地考虑所有可能的情况 A_i，然后计算某种情况发生的概率 $P(A_i)$，乘上 $P(B|A_i)$ 就得到情况 A_i 与事件 B 同时发生的概率 $P(BA_i)$。对 i 求和就得到了 B 发生的所有可能情况，就是 B 的概率。

如果再进一步，我们就可以通过诸 $P(B|A_i)$ 反过来算出诸 $P(A_i|B)$。因为 $P(A_i|B) = \frac{P(A_iB)}{P(B)}$。上面的式子可以按定义展开，下面的式子可以由刚才的全概率公式给出。也就有：

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{\infty} P(B|A_i)P(A_i)}$$

有人喜欢把这个称为先后验概率来进行阐释，赋予它一种形而上学意味。我则严格认为，它应该与任何这种解释撇清关系。这会把主观性重新带回我们费尽心思公理化的概率论中，让人感到错误的惊异，并无助于解释的清晰性。

2 随机变量

当我们把样本空间映射成数字来处理时，这时我们才有了更丰富的工具。

一个**随机变量**就是 $\Omega \to \mathbb{R}$ 的一个函数，而且是可测的。这样就可以在实数轴上很好地研究它的性质了。

对于一个随机变量 X，给出一个实数上的可测集合 S，定义概率 $P(X \in S) = P(\{e \in \Omega : X(e) \in S\})$。我们同样可以通过分离公理，用一个和 X 取值相关的谓词 $\mathcal{P}(X)$ 描述集合 $S = \{x \in \mathbb{R} : \mathcal{P}(x)\}$，这时我们简写 $P(X \in S)$ 为 $P(\mathcal{P}(X))$。例如 $P(1 < X \leq 2)$ 表示 X 取值大于 1 小于等于 2 的概率。

随机变量有离散型和连续性之分，但本质上都可以用同一套理论描述：离散随机变量使用离散测度处理，连续随机变量使用 Lebesgue 测度处理。它们的结论基本上通过互换 $\int$ 和 Σ，D 和 Δ 就能互相转换。不过我在这里仍然使用不同的记号。

2.1 随机变量的分布

随机变量 X 的**累计分布函数**是这样一个函数 $F_X(x) = P(X \leq x)$。很明显，它有这样的性质：

$$\lim_{x \to -\infty} F_X(x) = 0,\ \lim_{x \to +\infty} F_X(x) = 1,\ \lim_{x \to x_0^+} F_X(x) = F(x_0)$$

它们都是从前面概率的连续性，归一性推导出来的。

由概率的非负性，

$$\forall x_1 < x_2, F(x_2) - F(x_1) = P(x_1 < X \leq x_2) \geq 0$$

并且

$$0 \leq F(x) \leq 1$$

从测度的视角看，概率密度函数 f_X 是随机变量 X 的分布测度 P_X 相对于 Lebesgue 测度 λ 的 Radon-Nikodym 导数。即：

若 P_X 关于 Lebesgue 测度绝对连续（记为 $P_X \ll \lambda$），则存在非负可测函数 $f_X : \mathbb{R} \to [0, \infty)$，使得对任意可测集 $B \subseteq \mathbb{R}$，

$$P_X(B) = \int_B f_X(t)\, d\lambda(t) = \int_B f_X(t)\, dt.$$

此时 f_X 称为 X 的概率密度函数，它在几乎处处意义下唯一确定。下面给出的是更简单情况下的概率密度函数。

2.1.1 离散型随机变量

离散型随机变量只能取至多可数个值。我们可以通过列举所有可能取值发生的概率，来确定这个随机变量的统计规律。

如果一个离散型随机变量 X 可能能取值为 $\{x_k\}, k = 1, 2, \cdots$，则我们可以知道 $P(X = x_k)$。我们知道，一个函数的纤维划分定义域，$\bigsqcup_{k=1}^{\infty} X^{-1}(x_k) = \Omega$。所以根据可数可加性：

$$\sum_{k=1}^{\infty} P(X = x_k) = 1$$

虽然很简单，但是还是值得注意：$0 \leq P(X = x_k) \leq 1$。

$P(X = x_k) = p_k$ 这种表现形式叫做随机变量的**分布律**，也叫**概率质量函数**。

2.1.2 连续型随机变量

描述连续型随机变量时，有概率密度函数这种更好的工具。对于一个离散型随机变量，它的累积分布函数一定是具有各种跳变不连续点的，并且在两个跳变点之间都是常值的。连续型随机变量的累计分布函数如果很大程度上（“几乎处处”）是可导的，此时也就可以用**概率密度函数**来描述。

一个连续型随机变量在某一点处的的概率密度函数是累计分布函数的导数（当然它也可能不存在）。有了概率密度函数 f_X，我们就可以从积分来得出累积分布函数：

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

它被称为概率密度函数也就是因为，$f_X(t)$ 表示了在这一处一个无限小的区间内的“概率质量”的集中程度。$f_X(t)dt$ 就可以认为是 $P(t < X < t + dt)$。

很明显，如果概率密度函数存在，那么它满足：

非负：

$$f_X(x) \geq 0$$

归一：

$$\int_{-\infty}^{\infty} f_X(t)dt = 1$$

事实上对于任意一个可积的非负函数，我都可以乘上一个系数让它归一，从而满足概率密度的要求。利用这一事实往往可以节约计算。

并且此时，我可以通过它求随机变量落在任意可测集合内的概率：

$$P(X \in S) = \int_S f_X(t)dt$$

特别地，$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$。

如刚才的叙述，它们在测度视角下都是一样的，只不过是换了种测度。

2.1.3 随机变量的函数

给一个随机变量 X 后复合上一个可测函数 $\phi : \mathbb{R} \to \mathbb{R}$，我仍然会得到一个随机变量 $\phi \circ X$。我们会把这个新的随机变量写成 $\phi(X)$。

这个新的随机变量显然也存在自己的分布分布规律。它的累积分布函数也就可以表示成 $F_{\phi(X)}(x) = P(\phi(X) \leq x)$。对于一般的连续型随机变量，想要直接通过概率密度函数来计算 $\phi(X)$ 的概率密度函数是不太现实的，我们须要积分得到累积分布函数再找到 $\phi(X)$ 的累计分布函数，再做微分。

不过当 ϕ 是个严格单调函数时，就能直接计算出概率密度了。此时

$$f_{\phi(X)}(y) = \frac{f_X(\phi^{-1}(y))}{|\phi'(\phi^{-1}(y))|} = f_X(\phi^{-1}(y))|(\phi^{-1})'(y)|$$

我们在此不作证明，留在后面用一般性的结论处理。

作为一个简单的例子，数乘 $X \mapsto \lambda X$ 是一个非常简单的随机变量函数。如果 $\lambda = 0$，情况非常简单，我们跳过。如果 $\lambda \neq 0$，那么

$$f_{\lambda X}(x) = \frac{1}{|\lambda|} f_X(\frac{x}{\lambda})$$

对于离散型随机变量，$\phi(X)$ 的分布律可以通过对纤维直接求和得到：

$$P(\phi(X) = y) = \sum_{x \in \phi^{-1}(y)} P(X = x)$$

3 多元随机变量

所谓多元随机变量，就是多个随机变量组合在一起，像 (X,Y)。我们为了便利，可以把它们排列成向量的形式，也叫**随机向量**。

同样，我们也可以计算 n 元随机变量 $\mathbf{X}$ 的取值落在 $\mathbb{R}^n$ 中一个可测子集 S 内的概率 $P(\mathbf{X}\in S)$。

它的分布可以用**联合累积分布函数**来描述。连续型随机变量可以用**联合概率密度**描述，离散型随机变量可以用**联合分布律**描述。

对于 n 元随机变量 $\mathbf{X}=(X_1,X_2,\ldots,X_n)$ 和 n 元数对 $\mathbf{x}=(x_1,x_2,\ldots,x_n)$，它的联合累积分布函数定义为

$$F_{\mathbf{X}}(\mathbf{x})=P(X_1\leq x_1,X_2\leq x_2,\cdots,X_n\leq x_n)$$

可以确定，它仍然具有非负性，归一性，而且对于任意一个维度右连续且不减。

对于 n 元离散型随机变量，联合分布律也与一元有相同的形式，也就是 $P(\mathbf{X}=\mathbf{x}_k)=p_k$ 这种形式。

对于 n 元连续型随机变量，联合概率密度与一元有相同的形式，

$$f_{\mathbf{X}}(\mathbf{x})=\frac{\partial^n}{\partial x_1\partial x_2\cdots\partial x_n}F_{\mathbf{X}}(\mathbf{x})$$

即概率密度函数是累积分布函数的 n 重偏导数。当然这里前提是分布函数满足了某种光滑性，各阶偏导可交换。显然联合概率密度也有非负性和归一性。

如果给出多元随机变量的联合概率密度，我也就可以通过 n 重积分解答刚才的问题：

$$P(\mathbf{X}\in S)=\int_S f_{\mathbf{X}}(\mathbf{x})d\mathbf{x}$$

用测度的语言描述，就是说：

对于 n 维随机向量 $\mathbf{X}=(X_1,\ldots,X_n)^T$，其分布是 $\mathbb{R}^n$ 上的概率测度 $P_{\mathbf{X}}$，定义为：对任意可测集 $B\subseteq\mathbb{R}^n$，

$$P_{\mathbf{X}}(B)=P(\mathbf{X}\in B).$$

多元的情况的一元的情况是完全一样的，概率密度就是测度的 Radon-Nikodym 导数。

3.1 边缘分布

如果多元随机变量的联合分布 $F_{\mathbf{X}}$ 已知，那它其实已经蕴含了单个随机变量的分布 F_{X_1}。

$$F_{X_1}(x) = P(X_1 \leq x) = F_{\mathbf{X}}(x, \infty, \ldots, \infty)$$

同样，给出联合分布律，我也能求出单个随机变量的分布律（其中 $\pi_1 : \mathbb{R}^n \to \mathbb{R}$ 指第一个分量上的投影）：

$$P(X_1 = x) = \sum_{\mathbf{x} \in \pi_1^{-1}(x)} P(\mathbf{X} = \mathbf{x})$$

给出联合概率密度，可以求出单个随机变量的概率密度：

$$f_{X_1}(x) = \int_{\pi_1(\mathbf{x}) = x} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$$

以上这种单个变量的分布规律称为边缘分布规律，也就是**边缘累积分布函数**，**边缘分布律**，**边缘概率密度**。

3.2 条件分布

同样的，我们也可以给出一种条件分布，即已知某一（些）随机变量取特定值时，另一随机变量的分布规律。在离散情况下，这就是条件概率：因为对于任意 X_1 的取值，它的边缘概率质量都是不为 0 的。于是我们可以定义

$$P(\mathbf{X} = \mathbf{x} | X_1 = x_1) = \frac{P(\mathbf{X} = \mathbf{x})}{P(X_1 = x_1)}$$

分母的概率就是一个边缘的概率。

连续情况下，则会遇到一个问题，因为连续随机变量取单点的概率为零，无法适用条件概率的公式。但我们仍然可以仿照离散的形式给出：

$$f_{\mathbf{X}|X_1}(\mathbf{x}|x_1) = \frac{f_{\mathbf{X}}(\mathbf{x})}{f_{X_1}(x_1)}$$

这一定义确实在 $f_{X_1}(x_1) \neq 0$ 时是良好定义的，并且和我们的直觉是一致的。

我们最好把条件概率密度理解为在固定某个（些）随机变量的条件下诱导出的子空间上的概率测度。

写成累积分布函数的形式也是可以的，只需要一个积分。

3.3 独立性

同条件概率与事件独立性的关系一样，我们也可以通过条件分布研究随机变量独立性的关系。类比于事件的独立性，两个随机变量 X, Y 相互独立，当且仅当对于任意可测集 $A, B \subset \mathbb{R}$，

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

当然为了便利，用累积分布函数定义的形式是等价的。

$$\forall x, y \in \mathbb{R}, \quad F_{XY}(x, y) = F_X(x)F_Y(y)$$

对于离散随机变量，我们可以用：

$$\forall x, y \in \mathbb{R}, \quad P(X = x, Y = y) = P(X = x)P(Y = y)$$

对于连续随机变量，我们可以用：

$$\forall x, y \in \mathbb{R}, \quad f_{XY}(x, y) = f_X(x)f_Y(y)$$

用条件分布的语言来说，就是条件分布 $F_{Y|X}(y|x)$ 等于边缘分布 $F_Y(y)$。

多个随机变量 $X_1, X_2, \ldots, X_n$ **相互独立**是指：对于任意可测集 $A_1, A_2, \ldots, A_n \subset \mathbb{R}$，

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = P(X_1 \in A_1)P(X_2 \in A_2) \cdots P(X_n \in A_n).$$

用累积分布函数的形式就是，对于其中**任取**的 n 个随机变量，都有：

$$F_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} F_{X_i}(x_i)$$

这比它们两两独立要强很多。

这和张量积的表示是一致的。如果这些变量相互独立，那我就可以把这个联合分布写成诸边缘分布的张量积。

3.4 多元随机变量函数

类似于一元随机变量函数，我也可以定义多元随机变量函数。我可以用一个可测函数 $\phi : \mathbb{R}^n \to \mathbb{R}^m$ 后复合于一个随机向量 $\mathbf{X}$，又得到一个变换后的随机向量 $\phi(\mathbf{X})$。不过这里为了方便分析，我们先讨论 $\phi : \mathbb{R}^n \to \mathbb{R}$。讨论清楚之后，多维随机向量值变换也就可以分解成多个这样的函数来讨论了。

我们目前熟知的这样的函数有诸投影函数 π_i，它给出边缘分布。

另一类经典的函数是加减法和乘除法。我们先处理两个随机变量的情况，多个随机变量的情况也就能轻松归纳出了。

这是连续型随机变量的情况，在离散型的情况下只需要把积分换成求和，概率密度换成概率质量。我在这里姑且不写出了。

$$
\begin{aligned}
F_{X+Y}(z) &= P(X+Y \le z) \\
&= \iint_{x+y\le z} f_{XY}(x,y)dxdy \\
&= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{z-y} f_{XY}(x,y)dx\right]dy \\
&= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{z} f_{XY}(u-y,y)du\right]dy && \text{u=z-y} \\
&= \int_{-\infty}^{z}\left[\int_{-\infty}^{\infty} f_{XY}(u-y,y)dy\right]du && \text{Fubini 定理} \\
f_{X+Y}(z) &= \frac{dF_{X+Y}(z)}{dz} = \int_{-\infty}^{\infty} f_{XY}(z-t,t)dt && \text{微积分基本定理}
\end{aligned}
$$

这就是卷积公式。如果 X 和 Y 独立，那性质敢情更好了。里面的 $f_{XY}(z-t,t)$ 就可以拆成 $f_X(z-t)f_Y(t)$。顺带一提，虽然这个公式乍一看不对称，实际上是对称的，和加法的交换性是一致的。

做乘除法也有类似的公式，令 $Z = XY$：

$$
\begin{aligned}
F_Z(z) &= P(XY \le z) \\
&= \iint_{xy \le z} f_{XY}(x,y)\,dx\,dy \\
&= \int_{-\infty}^{0} \left[\int_{z/x}^{\infty} f_{XY}(x,y)\,dy \right] dx + \int_{0}^{\infty} \left[\int_{-\infty}^{z/x} f_{XY}(x,y)\,dy \right] dx \\
&= \int_{-\infty}^{0} \left[\int_{-\infty}^{z/x} f_{XY}(x,y)\,dy \right] dx + \int_{0}^{\infty} \left[\int_{-\infty}^{z/x} f_{XY}(x,y)\,dy \right] dx \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{z/x} f_{XY}(x,y)\,dy \right] dx \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{z} f_{XY}\left(x, \frac{u}{x}\right) \frac{1}{|x|}\,du \right] dx && u = xy \\
&= \int_{-\infty}^{z} \left[\int_{-\infty}^{\infty} f_{XY}\left(x, \frac{u}{x}\right) \frac{1}{|x|}\,dx \right] du && \text{Fubini 定理} \\
f_Z(z) &= \frac{dF_{XY}(z)}{dz} = \int_{-\infty}^{\infty} f_{XY}\left(x, \frac{z}{x}\right) \frac{1}{|x|}\,dx && \text{微积分基本定理}
\end{aligned}
$$

除法是类似的，令 $Z = Y/X$：

$$f_Z(z) = \int_{-\infty}^{\infty} f_{XY}(x, zx)|x|dx$$

不过这个公式就不对称了。

我们上面处理的手法都是求一次 n 重积分得到累积分布函数，然后求一次导数得到概率密度。在 n 大起来之后这种处理可能略微繁琐，不过大部分情况下我们都只能这么做。不过小部分情况下，我猜想我们可以只用做一个 $n-1$ 重的积分就可以了。这一想法最初始于从离散型随机变量的类比。

3.4.1 多维离散型随机变量的函数的概率密度

离散型随机变量通过其分布律被刻画。

设随机变量 $\mathbf{X} = (X_1, X_2, \ldots, X_n)$，其可能取值为 $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}, \ldots$（这里的 $\mathbf{x}^{(i)} \in \mathbb{R}^n$）。则其分布律可表示为：

$$P\{\mathbf{X} = \mathbf{x}^{(i)}\} = p_i, \qquad i = 1, 2, \ldots, m, \ldots$$

其分布函数则无需积分，只需要进行离散求和：

$$F_{\mathbf{X}}(\mathbf{x}) = P\left\{X_1 \leq x_1,\, X_2 \leq x_2,\, \ldots,\, X_n \leq x_n\right\}$$

即

$$F_{\mathbf{X}}(x_1, x_2, \ldots, x_n) = \sum_{\mathbf{x}^{(i)} \leq \mathbf{x}} P\{\mathbf{X} = \mathbf{x}^{(i)}\}$$

其中，

$$\mathbf{x}^{(i)} \leq \mathbf{x} \iff x_1^{(i)} \leq x_1,\, x_2^{(i)} \leq x_2,\, \ldots,\, x_n^{(i)} \leq x_n$$

对于多维离散型随机变量，其函数 $Y = g(\mathbf{X})$ 的分布律有更简单的表示。

设 $Y = g(\mathbf{X})$，那么 Y 的可能取值为 $y_1, y_2, \ldots, y_m, \ldots$，其分布律由下式给出：

$$P\{Y = y_j\} = \sum_{\mathbf{x} \in g^{-1}(y_j)} P\{\mathbf{X} = \mathbf{x}\}, \qquad j = 1, 2, \ldots, m, \ldots$$

其中 $g^{-1}(y_j)$ 是 y_j 的纤维。

同样也可以定义其分布函数，设随机变量 $Y = g(\mathbf{X})$，则其分布函数定义为：

$$F_Y(y) = P\{Y \leq y\}$$

由于 Y 是离散型随机变量，其分布函数可以写为：

$$F_Y(y) = \sum_{y_j \leq y} P\{Y = y_j\}$$

类比于连续型随机变量的函数概率密度计算方法，在离散型情形下，可以通过联合分布律先计算分布函数，然后差分得到函数的分布律。具体表达式如下：

$$F_Y(y) = P\{Y \leq y\} = \sum_{g(\mathbf{x}) \leq y} P\{\mathbf{X} = \mathbf{x}\}$$

$$P\{Y = y\} = F_Y(y) - F_Y(y^-) = \sum_{g(\mathbf{x}) = y} P\{\mathbf{X} = \mathbf{x}\}$$

当然，如果我们只想求得它的分布律，计算分布函数是个多余的步骤，我们只需要对 y 对应的纤维 $g^{-1}(y)$ 进行其求和即可。接下来我将把这种直觉类比到连续性随机变量函数的计算中。

3.4.2 多维连续随机变量一阶连续可微无临界点函数的概率密度

这个名字很长，不过足够保证我们安全地得到很好的结论，不过你也可以不看。离散情况下的处理是直接对函数纤维的分布律进行求和。能想到的最自然的一种类比就是直接将纤维的概率密度进行积分。直接进行 n 重积分是有问题的，因为纤维很有可能在 $\mathbb{R}^n$ 中零测。合理的做法是将积分限制在纤维这一低维结构上，也就是做一个第一类面（线）积分。

但是直接计算这一积分是不正确的，因为对于连续型随机变量，概率密度不只取决于纤维上的概率分布，还与每一点的局部性质密切相关。更准确地说，与离散随机变量不同，概率质量在从原始变量空间射到像空间时，会被变换函数 g 在每一点上的伸缩所影响。不过，这一效应可以用该函数在纤维方向正交方向上的雅可比行列式来刻画。因此，在进行积分时，必须引入恰当的权重项（即一个雅可比因子）以正确反映概率密度在变量变换下的变化情况。忽略这些因素的式子只能在某些特殊情况下成立（如仿射变换），而对于一般的函数，只有严格考虑这些因素，才能得到准确的新概率密度函数。

为了简洁性，我们在此限制 g 为 $\mathbb{R}^n \to \mathbb{R}$ 的一阶连续可微函数。由于它是一个 C^1 函数，对于几乎所有 z，它对应的纤维 $g^{-1}(z)$ 一定是一个低一维的子流形，也就是一条等值线（面）；只有在某些特殊的 z 值处，纤维会退化成 n 维结构或者离散点。更进一步，为了避免这些奇点带来的技术问题，我们可以假定对于每个 z，$\forall (x,y) \in g^{-1}(z), \nabla g(x,y) \neq 0$。这样一来，纤维在邻域内的结构就更加良好了：等值线将会是一条一阶连续可微的曲线（面）。于是使用余面积公式进行处理也就呼之欲出了。

定理 1 (n 维余面积公式)**.** *设 $g : \mathbb{R}^n \to \mathbb{R}$ 是 C^1 函数（$n \geq 2$），且其梯度 $\nabla g(\mathbf{x})$ 在 $g^{-1}([a,b])$ 上处处非零；$f(\mathbf{x})$ 是 $g^{-1}([a,b])$ 上的连续函数。则有：*

$$\int_{g^{-1}([a,b])} f(\mathbf{x})\, d\mathbf{x} = \int_{t=a}^{b} \left[\int_{g^{-1}(t)} \frac{f(\mathbf{x})}{|\nabla g(\mathbf{x})|}\, d\sigma(\mathbf{x}) \right] dt$$

其中 $d\sigma(\mathbf{x})$ 表示在纤维（超曲面）$g(\mathbf{x}) = t$ 上的 $(n-1)$ 维面积微元。

证明. 考虑积分区域 $\Omega := g^{-1}([a,b])$。对积分进行分层处理：每个 $t \in [a,b]$ 对应的纤维 $M_t := \{x \in \mathbb{R}^n \mid g(x) = t\}$ 是一个 $(n-1)$ 维一阶连续可微子流形。

由隐函数存在性定理，$g(\mathbf{x})$ 在 $\nabla g(\mathbf{x}) \neq 0$ 处可作为新的变量替换的一部分坐标，其余 $(n-1)$ 维由 M_t 上参数化给出。

设局部参数化为

$$x = \Phi(y_1, \ldots, y_{n-1}, t), \qquad (y_1, \ldots, y_{n-1}) \in D \subset \mathbb{R}^{n-1},\ t \in [a, b]$$

换元公式下，整体 n 维体积微元可拆解为：

$$dx = J_\Phi(y_1, \ldots, y_{n-1}, t)\, dy_1 \cdots dy_{n-1} dt$$

按照隐函数定理，Jacobi 因子正好为 $|\nabla g(\mathbf{x})|$，即

$$d\mathbf{x} = \frac{1}{|\nabla g(\mathbf{x})|} d\sigma(\mathbf{x}) dt$$

于是有

$$\int_\Omega f(x)\, d\mathbf{x} = \int_{t=a}^{b} \left[\int_{M_t} \frac{f(\mathbf{x})}{|\nabla g(\mathbf{x})|}\, d\sigma(\mathbf{x}) \right] dt$$

其中 $d\sigma(\mathbf{x})$ 为 M_t 上的 $(n-1)$ 维面积微元。即为所证。 □

那么 n 维连续型随机变量函数的概率密度就可以计算了。

设 $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ 是定义在概率空间 $(\Omega, \mathcal{F}, P)$ 上的 n 维连续型随机向量，其概率密度函数为 $f_{\mathbf{X}}(\mathbf{x})$。设 $Y = g(\mathbf{X})$ 的概率密度函数为 $f_Y(y)$，其中 g 是一阶连续可微函数，且 ∇g 在每个纤维 $g^{-1}(y)$ 上都不为 0。

由 n 维余面积公式，

$$F_Y(y) = \int_\Omega f(x)\, d\mathbf{x} = \int_{-\infty}^{y} \left[\int_{g(\mathbf{x})=t} \frac{f(\mathbf{x})}{|\nabla g(\mathbf{x})|}\, d\sigma(\mathbf{x}) \right] dt$$

其中，$\Omega = g^{-1}((-\infty, y]) = \{\mathbf{x} \in \mathbb{R}^n : g(\mathbf{x}) \le y\}$，$d\sigma(\mathbf{x})$ 为 $g^{-1}(t)$ 上的 $(n-1)$ 维面积微元。

求导，应用微积分基本定理，得到：

$f_Y(y) = \frac{dF_Y(y)}{dy} = \int_{g(\mathbf{x})=t} \frac{f(\mathbf{x})}{|\nabla g(\mathbf{x})|}\, d\sigma(\mathbf{x})$

一个简单的公式至此形成。它甚至覆盖了一元的情况：

$n = 1$ 的情况下，函数 g 的纤维基本都退化为了离散点。这种情况下纤维上积分的处理和高维有所不同。

不过幸运的是，我们仍然可以在这些离散点上定义合适的测度——即离散测度（或 Dirac 测度）。在这样的测度下，对这些点的积分自然就变成了单纯的加和：

$$\int_{g^{-1}(t)} f(x)\, d\mu(x) = \sum_{x \in g^{-1}(t)} f(x)$$

其中 μ 表示对点集的离散测度。这样，虽然在 Lebesgue 测度下单个点集为零测，但在离散测度下依然可以有意义地积分，因此仍然能够推广高维余面积公式到一维的情形。

若 $g:\mathbb{R}\to\mathbb{R}$ 为一阶可微函数，对于 $t\in[g(a),g(b)]$，其纤维 $g^{-1}(t)=\{x\in\mathbb{R}:g(x)=t\}$ 是离散点集（假设 g 可微且 $g\prime(x)\neq 0$）。在此情况下，余面积公式也可以得到一维积分的推广：

$$\int_{g^{-1}([a,b])} f(x)\,dx=\int_{t=a}^{b}\left(\sum_{x\in g^{-1}(t)}\frac{h(x)}{|g'(x)|}\right)dt$$

其中，$g^{-1}(t)$ 为所有满足 $g(x)=t$ 的点集，$f(x)$ 为待积分函数。

设 X 是一个随机变量，令 $Y=g(X)$，那么新变量 Y 的概率密度满足

$$f_Y(y)=\sum_{x\in g^{-1}(y)}\frac{f_X(x)}{|g'(x)|}$$

更进一步，由于 g 一阶连续可导，且 $g'(x)\neq 0$，我们可以得到 $g'(x)$ 始终是同号的，也就是说 g 是严格单调的。这时 $g^{-1}(y)$ 实际上只可能是单点集或者空集。

于是我们的式子就变成了：

$$f_Y(y)=\begin{cases}\left.\frac{f_X(x)}{|g'(x)|}\right|_{x=g^{-1}(y)}=\frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|}, & y\in Im(g)\\ 0, & \text{其他}\end{cases}$$

其中 $Im(g)$ 是 g 的值域，也就是 Y 的取值范围；$g^{-1}(y):Im(g)\to\mathbb{R}$ 在此表示严格意义的反函数。

这就是前面没有证明的一维随机变量函数的变换公式，与高维的公式具有一致的结构，只是此时纤维退化为点。

用这个公式重新阐释加法和乘法公式也会是更简单的，不过我们就不在这里写出了。后面我们还会在 χ^2 分布重新利用这个公式。

3.4.3 min,max

最大值与最小值也是经典的随机变量函数，但它也确实是我们刚才的公式所无法处理的。

设 $X_1, X_2, \ldots, X_n$ 为 n 个随机变量, 联合分布函数为 $F(x_1, x_2, \ldots, x_n)$。定义:

$$M = \max(X_1, X_2, \ldots, X_n), \quad m = \min(X_1, X_2, \ldots, X_n).$$

$M \le z$ 等价于 $\forall 1 \le i \le n, X_i \le z$, 因此最大值的累积分布函数为:

$$F_M(z) = P(M \le z) = P(X_1 \le z, X_2 \le z, \ldots, X_n \le z) = F(z, z, \ldots, z).$$

若 $X_1, X_2, \ldots, X_n$ 相互独立, 且各自分布函数为 $F_i(x)$, 则

$$F_M(z) = \prod_{i=1}^{n} F_i(z).$$

$m > z$ 等价于 $\forall 1 \le i \le n, X_i > z$, 于是最小值的累积分布函数为:

$$F_m(z) = P(m \le z) = 1 - P(m > z) = 1 - P(X_1 > z, X_2 > z, \ldots, X_n > z).$$

在独立情形下, 则有

$$1 - F_m(z) = \prod_{i=1}^{n} \big(1 - F_i(z)\big).$$

3.4.4 对独立变量分别变换保持独立性

如果说 $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$ 相互独立, 那么由它们构成的随机向量

$$\mathbf{X} = (X_1, X_2, \ldots, X_n), \quad \mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$$

也是相互独立的。这意味着对任意可测集合 $A \subseteq \mathbb{R}^n$ 和 $B \subseteq \mathbb{R}^m$, 都有

$$P(\mathbf{X} \in A, \mathbf{Y} \in B) = P(\mathbf{X} \in A) P(\mathbf{Y} \in B).$$

更一般地, 若 $\{X_{ij} : i = 1, \ldots, k; j = 1, \ldots, n_i\}$ 是一族相互独立的随机变量, 那么对任意 k 个（由不相交指标集构成的）随机向量

$$\mathbf{X}_1 = (X_{11}, \ldots, X_{1n_1}), \ \ldots, \ \mathbf{X}_k = (X_{k1}, \ldots, X_{kn_k}),$$

这些随机向量之间也是相互独立的。换句话说, 若原始随机变量全体相互独立, 则任意将它们分组后得到的随机向量组仍然相互独立。

可测变换下的保持性　进一步地，如果对每个 $i=1,\dots,k$，$g_i:\mathbb{R}^{n_i}\to\mathbb{R}^{d_i}$ 是一个可测函数，则变换后的随机向量

$$g_1(\mathbf{X}_1),g_2(\mathbf{X}_2),\dots,g_k(\mathbf{X}_k)$$

仍然是相互独立的。换句话说，对相互独立的随机向量分别施加可测变换，不会破坏它们的独立性。

证明. 由于 g_i 可测，$g_i^{-1}(B_i)\subseteq\mathbb{R}^{n_i}$ 可测。由 $\mathbf{X}_1,\dots,\mathbf{X}_k$ 的独立性，

$$\begin{aligned}P\left(\bigcap_{i=1}^{k}\{g_i(\mathbf{X}_i)\in B_i\}\right)&=P\left(\bigcap_{i=1}^{k}\{\mathbf{X}_i\in g_i^{-1}(B_i)\}\right)\\&=\prod_{i=1}^{k}P(\mathbf{X}_i\in g_i^{-1}(B_i))\\&=\prod_{i=1}^{k}P(g_i(\mathbf{X}_i)\in B_i).\end{aligned}$$

故变换后的向量仍保持独立。 □

3.4.5　一阶连续可微可逆无临界点变换的概率密度

如果我们不仅仅讨论 $\mathbb{R}^n\to\mathbb{R}$ 的多元随机变量函数，入手更多元的情况，我们一般没有特别好的结论，除了在满足条件的情况下对各分量应用余面积公式的推论。不过有一个比那个结论更弱一点，但是简单一些的结论：

定理 2. 设 $\mathbf{X}=(X_1,X_2,\dots,X_n)^T$ 是 n 维随机向量，具有联合概率密度函数 $f_{\mathbf{X}}(\mathbf{x})$。设 $\mathbf{g}:\mathbb{R}^n\to\mathbb{R}^n$ 为可逆变换，满足：

1. $\mathbf{g}$ 是连续可微的（C^1 函数）；
2. $\mathbf{g}$ 是可逆；
3. *Jacobi* 矩阵 $J_{\mathbf{g}}(\mathbf{x})=\left[\frac{\partial g_i}{\partial x_j}(\mathbf{x})\right]_{i,j=1}^{n}$ 的行列式 $\det J_{\mathbf{g}}(\mathbf{x})\neq 0$（几乎处处成立）。

则 $\mathbf{Y}$ 的概率密度函数为

$$f_{\mathbf{Y}}(\mathbf{y})=\frac{f_{\mathbf{X}}\big(\mathbf{g}^{-1}(\mathbf{y})\big)}{|\det J_{\mathbf{g}}(\mathbf{g}^{-1}(\mathbf{y}))|},\quad \mathbf{y}\in\mathbf{g}(\mathbb{R}^n).$$

我们后面在随机变量的数值特征这一部分应用并扩展这里的结论。

4 随机变量的数字特征

在应用中，我们往往感兴趣于某些能描述随机变量特征的常数。这种常数被称为随机变量的**数字特征**。对于随机变量，一些经典的数值特征是期望（均值），方差，标准差，协方差，矩。

4.1 均值

对任意随机变量 $X : \Omega \to \mathbb{R}$，定义其**均值**为：

$$E(X) = \int_{\Omega} X(\omega)\, dP(\omega).$$

这是个从测度角度的定义，我们也可以给出简单一点的连续型随机变量 X 的均值定义为：

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$

离散型就是把积分改为求和，就是换了个测度。注意到我们有处理无限积分/和的情况，所以有时均值不一定存在。

它的意义基本就是一个按概率的加权平均数。

对于一个存在均值的随机变量 X，我们把 $X - E(X)$ 叫做**中心化**的随机变量

均值是对于单个随机变量而言的，不过对于随机变量函数，我们也可以计算其均值。为了方便讨论，我们只研究连续随机变量函数。

定理 3. *设 $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ 是 n 维连续型随机向量，其联合概率密度函数为 $f_{\mathbf{X}}(\mathbf{x})$。设 $g : \mathbb{R}^n \to \mathbb{R}$ 是一个可测函数，定义新的随机变量 $Y = g(\mathbf{X})$。*

$$E(g(\mathbf{X})) = \int_{\mathbb{R}^n} g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$$

证明. 由定义

$$E(g(\mathbf{X})) = \int_{\Omega} g(\mathbf{X}(\omega)) dP(\omega)$$

由于 g 可测，应用积分变换：

$$\int_{\Omega} g(\mathbf{X}(\omega)) dP(\omega) = \int_{\mathbb{R}^n} g(\mathbf{x}) dP_{\mathbf{X}}(\mathbf{x})$$

再应用概率密度的定义：

$$\int_{\mathbb{R}^n} g(\mathbf{x}) dP_{\mathbf{X}}(\mathbf{x}) = \int_{\mathbb{R}^n} g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$$

综上即得

$$E(g(\mathbf{X})) = \int_{\mathbb{R}^n} g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$$

□

一个直观的理解就是我们可以直接把这个函数拿到积分内。对于线性函数来说，应用定理得到

$$\begin{aligned} E\left(\sum_{i=1}^{n} a_i X_i + b\right) &= \int_{\mathbb{R}^n} \left(\sum_{i=1}^{n} a_i x_i + b\right) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\ &= \left(\sum_{i=1}^{n} \int_{\mathbb{R}^n} a_i x_i f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}\right) + \int_{\mathbb{R}^n} b f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\ &= \sum_{i=1}^{n} a_i E\left(X_i\right) + b \end{aligned}$$

这就是期望的线性性。它还告诉我们：常数的均值是常数。注意线性性这一事实并不依赖于随机变量的任何独立性。

随机变量的独立性则体现在它们乘积的期望上。

$$\begin{aligned} E(XY) &= \int_{\mathbb{R}^2} xy f_{XY}(x,y) dx dy \\ E(X)E(Y) &= \left(\int_{\mathbb{R}} x f_X(x) dx\right)\left(\int_{\mathbb{R}} y f_Y(y) dy\right) \\ &= \int_{\mathbb{R}^2} xy f_X(x) f_Y(y) dx dy \end{aligned}$$

可见当 X 与 Y 独立时，上下两个式子相等。注意，反过来是不一定成立的。$E(X)E(Y) = E(XY)$ 只能得出 $f_X(x)f_Y(y) = f_{XY}(x,y)$ **几乎处处**相等。

4.2 方差

随机变量 X 的**方差**定义为 $E((X - E(X))^2)$。

连续型随机变量 X 的**方差**定义为：

$$D(X) = \int_{-\infty}^{\infty} (x - E(X)) f_X(x) dx$$

离散型就是把积分改为求和。注意到我们有处理无限积分/和的情况，所以有时方差也可能不存在。

它描述了随机变量的分布偏离均值的情况，并且总是非负的。

由于对于常数 C，$E(C) = C$，我们可以得到 $D(C) = 0$。这一结论反过来也**几乎**是成立的，但我们把证明留在弱大数定律的部分，我们先放心使用这个结论。请放心，弱大数定律的证明不会依赖这中间的任何东西。

对于一般的随机变量函数，它的方差没有什么太好的结论，只有在应用线性函数时才有比较好的结论。不过为了叙述的简洁性，我们把它的性质留到协方差这一部分来介绍。

标准差就是方差的平方根。

4.3 协方差

对于两个随机变量 X，Y，它们的协方差定义为

$$Cov(X, Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$$

按式子就有：$D(X) = Cov(X, X) = E(X^2) - (E(X))^2$。我们还可以验证这一表达式的对称性和双线性性：

$$Cov(X, Y) = Cov(Y, X)$$

$$Cov(\alpha X_1 + \beta X_2, Y) = \alpha Cov(X_1, Y) + \beta Cov(X_2, Y)$$

同时还有平移不变性 $Cov(X+C, Y) = Cov(X, Y)$，和非负性 $Cov(X, X) \geq 0$。

这些性质代入到方差中，就可以知道：

$$D(\lambda X) = \lambda^2 D(X), \quad D(X + C) = D(X)$$

所有这些性质的证明基本都依赖于均值的线性性质。

在适当的函数空间上，我们可以把协方差看作一个内积。如果我们把两个相差一个**几乎处处**常值随机变量的随机变量视为等价的（相当于我们在

处理中心化后的随机变量），那么协方差就有**正定性**了，也就构成一个实内积了。标准差于是就是它诱导的范数。

相应地，我们可以给出内积经典结论。

首先是勾股定理

$$\begin{aligned} D(X+Y) &= Cov(X+Y, X+Y) \\ &= Cov(X,X) + Cov(Y,Y) + 2Cov(X,Y) \\ &= D(X) + D(Y) + 2Cov(X,Y) \end{aligned}$$

于是 $D(X+Y) = D(X) + D(Y)$ 当且仅当 $Cov(X,Y) = 0$。

这种“协方差正交性”我们称为**不相关性**。

然后是柯西不等式：

$$Cov(X,Y)^2 \leq D(X)D(Y)$$

等号成立当且仅当一个是另一个的标量倍。注意这里我们是在零均值意义下讨论的，所以它们还可能相差一个常数。不过总归是 $Y = a + bX$ 这种线性关系就是了。

这里很自然地，这里就有一个类似余弦的量：

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X)D(Y)}} \leq 1$$

这就称为**相关系数**。反过来协方差也可以表示成 $\rho_{XY}\sqrt{D(X)D(Y)}$。

由前面的式子可以知道，X, Y 独立时有 $Cov(X,Y) = E(XY) - E(X)E(Y) = 0$。也就是说：**独立性蕴含不相关性**。但不相关的变量不一定独立。我们只能通过换位得到并非不相关的变量不是独立的。

对于一组随机变量 $\mathbf{X} = (X_1, \ldots, X_n)^T$，我们也可以给出一个类似 Gram 矩阵的构造：协方差矩阵。

我们称矩阵 $(C)_{ij} = Cov(X_i, X_j)$ 为它们的协方差矩阵，也记作 $Cov(\mathbf{X})$。这是个对称半正定矩阵，我们可以轻松用谱定理分解它。如果我们丢掉小特征值对应的部分，就能对数据进行降维分析。如果对 $\mathbf{X}$ 做一个线性变换得到 $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$，那么有 $Cov(\mathbf{Y}) = ACov(\mathbf{X})A^T$

4.4 矩

矩是一类更一般的数字特征，实际上上面的数字特征都属于特殊的矩。矩作为一类更一般的数字特征，统一描述了随机变量的位置、离散度、形状

等信息。

对单个随机变量 X，它的 k 阶原点矩 μ'_k 定义为 $E(X^k)$。均值就是 1 阶原点矩。

矩的一个好处是，随机变量的任意多项式函数的均值都可以表示成矩的线性函数。

k 阶中心矩 μ_k 定义为 $E((X - E(X))^k)$。方差就是 2 阶中心矩。

k 阶标准化矩 $\tilde{\mu}_k$ 定义为 $E\left(\left(\frac{X-\mu}{\sigma}\right)^k\right)$。3 阶标准化矩叫做偏度，4 阶标准化矩叫做峰度。

对随机变量 X 与 Y，其 (m, n) 阶或者 $m + n$ 阶混合矩定义为：

$$\begin{aligned} &\text{混合原点矩：} \quad \mu'_{mn} = E(X^m Y^n) \\ &\text{混合中心矩：} \quad \mu_{mn} = E\left[(X - \mu_X)^m (Y - \mu_Y)^n\right] \end{aligned}$$

当 $m = 1, n = 1$ 时，2 阶混合中心矩 $\mu_{11} = \mathrm{Cov}(X, Y)$ 即协方差。协方差矩阵中的诸元素就是诸 2 阶中心矩。

它真正强大的应用在矩母函数和矩估计处。

5 经典随机分布

我把这一部分放到很后面是为了在这里把它们的性质一起整合了。

5.1 离散分布

掌握一些组合恒等式和级数恒等式对处理离散分布非常有益。

5.1.1 两点分布

这是最简单的离散分布了。为了方便，我们常让变量只可能取值 $0,1$。它有一个参数 $0 \le p \le 1$，它的分布律是

$$\begin{aligned} &P(X=0) = 1-p \\ &P(X=1) = p \\ &P(X=k) = 0 \qquad\qquad k \notin \{0,1\} \end{aligned}$$

很简单地，$E(X)=p$，$D(X)=p(1-p)$。一个服从这个分布的随机试验叫做伯努利试验。

5.1.2 离散均匀分布

这是一种各个取值等概率的离散概率分布。设 $A=\{a_1,\dots,a_n\}$ 是一个有限集合，定义 X 服从离散均匀分布为

$$P(X=a) = \begin{cases} \frac{1}{n} & a \in A \\ 0 \end{cases}$$

可知 $E(X)=\frac{1}{n}\sum_{i=1}^{n} a_i = \overline{a}$，$D(X)=\frac{1}{n}\sum_{i=1}^{n}(a_i-\overline{a})^2$。

5.1.3 Bernoulli（二项）分布

若相互独立的随机变量 $X_1,\dots,X_n$ 均服从参数为定值 p 的两点分布，则称 $X=\sum_{i=1}^{n} X_i$ 服从二项分布，记作 $X \sim b(n,p)$。我们也把它理解为一个 n 重伯努利试验的成功次数。$n=1$ 的时候它也就退化成了两点分布。

很显然，按照定义，如果 $X \sim b(n_1, p), Y \sim b(n_2, p)$ 且它们相互独立，则有 $X + Y \sim b(n_1 + n_2, p)$。当然我们也可以按下面的分布律用卷积去验证。

其分布律为:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

这个表达式的关键在其二项式系数。我这里采取一种广义一些的定义，令 $\binom{n}{k}$ 在 $k > n$ 或 $k < 0$ 时为 0。

它的均值:

$$\begin{aligned}
E(X) &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} n \binom{n-1}{k-1} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} n \binom{n-1}{k-1} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n-1} n \binom{n-1}{k} p^{k+1} (1-p)^{n-1-k} \\
&= np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-1-k} \\
&= np
\end{aligned}$$

当然我们也可以用 n 个独立的两点分布之和来计算，结果就是立即的。方

差:

$$
\begin{aligned}
D(X) &= E(X^2) - (E(X))^2 \\
&= \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} - n^2 p^2 \\
&= n \sum_{k=0}^{n} k \binom{n-1}{k-1} p^k (1-p)^{n-k} - n^2 p^2 \\
&= np \sum_{k=0}^{n-1} (k+1) \binom{n-1}{k} p^k (1-p)^{n-1-k} - n^2 p^2 \\
&= np \sum_{k=0}^{n-1} k \binom{n-1}{k} p^k (1-p)^{n-1-k} + np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-1-k} - n^2 p^2 \\
&= np(n-1)p + np - n^2 p^2 = np(1-p)
\end{aligned}
$$

5.1.4 多项分布

如其名字所说，这是一个和多项式系数密切相关的分布。它可以被视作一个有多种可能结果的试验独立重复 n 次后各结果出现的次数。

记 k 元随机向量 $\mathbf{X}$ 服从多项分布为 $\mathbf{X} \sim Mult(n, k, p_1, \ldots, p_k)$，其中 $\sum_{i=1}^{k} p_i = 1$。

它的分布律为

$$
P(\mathbf{X} = (x_1, \ldots, x_k)) = \binom{n}{x_1 \cdots x_k} \prod_{i=1}^{k} p_i^{x_i}
$$

我这里采取一种广义一些的定义，令 $\binom{n}{x_1 \cdots x_k}$ 在 $\sum_{i=1}^{k} x_i \neq n$ 时为 0。

它的第 i 个分量的边缘分布显然符合二项分布 $b(n, p_i)$，而且还不止，把它的每一个分量加在一起，我们会得到 $\sum_{i=1}^{n} X_i = n$，做 n 次实验就会得到 n 次结果，不管结果是什么。

随便拿出来两个分量 X_i，X_j，有

$$
Cov(X_i, X_j) = \delta_{ij} p_i - p_i p_j
$$

证明就是组合恒等式，容我在此略过。

5.1.5 Poisson 分布

泊松分布是服从这样分布律的分布:

$$P(X=k)=\frac{\lambda^k e^{-\lambda}}{k!},\quad k\in\mathbb{N}$$

我们记作 $X\sim\pi(\lambda)$，其中 $\lambda>0$。它实际上是作为一族特殊的二项分布的极限而出现的。

设 $X\sim b(n,p)$，且假设 $n\to\infty$ 时 $np\to\lambda$。我们有：

$$\begin{aligned}
\lim_{n\to\infty}P(X=k)&=\lim_{n\to\infty}\binom{n}{k}p^k(1-p)^{n-k}\\
&=\lim_{n\to\infty}\frac{\prod_{i=0}^{k-1}(n-i)}{k!}p^k(1-p)^{n-k}\\
&=\lim_{n\to\infty}\frac{\prod_{i=0}^{k-1}(n-i)}{k!}p^k\left(1-\frac{\lambda}{n}\right)^{n-k}\\
&=\lim_{n\to\infty}\prod_{i=0}^{k-1}\left(1-\frac{i}{n}\right)\cdot\frac{\lambda^k}{k!}\cdot\left(1-\frac{\lambda}{n}\right)^{n-k}\\
&=\frac{\lambda^k}{k!}\cdot\lim_{n\to\infty}\prod_{i=0}^{k-1}\left(1-\frac{i}{n}\right)\cdot\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n-k}\\
&=\frac{\lambda^k}{k!}\cdot 1\cdot e^{-\lambda}\\
&=\frac{\lambda^k e^{-\lambda}}{k!}
\end{aligned}$$

当然我们让 $p\to 0$ 的效果也是一样的。

泊松分布常用于描述单位时间内随机事件发生的次数，λ 表示该事件的平均发生率。上面这一事实可以理解为把时间无限细分，每段最多发生一次事件。同一件事情的另一面可以由指数分布描述。这一定理同时意味着当 n 足够大时，我们也可以用泊松分布当作二项分布的近似。

泊松分布的均值：

$$\begin{aligned}
E(X)&=\sum_{k=0}^{\infty}k\frac{\lambda^k e^{-\lambda}}{k!}\\
&=\lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}e^{-\lambda}}{(k-1)!}\\
&=\lambda\sum_{k=0}^{\infty}\frac{\lambda^k e^{-\lambda}}{k!}\\
&=\lambda
\end{aligned}$$

方差：

$$\begin{aligned}
D(X) &= E(X^2) - (E(X))^2 \\
&= \sum_{k=0}^{\infty} k^2 \frac{\lambda^k e^{-\lambda}}{k!} - \lambda^2 \\
&= \lambda \sum_{k=1}^{\infty} k \frac{\lambda^{k-1} e^{-\lambda}}{(k-1)!} - \lambda^2 \\
&= \lambda \sum_{k=0}^{\infty} (k+1) \frac{\lambda^k e^{-\lambda}}{k!} - \lambda^2 \\
&= \lambda \sum_{k=0}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} + \lambda \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} - \lambda^2 \\
&= \lambda
\end{aligned}$$

泊松分布具有可加性，即几个独立的服从泊松的随机变量之和仍然服从泊松分布。这里证明两个随机变量的情况，多个随机变量容易归纳。设 $X \sim \pi(\lambda_1), Y \sim \pi(\lambda_2), Z = X + Y$，$X, Y$ 独立：

$$\begin{aligned}
P(Z = m) &= \sum_{k=0}^{m} P(X = k, Y = m - k) \\
&= \sum_{k=0}^{m} P(X = k) P(Y = m - k) \\
&= \sum_{k=0}^{m} \frac{\lambda_1^k e^{-\lambda_1}}{k!} \frac{\lambda_2^{m-k} e^{-\lambda_2}}{(m-k)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{m!} \sum_{k=0}^{m} \binom{m}{k} \lambda_1^k \lambda_2^{m-k} \\
&= \frac{(\lambda_1 + \lambda_2)^k e^{-(\lambda_1+\lambda_2)}}{m!}
\end{aligned}$$

于是 $Z \sim \pi(\lambda_1 + \lambda_2)$。它不光服从泊松分布，其参数还是二者的和。

泊松分布还有这样的稀释性：令 $X \sim \pi(\lambda)$，若条件分布 $Y|X = n \sim b(n, p)$，则 $Y \sim \pi(\lambda p)$。换句话说，就是每个事件以相同的独立的概率 p 导致另一事件发生，那么这一事件发生的次数 $Y \sim \pi(\lambda p)$。证明如下

$$
\begin{aligned}
P(Y=k) &= \sum_{n=k}^{\infty} P(Y=k|X=n)P(X=n) \\
&= \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} \frac{\lambda^n e^{-\lambda}}{n!} \\
&= \sum_{n=k}^{\infty} p^k (1-p)^{n-k} \frac{\lambda^k \lambda^{n-k} e^{-\lambda}}{k!(n-k)!} \\
&= \frac{(\lambda p)^k e^{-\lambda}}{k!} \sum_{n=k}^{\infty} \frac{(\lambda(1-p))^{n-k}}{(n-k)!} \\
&= \frac{(\lambda p)^k e^{-\lambda}}{k!} \sum_{n=0}^{\infty} \frac{(\lambda(1-p))^{n}}{n!} \\
&= \frac{(\lambda p)^k e^{-\lambda}}{k!} e^{\lambda(1-p)} \\
&= \frac{(\lambda p)^k e^{-\lambda p}}{k!}
\end{aligned}
$$

5.1.6 几何分布

几何分布的常见定义有两种，在这里我们采用这种定义：一系列独立的伯努利试验中，首次成功所需的试验次数服从几何分布。另一种定义是成功首次前的失败次数，很显然这种定义要比前面的定义少 1。我们记前面的总次数 $X \sim G(p)$。

很容易知道，

$$
P(X=k) = p(1-p)^{k-1}, \quad k = 1, 2, \cdots
$$

几何分布因几何级数而得名。几何分布的均值：

$$
\begin{aligned}
E(X) &= \sum_{k=1}^{\infty} kp(1-p)^{k-1} \\
&= p \sum_{k=0}^{\infty} (k+1)(1-p)^k \\
&= p \frac{1}{(1-(1-p))^2} = \frac{1}{p}
\end{aligned}
$$

方差:

$$
\begin{aligned}
D(X) &= E(X^2) - (E(X))^2 \\
&= \sum_{k=1}^{\infty} k^2 p(1-p)^{k-1} - \frac{1}{p^2} \\
&= \sum_{k=0}^{\infty} (k+1)^2 p(1-p)^k - \frac{1}{p^2} \\
&= \sum_{k=0}^{\infty} k^2 p(1-p)^k + 2\sum_{k=0}^{\infty} kp(1-p)^k + \sum_{k=0}^{\infty} p(1-p)^k - \frac{1}{p^2} \\
&= \sum_{k=0}^{\infty} k^2 p(1-p)^k + 2\sum_{k=1}^{\infty} (k-1)p(1-p)^{k-1} + \sum_{k=1}^{\infty} p(1-p)^{k-1} - \frac{1}{p^2} \\
&= p\sum_{k=0}^{\infty} k^2 (1-p)^k - \frac{1}{p^2} + \frac{2}{p} - 1 \\
&= p\frac{(1-p)(1+(1-p))}{p^3} + -\frac{1}{p^2} + \frac{2}{p} - 1 \\
&= \frac{1-p}{p^2}
\end{aligned}
$$

它和连续情况下的指数分布有很大的相似性。它是唯一在非负整数下具有无记忆性的离散分布。无记忆性是指:

$$P(X > m+n | X > n) = P(X > m)$$

已知前面 n 次没有成功，再进行 m 次还没成功的概率和从头开始做 m 次试验还没成功的概率相同。或者等价地，

$$P(X = m+n | X > n) = P(X = m)$$

简单计算一下 $P(X > k)$ 就能知道:

$$
\begin{aligned}
P(X > k) &= \sum_{n=k+1}^{\infty} p(1-p)^{n-1} \\
&= \sum_{n=0}^{\infty} p(1-p)^{n+k} \\
&= p(1-p)^k \sum_{n=0}^{\infty} (1-p)^n = (1-p)^k
\end{aligned}
$$

5.1.7 负二项分布

前面的几何分布是二项分布的一个特例。如果我们把刚才的问题扩展到试验恰好 r 次成功时的总次数，我们就的到了负二项分布。我们记次数 $X \sim nb(r,p)$。由定义很明显，$G(p) = nb(1,p)$。当然，也有一种替代的定义，就是令变量为恰好成功 r 次之前的失败次数，这也很简单地比前面的定义要小一个 r。

那么它的分布律应该是：

$$P(X=k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \quad k = r, r+1, \cdots$$

负二项就得名于此。负二项分布可用于过度离散（方差大于均值）的计数数据建模，是泊松分布的一种推广。取极限之后，**失败次数**形式的负二项分布也会像二项分布一样趋近一个泊松分布。

设 $Y + r \sim nb\,(r, 1-p)$，且假设 $r \to \infty$ 时 $r(1-p) \to \lambda$ 。则有：

$$\begin{aligned}
\lim_{r\to\infty} P(Y=y) &= \lim_{r\to\infty} \binom{r+y-1}{y} p^r (1-p)^y \\
&= \lim_{r\to\infty} \frac{\prod_{i=0}^{y-1}(r+i)}{y!} \left(1-\frac{\lambda}{r}\right)^r \left(\frac{\lambda}{r}\right)^y \\
&= \lim_{r\to\infty} \frac{r^y}{y!} \prod_{i=0}^{y-1} \left(1+\frac{i}{r}\right) \left(1-\frac{\lambda}{r}\right)^r \frac{\lambda^y}{r^y} \\
&= \frac{\lambda^y}{y!} \lim_{r\to\infty} \prod_{i=0}^{y-1} \left(1+\frac{i}{r}\right) \cdot \lim_{r\to\infty} \left(1-\frac{\lambda}{r}\right)^r \\
&= \frac{\lambda^y e^{-\lambda}}{y!}.
\end{aligned}$$

当然我们让 $p \to 1$ 也有同样的结论。

另一种看法是，由于几何分布的无记忆性，我们可以把它视为 r 个独立的服从 $G(p)$ 的随机变量之和。它很清晰地就具有可加性了：若 $X_1 \sim nb(r_1,p), X_2 \sim nb(r_2,p)$ 且相互独立，则 $X_1 + X_2 \sim nb(r_1+r_2,p)$。

于是就不难得到 $E(X) = \frac{r}{p}$，$D(X) = \frac{r(1-p)}{p^2}$。

5.1.8 超几何分布

超几何分布描述的是不放回抽样中抽到特定类别物品数量的概率分布。

设有总数为 N 的总体，其中包含 M 个“成功”物品，其余 $N-M$ 个为“失败”物品。从中不放回地随机抽取 n 个物品，$n \leq N$。令随机变量 X 表示在这 n 个物品中“成功”物品的数量，则 X 服从超几何分布，记作 $X \sim H(n, M, N)$。

它的分布律就是：

$$P(X=k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

超几何分布因超几何级数得名。

$$\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} = \frac{(-n)^{\overline{k}}(-M)^{\overline{k}}}{k!(N-M-n+1)^{\overline{k}}} \frac{\binom{N-M}{n}}{\binom{N}{n}}$$

这里的 $x^{\overline{k}}$ 指的是 k 次上升幂，也就是 $x(x+1)\cdots(x+k-1)$。这就是超几何级数中的一项。

从直觉上来说，当总体容量 N 很大时，无放回抽样可以近似为有放回抽样，事情也确实是这样的。设 $X \sim H(n, M, N)$，当 $N \to \infty$ 时，若 $\frac{M}{N} \to p$，则有：

$$\begin{aligned}
\lim_{N\to\infty} P(X=k) &= \lim_{N\to\infty} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \\
&= \lim_{N\to\infty} \frac{\frac{1}{k!}\prod_{i=0}^{k-1}(M-i) \cdot \frac{1}{(n-k)!}\prod_{j=0}^{n-k-1}(N-M-j)}{\frac{1}{n!}\prod_{l=0}^{n-1}(N-l)} \\
&= \binom{n}{k} \lim_{N\to\infty} \frac{\prod_{i=0}^{k-1}(M-i) \cdot \prod_{j=0}^{n-k-1}(N-M-j)}{\prod_{l=0}^{n-1}(N-l)} \\
&= \binom{n}{k} \lim_{N\to\infty} \left[\prod_{i=0}^{k-1} \frac{M-i}{N-i} \cdot \prod_{j=0}^{n-k-1} \frac{N-M-j}{N-k-j}\right] \\
&= \binom{n}{k} \prod_{i=0}^{k-1} \lim_{N\to\infty} \frac{M-i}{N-i} \cdot \prod_{j=0}^{n-k-1} \lim_{N\to\infty} \frac{N-M-j}{N-k-j} \\
&= \binom{n}{k} \prod_{i=0}^{k-1} \frac{M}{N} \cdot \prod_{j=0}^{n-k-1} \left(1-\frac{M}{N}\right) \\
&= \binom{n}{k} \left(\frac{M}{N}\right)^k \left(1-\frac{M}{N}\right)^{n-k}.
\end{aligned}$$

它确实逼近一个二项分布，当然我也可以进一步让其逼近泊松分布。类

似二项分布和多项分布，我也可以将其推广为多项超几何分布，这里就不再展开了。

它的均值与有放回的情况惊人地一致：

$$\begin{aligned} E(X) &= \sum_k k \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \\ &= M \sum_k \frac{\binom{M-1}{k-1}\binom{N-M}{n-k}}{\binom{N}{n}} \\ &= M \frac{\binom{N-1}{n-1}}{\binom{N}{n}} \\ &= n\frac{M}{N} \end{aligned}$$

而方差则相差一个因子：

$$\begin{aligned} D(X) &= E(X^2) - (E(X))^2 \\ &= \sum_k k^2 \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} - \frac{n^2M^2}{N^2} \\ &= \sum_k k(k-1) \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} + \sum_k k \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} - \frac{n^2M^2}{N^2} \\ &= M(M-1) \sum_k \frac{\binom{M-2}{k-2}\binom{N-M}{n-k}}{\binom{N}{n}} + \frac{nM}{N} - \frac{n^2M^2}{N^2} \\ &= M(M-1) \frac{\binom{N-2}{n-2}}{\binom{N}{n}} + \frac{nM}{N} - \frac{n^2M^2}{N^2} \\ &= \frac{n(n-1)M(M-1)}{N(N-1)} + \frac{nM}{N} - \frac{n^2M^2}{N^2} \\ &= n\frac{M}{N}(1 - \frac{M}{N})\frac{N-n}{N-1} \end{aligned}$$

这个因子 $\frac{N-n}{N-1}$ 也被称为有限总体校正因子。

5.2 连续分布

和处理离散分布一样，掌握一些积分恒等式是有益的。善用 Γ 函数可以让生活大部分变得更美好。

关于 Γ 函数，我们需要了解的基本就是下面这几件事：

定义:

$$\Gamma(s) = \int_0^{+\infty} t^{s-1} e^{-t} dt$$

递推性质:

$$\begin{aligned}
\Gamma(s+1) &= \int_0^{+\infty} t^s e^{-t} dt \\
&= -\int_0^{+\infty} t^s d(e^{-t}) \\
&= -\left[t^s e^{-t}\right]_0^{+\infty} + \int_0^{+\infty} e^{-t} d(t^s) \\
&= s\int_0^{+\infty} t^{s-1} e^{-t} dt = s\Gamma(s)
\end{aligned}$$

余割恒等式

$$\Gamma(s)\Gamma(1-s) = \frac{\pi}{\sin \pi s}$$

特殊值

$$\Gamma(1) = 1, \quad \Gamma(\frac{1}{2}) = \sqrt{\pi}$$

进一步应用递推性质，就有:

$$\Gamma(n+1) = n!, \quad \Gamma\left(n + \frac{1}{2}\right) = \frac{(2n)!}{4^n n!}\sqrt{\pi}, \quad n \in \mathbb{N}$$

实际上最初 Γ 函数就是作为阶乘的解析延拓而出现的。

作为补充，Beta 函数也有大用。Beta 函数定义为

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 t^{a-1}(1-t)^{b-1} dt$$

这里并不打算给出这两个定义等价的证明。

5.2.1 均匀分布

毫无疑问，连续分布中最简单的就是均匀分布了。均匀分布描述了一个随机变量在某个区间内取值的概率处处相等的分布。随机变量 X 服从区间 $[a,b]$ 上的均匀分布记作 $X \sim U(a,b)$。其概率密度函数为:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b, \\ 0 & \end{cases}$$

几乎没有任何可说的，不过需要注意的是，一定不能忘记随机变量在区间 $[a,b]$ 外取值为 0 这一事实。

它的累积分布函数是

$$F_X(x)=\begin{cases}0 & x<a\\ \frac{x-a}{b-a} & a\le x\le b\\ 0 & x>b\end{cases}$$

它的均值为 $\frac{a+b}{2}$，方差为 $\frac{(b-a)^2}{12}$。

它值得称道的一点是，可以通过对 $U(0,1)$ 进行变换来得到任意分布的随机变量。令 X 为随机变量，F_X 为其累积分布函数。如果 F_X 严格单调递增（也就是连续且可逆），则 $F_X(X)\sim U(0,1)$。反过来，若 $U\sim U(0,1)$，$F_X^{-1}(U)$ 服从原分布 F_X。其证明是浪费笔墨的。利用它可以通过均匀分布生成几乎任意分布的随机数。

5.2.2 指数分布

指数分布非常类似离散的几何分布，就像伯努利试验和泊松过程的类比。随机变量 X 服从指数分布记作 $X\sim Exp(\theta)$。

指数分布的概率密度函数是

$$f_X(x)=\begin{cases}\frac{1}{\theta}e^{-\frac{x}{\theta}} & x\ge 0\\ 0 & x<0\end{cases}$$

累积分布函数是

$$F_X(x)=\begin{cases}1-e^{-\frac{x}{\theta}} & x\ge 0\\ 0 & x<0\end{cases}$$

注意，有另一种等价的定义是 $\lambda e^{-\lambda x}$ 的形式，它和我们的定义只相差一个倒数关系，不过我们不在这里使用这种定义。

同样，它也是唯一具有无记忆性的连续分布。对于 $t_1,t_2\ge 0$，

$$P(X>t_1+t_2|X>t_1)=P(X>t_2)$$

当然这也是因为

$$P(X>t)=\int_t^{+\infty}\frac{1}{\theta}e^{-\frac{x}{\theta}}dx=e^{-\frac{t}{\theta}},\quad t\ge 0$$

它的均值

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} x f_X(x) dx \\ &= \int_0^{+\infty} \frac{x}{\theta} e^{-\frac{x}{\theta}} dx \\ &= \theta \int_0^{+\infty} u e^{-u} du \qquad\qquad u = \frac{x}{\theta} \\ &= \theta \Gamma(2) = \theta \end{aligned}$$

方差

$$\begin{aligned} D(X) &= E(X^2) - (E(X))^2 \\ &= \int_{-\infty}^{+\infty} x^2 f_X(x) dx - \theta^2 \\ &= \int_0^{+\infty} \frac{x^2}{\theta} e^{-\frac{x}{\theta}} dx - \theta^2 \\ &= \theta^2 \int_0^{+\infty} u^2 e^{-u} du - \theta^2 \qquad\qquad u = \frac{x}{\theta} \\ &= \theta^2 \Gamma(3) - \theta^2 = \theta^2 \end{aligned}$$

在前面关于的泊松分布的问题中，如果说事件发生的次数服从泊松分布，那么相邻两个事件的间隔服从指数分布，反过来也是一样的，这就是泊松过程。

5.2.3 Erlang 分布

爱尔朗分布是和指数分布密切相关的一种分布。如果相互独立的随机变量 $X_1, \ldots, X_n$ 都服从 $Exp(\theta)$，那么我们称它们的和 $X = \sum_{i=1}^{n} X_i$ 服从爱尔朗分布，记作 $X \sim Erlang(n, \theta)$。

它的概率密度函数是

$$f_X(x) = \begin{cases} \frac{x^{n-1} e^{-\frac{x}{\theta}}}{\theta^n (n-1)!} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

你会发现这个式子和 s 取整数的 Γ 函数几乎相同。它确实是我们后面会提到的 Gamma 分布的特例。指数分布也就是 $n=1$ 的特例。

我们可以用卷积公式归纳验证它的正确性，不过这不是最简单的方法，因为用傅里叶变换做卷积才快。

由定义可以看出它的可加性，这里姑且验证一下，顺带把上面的式子验证了。令 $X \sim Erlang(n_1, \theta)$，$Y \sim Erlang(n_2, \theta)$，$X, Y$ 独立，$Z = X + Y$。那么有：

$$
\begin{aligned}
f_Z(z) &= \int_{-\infty}^{+\infty} f_X(t) f_Y(z-t) dt \\
&= \int_0^z \frac{t^{n_1-1} e^{-\frac{t}{\theta}}}{\theta^{n_1}(n_1-1)!} \frac{(z-t)^{n_2-1} e^{-\frac{z-t}{\theta}}}{\theta^{n_2}(n_2-1)!} dt \\
&= \frac{e^{-\frac{z}{\theta}}}{\theta^{n_1+n_2}(n_1-1)!(n_2-1)!} \int_0^z t^{n_1-1}(z-t)^{n_2-1} dt \\
&= \frac{z^{n_1+n_2-1} e^{-\frac{z}{\theta}}}{\theta^{n_1+n_2}(n_1-1)!(n_2-1)!} \int_0^1 u^{n_1-1}(1-u)^{n_2-1} du \qquad t = zu \\
&= \frac{z^{n_1+n_2-1} e^{-\frac{z}{\theta}}}{\theta^{n_1+n_2}(n_1-1)!(n_2-1)!} B(n_1, n_2) \\
&= \frac{z^{n_1+n_2-1} e^{-\frac{z}{\theta}}}{\theta^{n_1+n_2}(n_1-1)!(n_2-1)!} \frac{(n_1-1)!(n_2-1)!}{(n_1+n_2-1)!} \\
&= \frac{z^{n_1+n_2-1} e^{-\frac{z}{\theta}}}{\theta^{n_1+n_2}(n_1+n_2-1)!}
\end{aligned}
$$

由定义，我们可以知道其均值 $E(X) = n\theta$，方差 $D(X) = n\theta^2$。

它就是独立指数分布的叠加，在强度为 $\frac{1}{\theta}$ 的泊松过程中，第 n 个事件发生的等待时间服从 $Erlang(n, \theta)$。

5.2.4 正态分布

正态分布可谓是概率统计中最重要的分布了。它有非常多良好的性质，中心极限定理更是奠定了其无可动摇的地位。随机变量 X 服从正态分布记为 $X \sim N(\mu, \sigma^2)$。这两个参数分别叫做正态分布的位置参数和尺度参数的平方。

正态分布的概率密度函数是

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

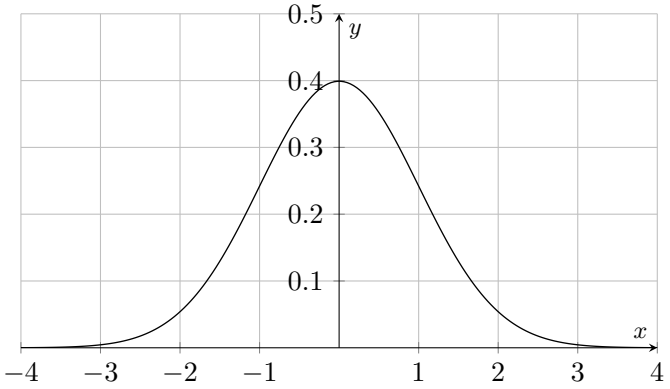
这是一个奇特的式子，它的核心是一个高斯积分。我们不难做一个变量代换 $U = \frac{X-\mu}{\sigma}$，这样就有

$$f_U(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

这就表明 $U \sim N(0,1)$。这种正态分布被称为标准正态分布。由于这种变换的存在，我在下面基本只用讨论标准正态分布了。我们可以反过来根据这一事实得出正态分布的线性变换性质。若 $X \sim N(\mu, \sigma^2)$，则 $a + bX \sim N(a + b\mu, b^2\sigma^2)$。

标准正态分布的图形是这样的：

标准正态分布



正态分布的累积分布函数没有初等的解析式形式，不过姑且有一些不错的性质。我们记标准正态分布的累计分布函数为 $\Phi(x)$，那么有：

$$\begin{aligned}
\Phi(x) &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \\
&= \int_{-x}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du && u = -t \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du - \int_{-\infty}^{-x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \\
&= 1 - \Phi(-x)
\end{aligned}$$

代入 $x = 0$，就有 $\Phi(0) = \frac{1}{2}$。

虽然没有初等形式，但是我们有一个级数表达式：

$$\begin{aligned}\Phi(x) &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \\ &= \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt + \int_{0}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \\ &= \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{-\frac{t^2}{2}} dt \\ &= \frac{1}{2} + \frac{1}{2\sqrt{\pi}} \int_{0}^{\frac{x^2}{2}} u^{-\frac{1}{2}} e^{-u} du \qquad\qquad u = \frac{t^2}{2}\end{aligned}$$

右边很像 Γ 函数，但是被不幸截断了！不过对 e^{-u} 做级数展开就好了：

$$\begin{aligned}&= \frac{1}{2} + \frac{1}{2\sqrt{\pi}} \int_{0}^{\frac{x^2}{2}} u^{-\frac{1}{2}} e^{-u} du \\ &= \frac{1}{2} + \frac{1}{2\sqrt{\pi}} \int_{0}^{\frac{x^2}{2}} u^{-\frac{1}{2}} \sum_{n=0}^{+\infty} \frac{(-1)^n u^n}{n!} du \\ &= \frac{1}{2} + \frac{1}{2\sqrt{\pi}} \sum_{n=0}^{+\infty} \frac{(-1)^n}{n!} \int_{0}^{\frac{x^2}{2}} u^{n-\frac{1}{2}} du \\ &= \frac{1}{2} + \frac{1}{2\sqrt{\pi}} \sum_{n=0}^{+\infty} \frac{(-1)^n}{n!\left(n+\frac{1}{2}\right)} \left(\frac{x^2}{2}\right)^{n+\frac{1}{2}} \\ &= \frac{1}{2} + \frac{1}{\sqrt{\pi}} \sum_{n=0}^{+\infty} \frac{(-1)^n x^{2n+1}}{n!(2n+1)(\sqrt{2})^{2n+1}}\end{aligned}$$

我们把后面这部分拿出来，定义为**误差函数**

$$Erf(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt$$

注意，它和我们正态分布在指数上差一个 2！不过古人就是这么定义的，重新定义一个只会造成混乱。其级数展开于是就是

$$Erf(x) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{+\infty} \frac{(-1)^n x^{2n+1}}{n!(2n+1)}$$

有了 Erf，刚才我们的式子就可以写成：

$$\Phi(x) = \frac{1}{2}\left(1 + Erf\left(\frac{x}{\sqrt{2}}\right)\right)$$

我们可以快乐地计算标准正态分布的均值，然后一般的正态分布就呼之欲出了：

$$\begin{aligned}
E(X) &= \int_{-\infty}^{+\infty} x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx \\
&= \int_{0}^{+\infty} x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx - \int_{0}^{+\infty} x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx \\
&= \frac{\Gamma(1)}{\sqrt{2\pi}} - \frac{\Gamma(1)}{\sqrt{2\pi}} \\
&= 0
\end{aligned}$$

我为什么不用奇函数对称区间积分得 0？因为这是广义积分，两边的极限不是对称取得的。对称取得到的是柯西主值，它只在两边都收敛的情况下和广义积分相等。所幸它在这里确实是正确的，但是不加考虑地使用会在柯西分布被偷袭！

标准正态分布的方差：

$$\begin{aligned}
D(X) &= E(X^2) - (E(X))^2 \\
&= \int_{-\infty}^{+\infty} x^2\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx \\
&= 2\int_{0}^{+\infty} x^2\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx \\
&= \frac{2}{\sqrt{\pi}}\int_{0}^{+\infty} u^{\frac{1}{2}}e^{-u}du \qquad\qquad u = \frac{x^2}{2} \\
&= \frac{2}{\sqrt{\pi}}\Gamma\left(\frac{3}{2}\right) = 1
\end{aligned}$$

于是对于一般的正态分布，如果 $Y \sim N(\mu, \sigma^2)$，也就是 $\frac{Y-\mu}{\sigma} \sim N(0,1)$，那么 $E(Y) = \mu$,$D(Y) \sim \sigma^2$。它们被称为位置参数和尺度参数的平方也就很合理了。

我们可以把结论推广到标准正态分布任意阶原点矩。首先，它的任意阶矩都是存在的。

它的奇数阶矩都可以用刚才处理均值的方法得到 0。它的偶数阶矩就需

要稍加计算了，让我们算下 $2m$ 阶原点矩：

$$\begin{aligned}
\mu_{2m} &= \int_{-\infty}^{+\infty} x^{2m} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\
&= 2\int_{0}^{+\infty} x^{2m} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\
&= \frac{2^m}{\sqrt{\pi}} \int_{0}^{+\infty} u^{m-\frac{1}{2}} e^{-u} du \qquad\qquad u = \frac{x^2}{2} \\
&= \frac{2^m}{\sqrt{\pi}} \Gamma\left(m + \frac{1}{2}\right) \\
&= \frac{2^m}{\sqrt{\pi}} \frac{(2m)!}{4^m m!} \sqrt{\pi} \\
&= \frac{(2m)!}{2^m m!}
\end{aligned}$$

另一种写法是双阶乘 $(2m-1)!!$，即 $(2m-1)(2m-3)\cdots 1$。

正态分布是一种稳定的分布，或者说独立的正态分布具有可加性。令 $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$，如果他们独立，则 $Y = X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$。注意是平方的和，不是和的平方。证明如下：

$$\begin{aligned}
f_Y(y) &= \int_{-\infty}^{+\infty} f_{X_1}(t) f_{X_2}(y-t) dt \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(t-\mu_1)^2}{2\sigma_1^2}} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(y-t-\mu_2)^2}{2\sigma_2^2}} dt \\
&= \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{+\infty} Exp\left(-\frac{(t-\mu_1)^2}{2\sigma_1^2} - \frac{(y-t-\mu_2)^2}{2\sigma_2^2}\right) dt
\end{aligned}$$

看上去不是很简单，但是事实上这里没有什么困难，主要是配平方。为了方便，设 $A = \frac{1}{2\sigma_1^2}$，$B = \frac{1}{2\sigma_2^2}$，则有

$$\begin{aligned}
&-A(t-\mu_1)^2 - B(y-t-\mu_2)^2 \\
&= -(A+B)\left[t - \frac{A\mu_1 + B(y-\mu_2)}{A+B}\right]^2 - \frac{(y-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)} \\
&= -(\frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2})\left[t - \frac{\sigma_2^2\mu_1 + \sigma_1^2(y-\mu_2)}{\sigma_1^2+\sigma_2^2}\right]^2 - \frac{(y-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}
\end{aligned}$$

我们先把 $\frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}$ 称为 k，$\left[t-\frac{\sigma_2^2\mu_1+\sigma_1^2(y-\mu_2)}{\sigma_1^2+\sigma_2^2}\right]^2$ 称为 d。代回到原式，我们得到：

$$\begin{aligned}
f_Y(y) &= \frac{1}{2\pi\sigma_1\sigma_2}e^{-\frac{(y-(\mu_1+\mu_2))^2}{2(\sigma_1^2+\sigma_2^2)}}\int_{-\infty}^{+\infty}e^{-k(t-d)^2}dt \\
&= \frac{1}{2\pi\sigma_1\sigma_2}e^{-\frac{(y-(\mu_1+\mu_2))^2}{2(\sigma_1^2+\sigma_2^2)}}\sqrt{\frac{\pi}{k}} \\
&= \frac{1}{2\pi\sigma_1\sigma_2}e^{-\frac{(y-(\mu_1+\mu_2))^2}{2(\sigma_1^2+\sigma_2^2)}}\sqrt{\frac{\pi\sigma_1^2\sigma_2^2}{2(\sigma_1^2+\sigma_2^2)}} \\
&= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2+\sigma_2^2}}e^{-\frac{(y-(\mu_1+\mu_2))^2}{2(\sigma_1^2+\sigma_2^2)}}
\end{aligned}$$

对任意有限多个独立的正态分布也有这样的性质。

5.2.5 多元正态分布

多个正态变量的联合概率分布是有点说法的，我们只需要掌握它们的均值向量和协方差矩阵，它们之间的关系就可以把握了。

我们先给出多元正态分布的定义，再给出我们这样做的合理性。随机向量 $\mathbf{X}=(X_1,\ldots,X_n)^T$，服从多元正态分布记作 $\mathbf{X}\sim N_n(\boldsymbol{\mu},\Sigma)$，其中 Σ 是一个正定对称矩阵。其概率密度为：

$$f_{\mathbf{X}}(\mathbf{x})=\frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det\Sigma}}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

边缘分布正态性 这样定义自然是有内在的合理性的。首先它的每个边缘分布都是正态分布：

将 $\mathbf{X}$ 分块为

$$\mathbf{X}=\begin{pmatrix}X_1\\ \mathbf{X}_{(2)}\end{pmatrix},\quad \boldsymbol{\mu}=\begin{pmatrix}\mu_1\\ \boldsymbol{\mu}_{(2)}\end{pmatrix},\quad \Sigma=\begin{pmatrix}\sigma_{11} & \mathbf{r}^T\\ \mathbf{r} & \Sigma_{22}\end{pmatrix},$$

其中 $\mathbf{r}=\mathrm{Cov}(X_1,\mathbf{X}_{(2)})$ 为 $(n-1)\times 1$ 向量。

根据分块矩阵求逆公式，有

$$(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})=\frac{(x_1-\mu_1)^2}{\sigma_{11}}+\big(\mathbf{x}_{(2)}-\boldsymbol{\mu}_{(2)}-\tfrac{x_1-\mu_1}{\sigma_{11}}\mathbf{r}\big)^T\Sigma_{22|1}^{-1}\big(\mathbf{x}_{(2)}-\boldsymbol{\mu}_{(2)}-\tfrac{x_1-\mu_1}{\sigma_{11}}\mathbf{r}\big),$$

其中 $\Sigma_{22|1} = \Sigma_{22} - \dfrac{1}{\sigma_{11}}\mathbf{r}\mathbf{r}^T$。

X_1 的边缘概率密度为

$$\begin{aligned} f_{X_1}(x_1) &= \int_{\mathbb{R}^{n-1}} f_{\mathbf{X}}(x_1, \mathbf{x}_{(2)})\, d\mathbf{x}_{(2)} \\ &= \frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}} \exp\Big[-\frac{(x_1-\mu_1)^2}{2\sigma_{11}}\Big] \int_{\mathbb{R}^{n-1}} \exp\Big[-\frac{1}{2}Q(\mathbf{x}_{(2)}; x_1)\Big] d\mathbf{x}_{(2)}, \end{aligned}$$

其中

$$Q(\mathbf{x}_{(2)}; x_1) = \big(\mathbf{x}_{(2)} - \boldsymbol{\mu}_{(2)} - \tfrac{x_1-\mu_1}{\sigma_{11}}\mathbf{r}\big)^T \Sigma_{22|1}^{-1} \big(\mathbf{x}_{(2)} - \boldsymbol{\mu}_{(2)} - \tfrac{x_1-\mu_1}{\sigma_{11}}\mathbf{r}\big).$$

这就是剥离第一行第一列成分后得到的二次型

对 $\mathbf{x}_{(2)}$ 的积分是多元正态 $N\big(\boldsymbol{\mu}_{(2)} + \frac{x_1-\mu_1}{\sigma_{11}}\mathbf{r},\ \Sigma_{22|1}\big)$ 的归一化积分：

$$\int_{\mathbb{R}^{n-1}} \exp\Big[-\frac{1}{2}Q(\mathbf{x}_{(2)}; x_1)\Big] d\mathbf{x}_{(2)} = (2\pi)^{(n-1)/2}\sqrt{\det\Sigma_{22|1}}.$$

代入并利用行列式公式 $\det\Sigma = \sigma_{11}\cdot\det\Sigma_{22|1}$，得

$$\begin{aligned} f_{X_1}(x_1) &= \frac{1}{(2\pi)^{n/2}\sqrt{\sigma_{11}\det\Sigma_{22|1}}} \exp\Big[-\frac{(x_1-\mu_1)^2}{2\sigma_{11}}\Big] \cdot (2\pi)^{(n-1)/2}\sqrt{\det\Sigma_{22|1}} \\ &= \frac{1}{\sqrt{2\pi\sigma_{11}}} \exp\Big[-\frac{(x_1-\mu_1)^2}{2\sigma_{11}}\Big]. \end{aligned}$$

这正是 $N(\mu_1, \sigma_{11})$ 的概率密度函数。

可逆线性变换不变性　多元正态分布还有一个很好的性质，即线性变换不变性。若 $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$，线性变换 $A : R^n \to R^m$ 为满射，$\mathbf{b}$ 为常向量，则 $A\mathbf{X} + \mathbf{b} \sim N_m(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T)$。

我们先证明一个弱一点的结论，即 $n = m$ 的情况，也就是 A 是一个 $\mathbb{R}^n$ 的线性自同构。不过这里我们仍然可以类比一元时的思路：我们可以把任意正态分布做线性变换变成标准正态分布，把标准正态分布变成任何正态分布。在多元的情况下，具有标准正态分布地位的这个分布是 $N_n(\mathbf{0}, I_n)$。由于 Σ 是一个正算子，所以我可以得出其唯一的正平方根分解 $\sqrt{\Sigma}$。和一元的情况简直如出一辙，令 $\mathbf{U} = (\sqrt{\Sigma})^{-1}(\mathbf{X} - \boldsymbol{\mu})$ 就有了！这是可逆的情况，所以用超棒的变量代换公式就可以验证。反过来，我刚才给出的变换也是可逆的，这可太棒了。

协方差 插上一嘴，$\mathbf{U} \sim N_n(\mathbf{0}, I_n)$ 这个分布的诸分量是相互独立的，证明也是直接的，因为 $\mathbf{x}^T I_n^{-1} \mathbf{x}$ 就是 $\sum_{i=1}^n x_i^2$，很容易被拆开。类似而更一般地，如果 Σ 是个对角矩阵，诸分量也是独立的。我们也就可以把 $N_n(\mathbf{0}, I_n)$ 叫做独立标准正态分布。

甚至更进一步地，这告诉我们 Σ 恰好就是 $Cov(\mathbf{X})$。按刚才的推导，$\mathbf{X} = \boldsymbol{\mu} + \sqrt{\Sigma}\mathbf{U}$，其中 $\mathbf{U} \sim N_n(\mathbf{0}, I_n)$，于是乎

$$\begin{aligned} Cov(\mathbf{X}) &= E((\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T) \\ &= E((\sqrt{\Sigma}\mathbf{U})(\sqrt{\Sigma}\mathbf{U})^T) \\ &= \sqrt{\Sigma} E(\mathbf{U}\mathbf{U}^T) \sqrt{\Sigma}^T \\ &= \sqrt{\Sigma} I_n \sqrt{\Sigma}^T \\ &= \Sigma \end{aligned}$$

独立性的等价条件 关于多元正态分布，还有更好的性质：两个边缘分布不相关当且仅当两个边缘分布独立。这一结论推广到任意多个变量也是成立的。我们下面证明更困难的那个方向：

设 $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$。将 $\mathbf{X}$ 分成两个子向量 $\mathbf{X_1}, \mathbf{X_2}$。设 $\Sigma_{12} = \mathbf{0}$，我们来证明 $\mathbf{X}_1$ 与 $\mathbf{X}_2$ 独立。

由于 $\Sigma_{12} = \mathbf{0}$，协方差矩阵为分块对角矩阵：

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \mathbf{0} \\ \mathbf{0} & \Sigma_{22} \end{pmatrix}.$$

其逆矩阵为

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \Sigma_{22}^{-1} \end{pmatrix}.$$

行列式为

$$\det \Sigma = \det \Sigma_{11} \cdot \det \Sigma_{22}.$$

$\mathbf{X}$ 的联合概率密度为

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\Big[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big] \\ &= \frac{1}{(2\pi)^{p/2}|\Sigma_{11}|^{1/2}} \exp\Big[-\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu}_1)^{\top}\Sigma_{11}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1)\Big] \\ &\quad \cdot \frac{1}{(2\pi)^{q/2}|\Sigma_{22}|^{1/2}} \exp\Big[-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu}_2)^{\top}\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\Big] \\ &= f_{\mathbf{X}_1}(\mathbf{x}_1)\cdot f_{\mathbf{X}_2}(\mathbf{x}_2) \end{aligned}$$

这正是 $\mathbf{X}_1$ 的边缘概率密度 $f_{\mathbf{X}_1}(\mathbf{x}_1)$ 与 $\mathbf{X}_2$ 的边缘概率密度 $f_{\mathbf{X}_2}(\mathbf{x}_2)$ 的乘积! 二者独立。

更一般的情况：满线性映射不变性 现在考虑一般的满射线性映射 $A:\mathbb{R}^n\to\mathbb{R}^m$。我们可以将 $A\mathbf{X}+\mathbf{b}$ 分解为三个线性变换的复合:

1. **标准化：** $\mathbf{X}\mapsto\mathbf{U}=(\sqrt{\Sigma})^{-1}(\mathbf{X}-\boldsymbol{\mu})$，得到 $N_n(\mathbf{0},I_n)$。

2. **投影：** 由于 $\mathbf{U}$ 的分量独立同分布 $N(0,1)$，取它的前 m 个分量 $\mathbf{U}_m=(U_1,\dots,U_m)^T$ 即得到 $N_m(\mathbf{0},I_m)$。更一般地，满射 A 作用在 $\mathbf{X}$ 上相当于作用在 $\mathbf{U}$ 上:

$$A\mathbf{X}+\mathbf{b}=A(\boldsymbol{\mu}+\sqrt{\Sigma}\,\mathbf{U})+\mathbf{b}=(A\boldsymbol{\mu}+\mathbf{b})+A\sqrt{\Sigma}\,\mathbf{U}.$$

记 $\tilde{A}=A\sqrt{\Sigma}\in\mathbb{R}^{m\times n}$，则 $\tilde{A}\mathbf{U}$ 是独立标准正态变量的线性组合。

3. **重新参数化：** $\tilde{A}\mathbf{U}\sim N_m(\mathbf{0},\tilde{A}\tilde{A}^T)$，且

$$\tilde{A}\tilde{A}^T=A\sqrt{\Sigma}(A\sqrt{\Sigma})^T=A\Sigma A^T.$$

于是

$$A\mathbf{X}+\mathbf{b}\sim N_m(A\boldsymbol{\mu}+\mathbf{b},A\Sigma A^T).$$

这个分解展示了: 任意多元正态分布的线性变换只是对独立标准正态变量做线性组合，从而结果仍是多元正态。

标准独立正态的正交变换不变性 对标准独立正态分布应用正交变换 Q, 还有更好的结果。若 $\mathbf{X}\sim N_n(\mathbf{0},I_n)$, 则有 $Q\mathbf{X}\sim N_n(Q\mathbf{0},QI_nQ^T)=N_n(\mathbf{0},I_n)$。也就是说正交变换保持标准独立正态分布不变。

服从多元正态分布的等价条件 这一结论应用在 $\mathbb{R}^n$ 上的线性泛函时有奇效。正向结论是成立的，反向结论也是成立的！就是说：$\mathbf{X} = (X_1, \ldots, X_n)^T$ 服从多元正态分布**当且仅当**对于任意线性泛函 $\phi : \mathbb{R}^n \to \mathbb{R}$，随机变量 $\phi(\mathbf{X})$ 服从一元正态分布。

反方向的证明如下：

证明. 由 Riesz 表示定理，任意线性泛函 ϕ 都可以表示成和某个向量 $\mathbf{v}$ 的内积，即 $\phi(\mathbf{X}) = \mathbf{c}^T\mathbf{X}$。

按前面的结论,我们可以猜想它其实就服从 $N_n(\boldsymbol{\mu}, \Sigma)$,其中 $\boldsymbol{\mu} = E(\mathbf{X}), \Sigma = Cov(\mathbf{X})$，但是要证明它还是需要一点技巧的。

我们可以先尝试构造相互独立的正态变量，然后再达到它。因为我们知道：独立的正态分布的联合分布确实是多元独立正态分布（只需要做个简单的乘法）。

它的线性代数表达就是对 Σ 做谱分解得到 $Q\Lambda Q^T$，这里 Q 是正交的，Λ 是对角的特征值矩阵。我们简单地令 $\mathbf{Z} = Q^T(\mathbf{X} - \boldsymbol{\mu})$，那么它均值为 $\mathbf{0}$，其协方差

$$Cov(\mathbf{Z}) = Cov(Q^T(\mathbf{X} - \boldsymbol{\mu})) = Q^T Cov(\mathbf{X})Q = Q^T Q\Lambda Q^T Q = \Lambda$$

太好了，$\mathbf{Z}$ 各分量是彼此不相关的！我们只需要证明各分量是正态的，就知道各分量是相互独立的！而我做的变换也确实能保证 $\mathbf{Z}$ 诸分量是正态的：

$$\begin{aligned} Z_i &= \mathbf{e}_i^T\mathbf{Z} \\ &= \mathbf{e}_i^T Q^T(\mathbf{X} - \boldsymbol{\mu}) \\ &= (Q\mathbf{e}_i)^T\mathbf{X} - (Q\mathbf{e}_i)^T\boldsymbol{\mu} \end{aligned}$$

由假设，$(Q\mathbf{e}_i)^T\mathbf{X}$ 是正态，加上常数仍然正态。于是乎，$\mathbf{Z}$ 就确定是服从 $N_n(\mathbf{0}, \Lambda)$ 的。$\mathbf{X} = Q\mathbf{Z} + \boldsymbol{\mu}$，于是就有 $\mathbf{X} \sim N_n(\boldsymbol{\mu}, Q\Lambda Q^T) = N_n(\boldsymbol{\mu}, \Sigma)$。 □

有了这些性质，我们就能更向前迈进了。

5.2.6 Rayleigh 分布

瑞利分布常用于描述二维向量的模长分布。它在形式上类似于指数分布的升级版。

若随机变量 X 的概率密度函数为

$$f_X(x) = \begin{cases} \dfrac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

其中 $\sigma > 0$ 为尺度参数，则称 X 服从参数为 σ 的 Rayleigh 分布，记作 $X \sim \text{Rayleigh}(\sigma)$。

其累积分布函数为

$$F_X(x) = \begin{cases} 1 - e^{-\frac{x^2}{2\sigma^2}}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

如果随机变量 Y_1, Y_2 均服从 $N(0, \sigma^2)$ 且相互独立，则 $R = \sqrt{Y_1^2 + Y_2^2} \sim \text{Rayleigh}(\sigma)$。

它的均值

$$\begin{aligned} E(X) &= \int_0^{+\infty} \frac{x^2}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} dx \\ &= \sqrt{2}\sigma \int_0^{+\infty} u^{\frac{1}{2}} e^{-u} du \qquad\qquad u = \frac{x^2}{2\sigma^2} \\ &= \sqrt{2}\sigma \Gamma(\frac{3}{2}) = \sigma\sqrt{\frac{\pi}{2}} \end{aligned}$$

方差

$$\begin{aligned} D(X) &= E(X^2) - (E(X))^2 \\ &= \int_0^{+\infty} \frac{x^3}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} dx - \sigma^2 \frac{\pi}{2} \\ &= 2\sigma^2 \int_0^{+\infty} u e^{-u} du - \sigma^2 \frac{\pi}{2} \\ &= 2\sigma^2 \Gamma(2) - \sigma^2 \frac{\pi}{2} \\ &= \frac{4 - \pi}{2} \sigma^2 \end{aligned}$$

通信中平坦衰落信道的包络，雷达信号中杂波幅值，风速、海洋波高等物理量都可以用瑞利分布拟合。

5.2.7 Weibull 分布

韦布尔分布是可靠性工程与生存分析中极为重要的连续分布，它是指数分布与 Rayleigh 分布的推广，也是三类极值分布之一。

若随机变量 X 的概率密度函数为

$$f_X(x)=\begin{cases}\dfrac{k}{\lambda}\left(\dfrac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^k}, & x\geq 0,\\ 0, & x<0,\end{cases}$$

其中 $k>0$ 为形状参数，$\lambda>0$ 为尺度参数，则称 X 服从 Weibull 分布，记作 $X\sim\text{Weibull}(k,\lambda)$。

其累积分布函数为

$$F_X(x)=\begin{cases}1-e^{-(x/\lambda)^k}, & x\geq 0,\\ 0, & x<0.\end{cases}$$

可以看到，当 $k=1$ 时，退化为指数分布 $Exp(\lambda)$；当 $k=2$ 时，即为 Rayleigh 分布 $\text{Rayleigh}\left(\frac{\lambda}{\sqrt{2}}\right)$；当 $k\approx 3.6$ 时，分布形状很接近正态分布。

其均值:

$$\begin{aligned}E(X)&=\int_0^{+\infty}k\left(\frac{x}{\lambda}\right)^k e^{-(x/\lambda)^k}dx\\ &=\lambda\int_0^{+\infty}u^{\frac{1}{k}}e^{-u}du && u=\left(\frac{x}{\lambda}\right)^k\\ &=\lambda\Gamma\left(1+\frac{1}{k}\right)\end{aligned}$$

其方差:

$$\begin{aligned}D(X)&=E(X^2)-(E(X))^2\\ &=\int_0^{+\infty}kx\left(\frac{x}{\lambda}\right)^k e^{-(x/\lambda)^k}dx-\lambda^2\Gamma^2\left(1+\frac{1}{k}\right)\\ &=\lambda^2\int_0^{+\infty}u^{\frac{2}{k}}e^{-u}du-\lambda^2\Gamma^2\left(1+\frac{1}{k}\right) && u=\left(\frac{x}{\lambda}\right)^k\\ &=\lambda^2\left(\Gamma\left(1+\frac{2}{k}\right)-\Gamma^2\left(1+\frac{1}{k}\right)\right)\end{aligned}$$

5.2.8 Gamma 分布

Gamma 分布由 Γ 函数标准化而来，也可以作为泊松过程中第 s（非整数）个事件的等待时间分布的推广

若随机变量 X 的概率密度函数为

$$f_X(x) = \begin{cases} \dfrac{x^{s-1}e^{-\frac{x}{\theta}}}{\theta^s\Gamma(s)}, & x > 0, \\ 0, & x \le 0, \end{cases}$$

其中 $s > 0$ 为形状参数，$\theta > 0$ 为尺度参数，则称 X 服从 Gamma 分布，记作 $X \sim \Gamma(s, \theta)$。

当 $s = 1$ 时，即为指数分布 $Exp(\theta)$；当 $s = n$（正整数）时，即为 Erlang 分布 $Erlang(n, \theta)$。

Gamma 分布具有可加性，若 $X \sim \Gamma(s_1, \theta)$，$Y \sim \Gamma(s_2, \theta)$ 且独立，则 $X + Y \sim \Gamma(s_1 + s_2, \theta)$。这一证明和前面 Erlang 分布的可加性如出一辙，只不过把阶乘换成 Γ 函数就是了。

也很容易推广出其均值 $E(X) = s\theta$，方差 $D(X) = s\theta^2$。

5.2.9 Beta 分布

Beta 分布是定义在区间 $(0, 1)$ 上的连续分布，由 Beta 函数标准化而来。

若随机变量 X 的概率密度函数为

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

其中 $\alpha > 0$，$\beta > 0$ 为形状参数，则称 X 服从 Beta 分布，记作 $X \sim \text{Beta}(\alpha, \beta)$。

其均值

$$\begin{aligned} E(X) &= \int_0^1 \frac{x^{\alpha}(1-x)^{\beta-1}}{B(\alpha, \beta)} dx \\ &= \frac{B(\alpha+1, \beta)}{B(\alpha, \beta)} = \frac{\alpha}{\alpha+\beta} \end{aligned}$$

方差

$$\begin{aligned} D(X) &= E(X^2) - (E(X))^2 \\ &= \int_0^1 \frac{x^{\alpha+1}(1-x)^{\beta-1}}{B(\alpha,\beta)}dx - \frac{\alpha^2}{(\alpha+\beta)^2} \\ &= \frac{B(\alpha+2,\beta)}{B(\alpha,\beta)} - \frac{\alpha^2}{(\alpha+\beta)^2} \\ &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \end{aligned}$$

从 Beta 函数的定义就能很自然得出：若 $Y_1 \sim \Gamma(\alpha,1)$，$Y_2 \sim \Gamma(\beta,1)$ 且独立，则

$$X = \frac{Y_1}{Y_1+Y_2} \sim \text{Beta}(\alpha,\beta).$$

Beta 分布可以用来描述顺序统计量的分布：

对于 $U_{(1)},\ldots,U_{(n)}$ 为来自 $U(0,1)$ 的次序统计量（即我们从小到大排了个序），有 $U_{(k)} \sim \text{Beta}(k,n-k+1)$。这也是因为 Beta 函数就可以理解为组合数的一种解析延拓。

5.2.10 Cauchy 分布

柯西分布一般是作为一个经典反例而出现的。我们记随机变量 X 服从柯西分布为 $X \sim Cauchy(a,b)$。a 和 b 分别被称为位置参数和尺度参数。

柯西分布的概率密度函数是：

$$f_X(x) = \frac{1}{b\pi}\frac{1}{1+(\frac{x-a}{b})^2}$$

令 $a=0$，$b=1$ 就得到标准柯西分布 $\frac{1}{\pi}\frac{1}{1+x^2}$。显然任意柯西分布 $X \sim Cauchy(a,b)$ 可以通过变量代换得出 $a+bX \sim Cauchy(0,1)$。

它最显著的性质就是：没有均值！不妨对标准柯西分布尝试一下：

$$\int_{-\infty}^{\infty} x\frac{1}{\pi}\frac{1}{1+x^2}dx$$

看上去是个奇函数，可以直接对称区间积分得零。**但这是广义积分**。任意一边拿出来积分都不收敛：

$$\frac{1}{\pi}\frac{x}{1+x^2} > \frac{1}{\pi x}$$

后面的几个经典不等式和大数定律也将与它无缘。

不过蛮好,柯西分布稳定可加。令 $X \sim Cauchy(a_1, b_1)$, $Y \sim Cauchy(a_2, b_2)$, X, Y 独立则有 $X + Y \sim Cauchy(a_1 + a_2, b_1 + b_2)$。如果缺乏复分析技巧的话，仅用积分技巧很难得到这个结论。而使用复分析的手法来计算的话，思路非常简单，只是有一些烦人的代数化简，所以这里就姑且跳过证明。真正好使的处理方法是用特征函数。

6 经典定理

6.1 马尔可夫不等式

该不等式不要求方差存在，只需非负性与期望存在。其形式简单但威力强大，是后续更精细不等式的基础。马尔可夫不等式给出了非负随机变量超出某正数的概率上界。

定理 4 (马尔可夫不等式)**.** *设 X 为非负随机变量（即 $P(X \geq 0) = 1$），且均值 $E(X)$ 存在。则对任意实数 $a > 0$，有*

$$P(X \geq a) \leq \frac{E(X)}{a}.$$

证明. 证明是非常简单直接的。这里用连续的形式表示，离散的情况是类似的。

$$\begin{aligned} P(X \geq a) &= \int_a^{+\infty} f(x)dx \\ &\leq \int_a^{+\infty} \frac{x}{a} f(x)dx \\ &\leq \int_0^{+\infty} \frac{x}{a} f(x)dx = \frac{E(X)}{a} \end{aligned}$$

□

6.2 切比雪夫不等式

切比雪夫不等式利用方差对随机变量偏离其均值的程度给出概率上界。

定理 5 (切比雪夫不等式)**.** *设随机变量 X 的数学期望 $E(X) = \mu$ 和方差 $D(X) = \sigma^2$ 均存在（$\sigma^2 > 0$）。则对任意实数 $k > 0$，有*

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

证明. 对非负随机变量 $(X - \mu)^2$ 应用马尔可夫不等式，取 $a = \epsilon^2$，得

$$P\left((X - \mu)^2 \geq \epsilon^2\right) \leq \frac{E((X - \mu)^2)}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}.$$

□

该不等式给出了分布未知时偏差概率的保守估计。特别地，取 $\epsilon = m\sigma$ 可得 $P(|X-\mu| \geq m\sigma) \leq \frac{1}{m^2}$。它表明方差越小，随机变量越集中在均值附近。

如果此时 $\sigma = 0$，那么 $P(X = \mu) = 1$，也就是 X **几乎处处**为常数!

6.3 弱大数定律

弱大数定律描述了独立同分布随机变量序列的样本均值依概率收敛于总体均值的经典结论。

定理 6 (弱大数定律)**.** *设 $X_1, X_2, \ldots, X_n, \ldots$ 是独立同分布的随机变量序列，且 $E(X_i) = \mu$，$D(X_i) = \sigma^2 < \infty$。记样本均值为*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

则对任意 $\varepsilon > 0$，有

$$\lim_{n\to\infty} P\left(\left|\overline{X}_n - \mu\right| \geq \varepsilon\right) = 0.$$

即 $\overline{X}_n$ 依概率收敛于 μ，记作 $\overline{X}_n \xrightarrow{P} \mu$。

证明. 由独立同分布性，

$$E(\overline{X}_n) = \mu, \quad D(\overline{X}_n) = \frac{\sigma^2}{n}.$$

对 $\overline{X}_n$ 应用切比雪夫不等式得

$$P\left(\left|\overline{X}_n - \mu\right| \geq \varepsilon\right) \leq \frac{D(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

令 $n \to \infty$，右端趋于 0，即得

$$\lim_{n\to\infty} P\left(\left|\overline{X}_n - \mu\right| \geq \varepsilon\right) = 0.$$

□

注意这里给出的该定律是方差有限的形式。若去掉方差有限的条件，仅保留期望存在，仍有一个更一般的弱大数定律（辛钦大数定律），但证明需用特征函数等工具。

并且注意这里的收敛性是依概率收敛。这里有几种不同的收敛性：依分布收敛，依概率收敛，几乎必然收敛，均方收敛。强大数定律的收敛是几乎必然收敛。这几种收敛性的定义和区别如下：

1. **依分布收敛**：分布函数逐点收敛到极限分布函数（在连续点处）。
2. **依概率收敛**：随机变量序列与极限随机变量的偏差超过任意给定值的概率趋于零。
3. **几乎必然收敛**：随机变量序列以概率 1 收敛到极限随机变量。
4. **均方收敛**：随机变量序列与极限随机变量的均方差趋于零。

它们之间的强弱关系为：

$$\text{几乎必然收敛} \implies \text{依概率收敛} \implies \text{依分布收敛},$$

$$\text{均方收敛} \implies \text{依概率收敛}.$$

收敛强弱关系的证明就不在这里给出了。

这一大数定律也给出了伯努利大数定律，联系概率与频率的理论基础。

定理 7 (伯努利大数定律)**.** *设 n_A 为 n 次独立伯努利试验中事件 A 发生的次数，每次试验中 A 发生的概率为 p $(0<p<1)$。则对任意 $\varepsilon>0$，有*

$$\lim_{n\to\infty} P\left(\left|\frac{n_A}{n}-p\right|\geq\varepsilon\right)=0.$$

即事件 A 发生的频率 $f_A=\frac{n_A}{n}$ 依概率收敛于其概率 p。

7 中心极限定理

中心极限定理是概率论中最重要的定理之一，它揭示了大量独立（得有均值方差）随机变量和的标准化形式依分布收敛于标准正态分布的规律。

7.1 矩母函数和特征函数（傅里叶变换）

要严格证明中心极限定理，我们需要引入刻画随机变量分布特征的工具——矩母函数和特征函数。

矩母函数 设随机变量 X 的分布函数为 $F_X(x)$。若存在 $h>0$，使得对任意 $|t|<h$，期望 $E(e^{tX})$ 存在，则称

$$M_X(t)=E(e^{tX})=\int_{-\infty}^{+\infty}e^{tx}dF_X(x)$$

为 X 的矩母函数。

事实上另一个等价的定义才是其得名的原因：

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{+\infty} e^{tx} dF_X(x) \\ &= \int_{-\infty}^{+\infty} \sum_{n=0}^{+\infty} \frac{(tx)^n}{n!} dF_X(x) \\ &= \sum_{n=0}^{\infty} \frac{t^n}{n!} \int_{-\infty}^{+\infty} x^n dF_X(x) \\ &= \sum_{n=0}^{\infty} \frac{\mu_n t^n}{n!} \end{aligned}$$

如果它的各阶矩存在的话，这就是其各阶矩给出的一个泰勒级数。

有一件好事是，若两个随机变量的矩母函数在包含 0 某个区间内相等，则它们服从同一分布。这事实上就是解析函数的性质使然。至于为什么包含 0 呢？因为在 $t = 0$ 处，矩母函数的诸阶导数就是诸阶矩。

然而，仅凭矩相等（而不假定矩母函数在 0 点邻域存在）并不能保证分布相同。著名的反例由 Stieltjes 和 Heyde 给出：

考虑密度函数

$$f_0(x) = \frac{1}{\sqrt{2\pi}} x^{-1} e^{-(\ln x)^2/2}, \quad x > 0,$$

以及对于 $\alpha \in [-1, 1]$，

$$f_\alpha(x) = f_0(x)\big[1 + \alpha \sin(2\pi \ln x)\big], \quad x > 0.$$

这些分布具有以下性质：

- 对任意 $\alpha \in [-1, 1]$，$f_\alpha(x)$ 是合法的概率密度函数（非负且归一）。
- 所有 f_α 对应的分布具有完全相同的各阶矩：

$$\int_0^\infty x^k f_\alpha(x) dx = e^{k^2/2}, \quad k = 0, 1, 2, \ldots$$

- 但当 $\alpha \neq 0$ 时，f_α 与 f_0 是不同的分布。

这个反例之所以成立，是因为这些分布的矩母函数 $M(t) = E(e^{tX})$ 在 $t > 0$ 时不收敛（矩增长太快，$E(e^{tX}) = \infty$ 对任意 $t > 0$），因此定理中“矩

母函数在包含 0 的区间内存在”的条件不满足。此时仅凭矩序列无法唯一确定分布。

这一现象也出现在具有**本性奇点**的复变函数中：尽管函数在某点所有导数都相同（对应矩相等），但函数本身在该点邻域内可能有不同的解析延拓（对应不同分布）。解析性始终是这个问题的关键。

不过在性质良好的分布情况下，我们有更精彩的结论：**卷积定理**。如果随机变量 X, Y 相互独立，且矩母函数分别为 $M_X(t), M_Y(t)$，则 $Z = X + Y$ 的矩母函数就有 $M_Z(t) = M_X(t)M_Y(t)$。我在这里给出一个连续情况的证明：

$$
\begin{aligned}
M_Z(t) &= \int_{-\infty}^{+\infty} e^{tz} f_Z(z) dz \\
&= \int_{-\infty}^{+\infty} e^{tz} \left(\int_{-\infty}^{+\infty} f_X(\tau) f_Y(z-\tau) d\tau \right) dz \\
&= \int_{-\infty}^{+\infty} e^{t\tau} f_X(\tau) \left(\int_{-\infty}^{+\infty} e^{t(z-\tau)} f_Y(z-\tau) dz \right) d\tau \\
&= \int_{-\infty}^{+\infty} e^{t\tau} f_X(\tau) \left(\int_{-\infty}^{+\infty} e^{tu} f_Y(u) du \right) d\tau \qquad\qquad u = z - \tau \\
&= \left(\int_{-\infty}^{+\infty} e^{t\tau} f_X(\tau) d\tau \right) \left(\int_{-\infty}^{+\infty} e^{tu} f_Y(u) du \right) \\
&= M_X(t) M_Y(t)
\end{aligned}
$$

特征函数　坏消息是，矩母函数不一定收敛。不过好消息是，我们在实的解析性难以取得的时候，可以考虑转向复的解析性。

设随机变量 X 的分布函数为 $F_X(x)$，称

$$\phi_X(t) = E(e^{itX}) = \int_{-\infty}^{+\infty} e^{itx} dF_X(x)$$

为 X 的特征函数。

特征函数的好处是：对任意随机变量 X 和任意 $t \in \mathbb{R}$，有 $|e^{itX}| = 1$，故 $|E(e^{itX})| \leq 1$，特征函数总是存在。

而且显然它存在就具有解析性，因为指数就是解析的。于是就有特征函数与分布函数一一对应。只要在包含 0 的足够稠密的区间上两个特征函数处处相等，则两个分布也相等。我们可以找出一种从特征函数反过来求分布的逆变换，不过我们先把它放在后面。

如果矩都收敛的话，这是个复的泰勒级数：

$$\begin{aligned}\phi_X(t) &= \int_{-\infty}^{+\infty} e^{itx} dF_X(x) \\ &= \sum_{n=0}^{\infty} \frac{(it)^n}{n!} \int_{-\infty}^{+\infty} x^n dF_X(x) \\ &= \sum_{n=0}^{\infty} \mu_n i^n \frac{t^n}{n!}\end{aligned}$$

此时能求导得回 $\mu_k = i^{-k}\phi_X^{(k)}(0)$。

眼尖的朋友会发现，这不就是傅里叶变换吗（不过差了一个符号）？掏出我们的二重积分来验证，卷积定理仍然成立。如果随机变量 X, Y 相互独立，且特征函数分别为 $\phi_X(t), \phi_Y(t)$，则 $Z = X + Y$ 的特征函数就有 $\phi_Z(t) = \phi_X(t)\phi_Y(t)$。

$$\begin{aligned}\phi_Z(t) &= \int_{-\infty}^{+\infty} e^{itz} f_Z(z) dz \\ &= \int_{-\infty}^{+\infty} e^{itz} \left(\int_{-\infty}^{+\infty} f_X(\tau) f_Y(z-\tau) d\tau \right) dz \\ &= \int_{-\infty}^{+\infty} e^{it\tau} f_X(\tau) \left(\int_{-\infty}^{+\infty} e^{it(z-\tau)} f_Y(z-\tau) dz \right) d\tau \\ &= \int_{-\infty}^{+\infty} e^{it\tau} f_X(\tau) \left(\int_{-\infty}^{+\infty} e^{itu} f_Y(u) du \right) d\tau \qquad u = z - \tau \\ &= \left(\int_{-\infty}^{+\infty} e^{it\tau} f_X(\tau) d\tau \right) \left(\int_{-\infty}^{+\infty} e^{itu} f_Y(u) du \right) \\ &= \phi_X(t)\phi_Y(t)\end{aligned}$$

非常容易归纳到任意多相互独立随机变量，卷积计算大大简化，而且总能这么做。

傅里叶变换还提醒我们尺度收缩和平移性质，放在一起就是线性变换性质。若 X 的特征函数为 $\phi_X(t)$，则 $Y = a + bX \quad (b \neq 0)$ 的特征函数 $\phi_Y(t)$

有：

$$\begin{aligned}
\phi_Y(t) &= \int_{-\infty}^{+\infty} e^{ity} f_Y(y) dy \\
&= \int_{-\infty}^{+\infty} e^{ity} \frac{1}{|b|} f_X\left(\frac{y-a}{b}\right) dy \\
&= \int_{-\infty}^{+\infty} e^{it(a+bu)} f_X(u) du \qquad\qquad u = \frac{y-a}{b} \\
&= e^{ita} \int_{-\infty}^{+\infty} e^{i(bt)u} f_X(u) du \\
&= e^{ita} \phi_X(bt)
\end{aligned}$$

傅里叶变换告诉我们逆变换是怎样的（只要 ϕ_X 绝对可积）：

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itx} \phi_X(t) dt.$$

上面给出的一些分布的特征函数是这样的：

- **正态分布** $N(\mu, \sigma^2)$：

$$\phi(t) = \exp\left(i\mu t - \frac{\sigma^2 t^2}{2}\right).$$

- **泊松分布** $\pi(\lambda)$：

$$\phi(t) = \exp\left[\lambda(e^{it} - 1)\right].$$

- **指数分布** $\mathrm{Exp}(\theta)$：

$$\phi(t) = \frac{1}{1 - i\theta t}.$$

- **伽马分布** $\Gamma(s, \theta)$：

$$\phi(t) = \frac{1}{(1 - i\theta t)^s}.$$

- **柯西分布** $Cauchy(a, b)$：

$$\phi(t) = e^{iat - b|t|}.$$

在走向中心极限定理之前，我们只差一个结论了，即特征函数连续性定理。

定理 8 (连续性定理)**.** 分布函数列 $\{F_n\}$ 弱收敛（也就是**几乎处处**收敛）于分布函数 F 当且仅当相应的特征函数列 $\{\phi_n(t)\}$ 逐点收敛于 F 的特征函数 $\phi(t)$，且 $\phi(t)$ 在 $t=0$ 处连续。

证明涉及控制收敛定理，本人放弃在此说明。

7.2 中心极限定理

这里我们终于可以真正严格地走向中心极限定理了。

定理 9 (Lindeberg–Lévy 中心极限定理)**.** 设 $X_1, X_2, \ldots$ 是**独立同分布**的随机变量序列，$E(X_i)=\mu$，$D(X_i)=\sigma^2<\infty$ $(\sigma>0)$。记

$$S_n=\sum_{i=1}^{n} X_i, \quad \overline{X}_n=\frac{S_n}{n}.$$

则标准化和

$$Z_n=\frac{S_n-n\mu}{\sigma\sqrt{n}}=\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}$$

依分布收敛于标准正态分布：

$$Z_n \xrightarrow{d} N(0,1), \quad n\to\infty.$$

即对任意 $x\in\mathbb{R}$，

$$\lim_{n\to\infty} P(Z_n\le x)=\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}dt.$$

证明. 设 $Y_i=\frac{X_i-\mu}{\sigma}$，则 $E(Y_i)=0$，$D(Y_i)=1$，且 $Z_n=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Y_i$。

记 $\phi(t)$ 为 Y_i 的特征函数。由于 $E(Y_i)=0$，$E(Y_i^2)=1$，且高阶矩有限，由泰勒展开：

$$\phi(t)=1-\frac{t^2}{2}+o(t^2), \quad t\to 0.$$

Z_n 的特征函数为：

$$\phi_{Z_n}(t)=E\left(e^{itZ_n}\right)=E\left(e^{i\frac{t}{\sqrt{n}}\sum_{i=1}^{n}Y_i}\right)=\left[\phi\left(\frac{t}{\sqrt{n}}\right)\right]^n.$$

取对数得：

$$\ln\phi_{Z_n}(t)=n\ln\left[1-\frac{t^2}{2n}+o\left(\frac{t^2}{n}\right)\right].$$

利用 $\ln(1+z)=z+o(z)\ (z\to 0)$，当 $n\to\infty$ 时：

$$\ln\phi_{Z_n}(t)=n\left[-\frac{t^2}{2n}+o\left(\frac{t^2}{n}\right)\right]=-\frac{t^2}{2}+o(1).$$

故

$$\lim_{n\to\infty}\phi_{Z_n}(t)=e^{-t^2/2}.$$

而 $e^{-t^2/2}$ 正是标准正态分布 $N(0,1)$ 的特征函数。

由连续性定理，特征函数逐点收敛且极限函数 $e^{-t^2/2}$ 在 $t=0$ 处连续（实际上在全实轴连续），故 Z_n 的分布函数弱收敛于标准正态分布函数 $\Phi(x)$。 □

干得漂亮！

随之有个简单的推论。如果我们把二项分布看成独立同分布两点分布之和，应用中心极限定理就得到

定理 10 (Laplace-De Moivre 中心极限定理)**.** *若 $X\sim b(n,p)$，则 $n\to\infty$ 时有：*

$$\frac{X-np}{\sqrt{np(1-p)}}\xrightarrow{d}N(0,1)$$

像这样的中心极限定理还很多，不过万变不离其宗。我们可以对泊松分布、Erlang 分布、卡方分布等也给出类似结论。

8 样本和抽样分布

8.1 随机样本和统计量

在统计中，我们将研究对象的全体称为**总体**，组成总体的每个基本单元称为**个体**。总体通常用一个随机变量 X 及其概率分布来描述，该分布称为**总体分布**。

为了解总体的性质，我们不可能研究所有个体，所以需要一个**抽样**的过程。

从总体 X 中随机抽取 n 个个体 $X_1, X_2, \ldots, X_n$，称之为一个**样本**，n 称为**样本容量**。

在简单随机抽样的过程中，我们为了使样本能有效代表总体，我们要求抽样满足：

- 代表性：每个 X_i 都与总体 X 同分布。
- 独立性：$X_1, X_2, \ldots, X_n$ 相互独立。

这样的样本叫做**简单随机样本**，我们会得到一些独立同分布的**随机变量**，用大写字母表示。实际得到的数值叫做**观察值**或者**样本值**，用小写字母表示。它们之间的意义区分是明确的。

因此,若总体 X 的累积分布函数为 $F(x)$,则一个简单随机样本 $X_1, X_2, \ldots, X_n$ 的联合分布函数为：

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

统计量则是**样本**的一个随机变量函数。注意这句话的意思是：它和样本有关，而且只和样本有关，不和其他未知参数相关。

设 $X_1, \ldots, X_n$ 是一组样本，$Y_1, \ldots, Y_n$ 是容量相同的另一组样本。其中 X_i 独立同分布，来自总体 X；Y_i 独立同分布，来自总体 Y。对于 $i \neq j$，X_i, Y_j 独立。或者说，(X_i, Y_i) 是来自总体 (X, Y) 的一组样本。

在这里我们有一些经典统计量。并且由于它们本性上是随机变量，我也可以得到它们的数字特征。我们这里假设它们来自的总体 X, Y 各自都有均值和方差，$E(X) = \mu_x, E(Y) = \mu_y, D(X) = \sigma_x^2, D(Y) = \sigma_y^2, Cov(X, Y) = \rho\sigma_x\sigma_y$。

样本均值　样本均值定义为：

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

样本均值的均值和原分布的均值正好一致

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

样本均值的方差则和样本量有一些有趣的关系：

$$D(\overline{X}) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{\sigma_x^2}{n}$$

可以很直观地理解：我们抽样的样本容量越大，其平均偏差大概就会更小。

对于两个样本均值，它们的协方差也是如此：

$$\begin{aligned} Cov(\overline{X}, \overline{Y}) &= Cov\left(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, Y_j) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Cov(X_i, Y_i) \\ &= \frac{\rho\sigma_x\sigma_y}{n} \end{aligned}$$

上面的推导全都基于它们独立同分布的假设，我们后面不再强调。

样本方差、标准差　样本方差和标准差定义为：

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \quad S_x = \sqrt{S_x^2}$$

注意这里分母是 $n-1!$（感叹号并不表示阶乘）这不是没有道理的。样本方差的均值是这样的：

$$
\begin{aligned}
E(S^2) &= \frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\overline{X})^2\right)\\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}((X_i-\mu_x)-(\overline{X}-\mu_x))^2\right)\\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\mu_x)^2-2(\overline{X}-\mu_x)\sum_{i=1}^{n}(X_i-\mu_x)+\sum_{i=1}^{n}(\overline{X}-\mu_x)^2\right)\\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\mu_x)^2-2(\overline{X}-\mu_x)\sum_{i=1}^{n}(X_i-\mu_x)+2(\overline{X}-\mu_x)\sum_{i=1}^{n}(\overline{X}-\mu_x)-\sum_{i=1}^{n}(\overline{X}-\mu_x)^2\right)\\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\mu_x)^2-2(\overline{X}-\mu_x)\sum_{i=1}^{n}(X_i-\overline{X})-\sum_{i=1}^{n}(\overline{X}-\mu_x)^2\right)\\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\mu_x)^2-2(\overline{X}-\mu_x)(\sum_{i=1}^{n}X_i-n\overline{X})-\sum_{i=1}^{n}(\overline{X}-\mu_x)^2\right)\\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\mu_x)^2-\sum_{i=1}^{n}(\overline{X}-\mu_x)^2\right)\\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}E\left((X_i-\mu_x)^2\right)-nE\left((\overline{X}-\mu_x)^2\right)\right]\\
&= \frac{1}{n-1}(n\sigma_x^2-n\frac{\sigma_x^2}{n})=\sigma_x^2
\end{aligned}
$$

这个分母的 $n-1$ 是为了让样本方差的均值和总体的方差一致，这很重要。

样本方差的方差不是什么简单东西，这里只给出一个式子。如果总体分布的四阶原点矩 μ_4 存在，

$$D(S^2)=\frac{1}{n}\left(\mu_4-\frac{n-3}{n-1}\sigma^4\right)$$

而对于正态总体而言，$\mu_4=3\sigma^4$，$D(S^2)=\frac{2\sigma^4}{n-1}$。我们会在卡方分布处想起它。

如果单把右边的和式拿出来，我们称之为**离差平方和**：

$$S_{xx}=\sum_{i=1}^{n}(X_i-\overline{X})=(n-1)S_x^2$$

样本协方差　对于混合起来的量，我们也有定义样本协方差：

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

这里分母仍然是 $n-1$，仍然是为了均值的无偏性，即 $E(S_{xy}) = \rho\sigma_x\sigma_y$，证明思路和刚才的求和一致：

$$\begin{aligned}
E(S_{xy}) &= \frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}\left((X_i - \mu_x) - (\overline{X} - \mu_x)\right)\left((Y_i - \mu_y) - (\overline{Y} - \mu_y)\right)\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i - \mu_x)(Y_i - \mu_y)\right) \\
&\quad - \frac{1}{n-1}E\left((\overline{X} - \mu_x)\sum_{i=1}^{n}(Y_i - \mu_y) - (\overline{Y} - \mu_y)\sum_{i=1}^{n}(X_i - \mu_x)\right) \\
&\quad + \frac{1}{n-1}E\left((\overline{X} - \mu_x)\sum_{i=1}^{n}(\overline{Y} - \mu_y) + (\overline{Y} - \mu_y)\sum_{i=1}^{n}(\overline{X} - \mu_x)\right) \\
&\quad - \frac{1}{n-1}E\left(\sum_{i=1}^{n}(\overline{X} - \mu_x)(\overline{Y} - \mu_y)\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i - \mu_x)(Y_i - \mu_y) - \sum_{i=1}^{n}(\overline{X} - \mu_x)(\overline{Y} - \mu_y)\right) \\
&= \frac{1}{n-1}(n\rho\sigma_x\sigma_y - n\frac{\rho\sigma_x\sigma_y}{n}) \\
&= \rho\sigma_x\sigma_y
\end{aligned}$$

注意：虽然它记号类似 S_{xx}，但是完全不同！

类似地，不除以 $n-1$ 的统计量叫做**样本离差积和**：

$$SP_{xy} = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = (n-1)S_{xy}$$

类似地，我们也可以定义**样本相关系数**，又叫 Pearson 相关系数：

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

当然我们也可以不只限于两组样本，我们可以像协方差矩阵一样将其推广，得到到任意有限维**样本协方差矩阵**和**样本相关系数矩阵**。

样本矩　刚才在维度上做了推广，下面在次数上做推广。

我们可以定义**样本 k 阶原点矩**：

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

由于 X_i^k 都是独立同分布的，应用样本均值的结论，就有 $E(A_k) = \mu_k$。其中 μ_k 为总体的 k 阶原点矩。

次序统计量　还有一些次序统计量，即把样本排序后得到的统计量。设样本 $X_1, \ldots, X_n$ 排序后得到 $X_{(1)} \le \cdots \le X_{(n)}$。

它们每一个拿出来都可以作为统计量，其中 $X_{(k)}$ 叫做**第 k 个次序统计量**。如果总体的累积分布函数是 F，则

$$\begin{aligned} F_{X_{(k)}}(x) &= P(\text{至少有 } k \text{ 个观察值} \le x) \\ &= \sum_{j=k}^{n} \binom{n}{j} (F(x))^j (1-F(x))^{n-j} \end{aligned}$$

这是**不完全 Beta 函数**，这里我不打算写出这种表达。如果总体的概率密度 f 也存在：

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} (F(x))^{k-1} (1-F(x))^{n-k} f(x)$$

这个式子推导的大概是：恰有 1 个数在 $(x, x+dx)$ 内概率为 $f(x)dx$，恰有 $k-1$ 个数在 $< x$ 概率为 $(F(x))^{k-1}$，恰有 $n-k$ 个数 $> x$ 概率为 $(1-F(x))^{n-k}$。这里的 $\frac{n!}{(k-1)!(n-k)!}$ 其实就是多项式系数 $\binom{n}{k-1\ 1\ n-k}$。

也有一个有趣的事实：如果次序统计量各不相同（连续情况下这是几乎必然的），则所有次序统计量的联合概率密度是 $n!\prod_{i=1}^{n} f(x_i)$，$(x_1 < x_2 < \cdots < x_n)$，比无序情况多了个全排列。

这个公式显然也包含了最特别的那两个次序统计量**样本最小值**和**样本最大值**。

$$X_{(1)} = \min(X_1, \ldots, X_n), \quad X_{(n)} = \max(X_1, \ldots, X_n),$$

这两个统计量的累积分布函数前面已经给出过了，这里也可以代入上面的公式验证一下。

样本极差：

$$R = X_{(n)} - X_{(1)}$$

样本中位数:

$$M_e = \begin{cases} X_{(\frac{n+1}{2})} & n = 2k+1 \\ \frac{1}{2}X_{\frac{n}{2}} + \frac{1}{2}X_{\frac{n}{2}+1} & n = 2k \end{cases}$$

样本 p 分位数：它其实没有一个统一的符号。这里的 p 需要满足 $0 < p < 1$。np 不为整数时取 $X_{(\lfloor np \rfloor)}$，否则取 $X_{(\lfloor np \rfloor)}$ 和 $X_{(\lfloor np \rfloor+1)}$ 的均值。样本中位数即样本 $\frac{1}{2}$ 分位数。

上面这些量没有特别简单的统一公式，必须根据特定的分布进行推导。其基础是任意位置次序统计量的分布。

8.2 χ^2 分布

卡方分布大概是抽样分布中最重要的分布之一。设 $X_1, \ldots, X_n$ 是独立的**标准**正态分布，令 $Y = \sum_{i=1}^n X_i^2$，则称 $Y \sim \chi^2(n)$，即随机变量 Y 服从**自由度**为 n 的卡方分布。所谓自由度就是独立变量个数。从这个定义来看它就具有可加性。

它的概率密度函数是:

$$f_Y(y) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} e^{-\frac{y}{2}} & y > 0 \\ 0 & y \le 0 \end{cases}$$

这一公式可以用多种方式验证。一种方法是我前面给出的概率密度函数的公式：联合概率密度

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}}$$

考虑 $g(x_1, \ldots, x_n) = \sum_{i=1}^n x_i^2 = y$，则

$$\nabla g(x_1, \ldots, x_n) = (2x_1, 2x_2, \ldots, 2x_n)$$

$$|\nabla g(x)| = 2\sqrt{\sum_{i=1}^n x_i^2} = 2\sqrt{y}$$

每个 y 对应的是 n 维球面 S^{n-1}，半径 $r = \sqrt{y}$。

由余面积公式，

$$f_Y(y) = \int_{g(\mathbf{x})=y} \frac{f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)}{|\nabla g(\mathbf{x})|} \, d\sigma(\mathbf{x})$$

由对称性，密度在球面上处处一致；故可化为球面积分：

$$f_Y(y) = \frac{1}{2\sqrt{y}} \cdot S_{n-1}(r) \cdot \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{y}{2}}$$

其中 $S_{n-1}(r)$ 为 $n-1$ 维球面面积，半径 $r = \sqrt{y}$。已知

$$S_{n-1}(r) = \frac{2\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} r^{n-1} = \frac{2\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} y^{\frac{n-1}{2}}$$

于是整理化简得：

$$\begin{aligned}
f_Y(y) &= \frac{1}{2\sqrt{y}} \cdot \frac{2\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} y^{\frac{n-1}{2}} \cdot \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{y}{2}} \\
&= \frac{1}{2\sqrt{y}} \cdot \frac{2\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} y^{\frac{n-1}{2}} \cdot (2\pi)^{-\frac{n}{2}} e^{-\frac{y}{2}} \\
&= \frac{1}{2\sqrt{y}} \cdot \frac{2\pi^{\frac{n}{2}} y^{\frac{n-1}{2}}}{(2\pi)^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} e^{-\frac{y}{2}} \\
&= \frac{1}{2\sqrt{y}} \cdot \frac{2y^{\frac{n-1}{2}}}{\Gamma\left(\frac{n}{2}\right)} e^{-\frac{y}{2}} \\
&= \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{\frac{n}{2}}} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}, \qquad y > 0
\end{aligned}$$

这就是 χ^2_n 分布的概率密度函数表达式。当然我们在计算时可以先抛掉常系数，得到 $Cy^{\frac{n}{2}-1}e^{-\frac{y}{2}}$，然后归一化计算出常系数 C。

另一种方法是用特征函数。对于自由度为 1 的卡方分布，它就是正态的平方，其概率密度很容易通过变量代换 $Z = X^2$ 得到：

$$f_Z(z) = \begin{cases} \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} e^{-\frac{x}{2}} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

它的特征函数是：

$$\phi_Z(t) = (1 - 2it)^{-\frac{1}{2}}$$

卡方分布可加性在这里也显而易见地体现出来了。对于刚才的卡方变量 $Y \sim \chi^2(n)$，由卷积公式就有

$$\phi_Y(t) = (1 - 2it)^{-\frac{n}{2}}$$

眼尖的朋友会发现，这就是个变量代换后的 $\Gamma(\frac{n}{2},\frac{1}{2})$，这就是自由度为 n 的卡方分布了。

坏消息是，我们没有其累积分布函数的解析式，但这不妨碍我们给出数值。

自由度为 n 的卡方分布的均值是什么？n 个自由度为 1 的卡方分布均值加起来。自由度为 n 的卡方分布的方差是什么？n 个自由度为 1 的卡方分布方差加起来。令 $X\sim N(0,1), Z=X^2$，或者说 $Z\sim\chi^2(1)$：

$$E(Z)=E(X^2)=1,\quad D(Z)=D(X^2)=E(X^4)-(E(X^2))^2=3-1=2$$

于是对于 $Y\sim\chi^2(n)$，$E(Y)=n, D(Y)=2n$。

8.3 t 分布

另一个重要的抽样分布是 t 分布。它又叫学生氏分布，有一段有趣的小故事。设 $X\sim N(0,1)$，$Y\sim\chi^2(n)$，且 X 与 Y 相互独立。令

$$T=\frac{X}{\sqrt{\frac{Y}{n}}}$$

则称 $T\sim t(n-1)$，即自由度为 n 的 t 分布。

对 $\phi(x,y)=\frac{x}{\sqrt{\frac{y}{n}}}$ 用用前面的随机变量函数公式，就能计算其概率密度。

$$\nabla\phi(x,y)=\left(\frac{\sqrt{n}}{\sqrt{y}},-\frac{x\sqrt{n}}{2y^{\frac{3}{2}}}\right)$$

$$|\nabla\phi|=\sqrt{\frac{n}{y}+\frac{x^2n}{4y^3}}$$

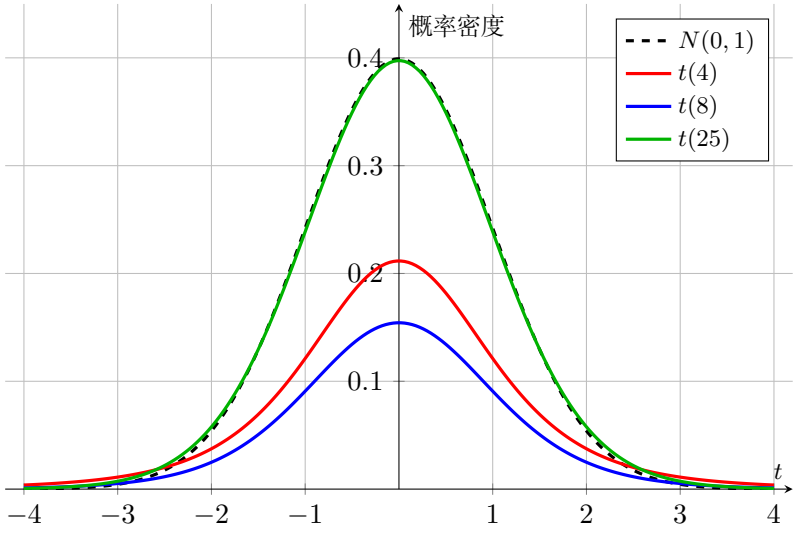
等值线 $\phi(x,y)=t$ 可以参数化为 $x=t\sqrt{\frac{y}{n}}$，$\left(\frac{dx}{dy}\right)^2=\frac{t^2}{2ny}$。

$$
\begin{aligned}
f_T(t) &= \int_{\phi(x,y)=t} \frac{f_X(x)f_Y(y)}{|\nabla\phi(x,y)|}d\sigma(x,y) \\
&= \int_0^{+\infty} \frac{f_X(x)f_Y(y)}{\sqrt{\frac{n}{y}+\frac{x^2 n}{4y^3}}}\sqrt{1+\frac{t^2}{2ny}}dy \\
&= \int_0^{+\infty} f_X\left(t\sqrt{\frac{y}{n}}\right) f_Y(y)\sqrt{\frac{y}{n}}dy \\
&= \int_0^{+\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2 y}{2n}} \cdot \frac{1}{2^{n/2}\Gamma(n/2)}y^{n/2-1}e^{-y/2} \cdot \sqrt{\frac{y}{n}}\,dy \\
&= \frac{1}{\sqrt{2\pi n}\,2^{n/2}\Gamma(n/2)}\int_0^{+\infty} y^{\frac{n+1}{2}-1}\exp\left[-\frac{y}{2}\left(1+\frac{t^2}{n}\right)\right]dy \\
&= \frac{1}{\sqrt{\pi n}\Gamma(n/2)}\int_0^{+\infty} y^{\frac{n+1}{2}-1}e^{-ay}\,dy, \quad a=\frac{1}{2}\left(1+\frac{t^2}{n}\right) \\
&= \frac{1}{\sqrt{\pi n}\Gamma(n/2)}\frac{\Gamma\left(\frac{n+1}{2}\right)}{a^{(n+1)/2}} \\
&= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma(n/2)} \cdot \left(1+\frac{t^2}{n}\right)^{-(n+1)/2}
\end{aligned}
$$

很艰难，但是完全是套公式和凑 Γ 积分。可见这个公式是很强大的。

这个式子告诉我们一个事情，即 t 分布是对称的。另一件事是，$n \to \infty$ 时，它逐点收敛于标准正态分布。具体证明需要 Stirling 公式，在此不作证明。我们一般在 $n > 30$ 时就可以用正态近似表示了。如图

标准正态分布与 t 分布对比（自由度 =4,8,25）



它的累积分布函数也是没有解析式的，我们需要数值计算。

t 分布的 k 阶矩存在当且仅当 $k < n$。特别地，均值 $E(T) = 0 \quad (n > 1)$，方差 $D(T) = \frac{n}{n-1} \quad (n > 2)$。$n \leq 2$ 时方差不存在，这大致在说 t 分布比正态分布具有更厚的尾部。

8.4 F 分布

F 分布由统计学家费希尔(R.A. Fisher)得名，主要用于比较两个正态总体的方差，也是方差分析和线性回归模型显著性检验的基础。设 $U \sim \chi^2(m)$，$V \sim \chi^2(n)$，且 U 与 V 相互独立。令

$$F = \frac{U/m}{V/n}$$

则称随机变量 F 服从自由度为 (m, n) 的 F 分布，记作 $F \sim F(m, n)$。其中 m 称为分子自由度，n 称为分母自由度。

由定义直接可以知道 $F(m, n)$ 和 $F(n, m)$ 之间有一种对偶关系。若 $F \sim F(m, n)$，则 $\frac{1}{F} \sim F(n, m)$。

仍然从定义，若 $T \sim t(n)$，则 $T^2 \sim F(1, n)$。

$F(m,n)$ 分布的概率密度函数为：

$$f_F(x)=\begin{cases}\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{m/2}\frac{x^{(m/2)-1}}{\left(1+\frac{m}{n}x\right)^{(m+n)/2}} & x>0\\ 0 & x\le 0\end{cases}$$

这一证明又需要我们大费周章用下随机变量函数的公式了。这个函数是

$$\phi(x,y)=\frac{nx}{my}$$

其梯度为

$$\nabla\phi(x,y)=\left(\frac{n}{my},\ -\frac{nx}{my^2}\right).$$

模长为

$$|\nabla\phi|=\sqrt{\left(\frac{n}{my}\right)^2+\left(\frac{nx}{my^2}\right)^2}=\frac{n}{my^2}\sqrt{y^2+x^2}.$$

等值线 $\phi(x,y)=f$ 给出：

$$\frac{nx}{my}=f\quad\Longleftrightarrow\quad x=\frac{mf}{n}y.$$

以 y 为参数，$x=\frac{mf}{n}y$，则

$$\frac{dx}{dy}=\frac{mf}{n}.$$

弧长微元为

$$d\sigma=\sqrt{1+\left(\frac{dx}{dy}\right)^2}\,dy=\sqrt{1+\left(\frac{mf}{n}\right)^2}\,dy.$$

注意到

$$\frac{d\sigma}{|\nabla\phi|}=\frac{\sqrt{1+(mf/n)^2}}{\frac{n}{my^2}\sqrt{y^2+\left(\frac{mf}{n}y\right)^2}}\,dy=\frac{m}{n}y\,dy$$

根据公式，

$$f_F(f)=\int_{\phi(x,y)=f}\frac{f_X(x)f_Y(y)}{|\nabla\phi(x,y)|}\,d\sigma(x,y)$$

代入参数化得

$$
\begin{aligned}
f_F(f) &= \int_0^{+\infty} f_X\left(\frac{mf}{n}y\right) f_Y(y)\frac{m}{n}y\,dy \\
&= \int_0^{\infty} \frac{1}{2^{(m+n)/2}\Gamma(m/2)\Gamma(n/2)} \left(\frac{mf}{n}\right)^{m/2-1} y^{m/2-1+n/2-1} e^{-\frac{y}{2}\left(1+\frac{mf}{n}\right)} \cdot \frac{m}{n}y\,dy \\
&= \frac{1}{2^{(m+n)/2}\Gamma(m/2)\Gamma(n/2)} \left(\frac{mf}{n}\right)^{m/2-1} \frac{m}{n} \int_0^{\infty} y^{(m+n)/2-1} e^{-\frac{y}{2}\left(1+\frac{mf}{n}\right)}\,dy \\
&= \frac{1}{2^{(m+n)/2}\Gamma(m/2)\Gamma(n/2)} \left(\frac{mf}{n}\right)^{m/2-1} \frac{m}{n} \int_0^{\infty} y^{(m+n)/2-1} e^{-ay}\,dy \quad a = \frac{1}{2}\left(1+\frac{mf}{n}\right) \\
&= \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{mf}{n}\right)^{m/2-1} \frac{m}{n} \cdot \left(1+\frac{mf}{n}\right)^{-(m+n)/2} \\
&= \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{m/2} \frac{f^{m/2-1}}{\left(1+\frac{m}{n}f\right)^{(m+n)/2}}
\end{aligned}
$$

这个随机变量函数公式确实很夸。

看着式子画个图，F 分布的密度图形是右偏的。随着自由度 m 和 n 的增大，分布逐渐对称并趋近正态分布。

8.5 关于统计量分布的重要定理

在开始之前，我们先给出一些常用的术语和值的定义。

上侧概率和上分位数 **上侧概率**指的是随机变量 X 的取值大于某个特定值的概率。也就是 $P(X > x_0) = 1 - F(x_0)$。

上分位数则反过来了，是指对于某个 $0 < p < 1$，使 $P(X > x_p) = p$ 的 x_p。严格来说，有时没有这种值，我们就取为 $\inf\{x \in \mathbb{R} : P(X > x) \geq p\}$。

上分位数常被用作临界值。正如前面几个抽样分布的累积分布函数无解析式一样，其上分位数也没有解析式子。对于这些重要的分布，其上分位数都有特定的记号。

标准正态分布 $N(0,1)$

- **记号**：z_α
- **定义**：$P(Z > z_\alpha) = \alpha$，或等价地 $P(Z \leq z_\alpha) = 1 - \alpha$
- **对称性**：由标准正态分布关于 0 对称，有 $z_{1-\alpha} = -z_\alpha$

- **示例**：$z_{0.05} \approx 1.645$，$z_{0.025} \approx 1.96$

t 分布（自由度 n）

- **记号**：$t_\alpha(n)$
- **定义**：$P(T > t_\alpha(n)) = \alpha$
- **对称性**：由 t 分布关于 0 对称，有 $t_{1-\alpha}(n) = -t_\alpha(n)$
- **示例**：
  - $t_{0.05}(10) \approx 1.812$
  - $t_{0.95}(10) = -t_{0.05}(10) \approx -1.812$

χ^2 分布（自由度 n）

- **记号**：$\chi^2_\alpha(n)$
- **定义**：$P(\chi^2 > \chi^2_\alpha(n)) = \alpha$
- **示例**：$\chi^2_{0.05}(5) \approx 11.07$

F 分布（自由度分别为 m, n）

- **记号**：$F_\alpha(m, n)$
- **定义**：$P(F > F_\alpha(m, n)) = \alpha$
- **性质**：$F_{1-\alpha}(m, n) = \dfrac{1}{F_\alpha(n, m)}$
- **示例**：$F_{0.05}(3, 10) \approx 3.71$

z 检验基础　由于中心极限定理的存在，我们遇到的很多随机总体都可以认为是服从正态分布的。如果总体 $X \sim N(\mu, \sigma^2)$，那么我们有样本均值 $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$。这个结论是简单的，因为独立正态变量之和仍然是正态变量。按我们前面的样本均值的均值方差推导，就得到了这个结果。

更进一步，我们可以把它标准化，得到：

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

这无疑对我们的计算来说是很方便的。

如果我们想比对两个**独立**的正态总体 $X \sim N(\mu_1, \sigma_1^2)$ 和 $Y \sim N(\mu_2, \sigma_2^2)$ 均值之差，那么我们也可以给出类似的构造：因为这两个变量的差也是正态变量。$X - Y \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$，$\overline{X} - \overline{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n})$。

于是标准化一下就得到：

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0, 1)$$

t 检验基础 不过刚才的情况还是太理想了。一般来说，我们无从知道随机总体的方差，我们必须用 χ^2 分布猜测方差，从而用 t 分布做猜测。

我们先给出我们要用到的核心定理：

定理 11. *设 $X_1, \ldots, X_n$ 独立，且均服从 $N(\mu, \sigma^2)$，样本均值为 $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$，样本方差为 $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$。则有：*

1. *$\bar{X}$ 与 S^2 相互独立。*
2. *$\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$。*

证明. 我们用多元独立同方差正态分布的正交变换性质来完成这个证明。

令 $\mathbf{X} = (X_1, \ldots, X_n)^T$，则 $\mathbf{X} \sim N_n(\mu \mathbf{1}_n, \sigma^2 I_n)$。其中 $\mathbf{1}_n = (1, \ldots, 1)^T$

我们的样本方差是一些平方和，样本均值是一个线性函数。我希望把样本方差表示成各自独立的正态变量平方和（这便成了卡方分布），把均值直接表示成一个和前面的变量都独立的正态变量。我们要从这个独立同方差多元正态分布得出另一组独立的正态变量，方法就是做正交变换。正交变换还有令一个好处，即保持内积，于是变换后变量的平方和仍然不变。

$\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i = \frac{1}{n}\mathbf{1}_n^T \mathbf{X}$。这提示我们这个正交变换的一个分量是什么了。只需要把 $(\frac{1}{n}, \ldots, \frac{1}{n})^T$ 标准化一下得到 $\mathbf{v} = (\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})^T$ 就好了。

剩下的分量是什么，似乎并不重要。我们只需要补全一组包含 $\mathbf{v}$ 的规范正交基就好了。我们把这样得到的正交变换叫做 Q。令 $\mathbf{Y} = (Y_1, \ldots, Y_n)^T = Q\mathbf{X}$，则有

1. $Y_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i = \sqrt{n}\overline{X}$。
2. $\mathbf{Y} \sim N_n(\mu Q\mathbf{1}_n, \sigma^2 Q I_n Q^T) = N_n(\mu Q\mathbf{1}_n, \sigma^2 I_n)$，仍然得到独立同方差的多元独立正态分布。
3. $\mu Q\mathbf{1}_n = (\mu\sqrt{n}, 0, \ldots, 0)^T$。

接下来就该处理 S^2 了。

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - n\overline{X}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}Y_i^2 - Y_1^2\right) \\ &= \frac{1}{n-1}\sum_{i=2}^{n}Y_i^2 \end{aligned}$$

功德圆满。现在 $\overline{X}$ 和 S^2 一点不沾边了。

后半部分的结论几乎就是直接的了。$\tilde{Y} = (Y_2, \ldots, Y_n) \sim N_{n-1}(\mathbf{0}_{n-1}, \sigma^2 I_{n-1})$，独立标准正态分布乘上个 σ^2。于是

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

□

这和 t 分布有什么关系呢？注意前面这个式子 $\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$ 里，σ^2 是要需要猜的，而刚才的式子告诉我们要用卡方分布猜。放在一起看：

$$\begin{aligned} \frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} &\sim N(0,1) \\ \frac{(n-1)S^2}{\sigma^2} &\sim \chi^2(n-1) \end{aligned}$$

t 分布就来了：

$$\frac{\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{\frac{\frac{(n-1)S^2}{\sigma^2}}{n-1}}} = \frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim t(n-1)$$

在方差未知的时候，比对两个独立正态变量的均值同样可以用 t 分布，道理是一样的。设样本 $X_1, \ldots, X_m$ 来自总体 X，方差为 σ_1^2；设样本 $Y_1, \ldots, Y_n$ 来自总体 Y，方差为 σ_2^2。

我们有

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}\sim N(0,1)$$

但是类比前面的式子，它的方差该怎么办呢？这需要一个小巧思。我们使用一个样本合并方差 S_W^2 来做到这一点，它可以反映两边合在一起的方差，下面先介绍引入它的动机。

我们知道：

$$\frac{(m-1)S_1^2}{\sigma_1^2}\sim\chi^2(m-1)$$
$$\frac{(n-1)S_2^2}{\sigma_2^2}\sim\chi^2(n-1)$$

二者独立，所以由可加性：

$$\frac{(m-1)S_1^2}{\sigma_1^2}+\frac{(n-1)S_2^2}{\sigma_2^2}\sim\chi^2(m+n-2)$$

我们就可以构造出 t 分布了。不过有一个问题是：$\sigma_1\neq\sigma_2$ 时，式子非常丑陋。我们先来考虑 $\sigma_1=\sigma_2=\sigma$ 的情况：

$$Z=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}}\sim N(0,1)$$
$$W=\frac{1}{\sigma^2}\left((m-1)S_1^2+(n-1)S_2^2\right)\sim\chi^2(m+n-2)$$

则

$$\frac{Z}{\sqrt{\frac{W}{m+n-2}}}=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{(m-1)S_1^2+(n-1)S_2^2}{m+n-2}}\sqrt{\frac{1}{m}+\frac{1}{n}}}\sim t(m+n-2)$$

我们把 $S_W^2=\frac{(m-1)S_1^2+(n-1)S_2^2}{m+n-2}$ 叫做样本合并方差。刚才的式子更紧凑一点就是：

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\frac{1}{m}+\frac{1}{n}}}\sim t(m+n-2)$$

那方差不齐怎么办呢？这个便唤作 Welch t 检验。如果方差不齐的话，刚才式子里的 σ_1,σ_2 很难被消掉，我们只能做近似得到一个 t 分布。这里给出结论而不作证明。

定理 12. 设样本 $X_1,\ldots,X_m$ 来自总体 X，方差为 σ_1^2；设样本 $Y_1,\ldots,Y_n$ 来自总体 Y，方差为 σ_2^2。则

$$\hat{\nu}\frac{S_X^2/m+S_Y^2/n}{\sigma_1^2/m+\sigma_2^2/n}\approx\chi^2(\hat{\nu})$$

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_X^2}{m}+\frac{S_Y^2}{n}}}\approx t(\hat{\nu})$$

其中 $\hat{\nu}$ 叫做近似自由度，式子为：

$$\hat{\nu}=\frac{\left(\frac{s_X^2}{m}+\frac{s_Y^2}{n}\right)^2}{\frac{(s_X^2/m)^2}{m-1}+\frac{(s_Y^2/n)^2}{n-1}}$$

$\hat{\nu}$ 不一定是整数。不过我们完全可以把 t 分布的式子延拓到非整数的情况去。

f 检验基础 如果我们想要比对两个不同的正态总体的方差，我们就需要用到 f 分布了。设样本 $X_1,\ldots,X_m$ 来自总体 X，方差为 σ_1^2；设样本 $Y_1,\ldots,Y_n$ 来自总体 Y，方差为 σ_2^2。

我们知道：

$$\frac{(m-1)S_1^2}{\sigma_1^2}\sim\chi^2(m-1)$$
$$\frac{(n-1)S_2^2}{\sigma_2^2}\sim\chi^2(n-1)$$

于是

$$\frac{\frac{(m-1)S_1^2}{\sigma_1^2}}{\frac{(n-1)S_2^2}{\sigma_2^2}}\frac{n-1}{m-1}=\frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2}\sim F(m-1,n-1)$$

9 参数估计

参数估计是这样一个问题：我们已知总体的分布，但是不知道其具体参数。我们需要从统计量的观察值来猜出它。

下面给出一些经典估计方法。

9.1 点估计

点估计是一种一发入魂的估计，我们直接猜我们认为最正确的那个数。**估计量**就是一个特殊的统计量，或者说关于样本的随机变量函数，它给出我们对某个参数的估计。对于参数 θ，我们一般计其估计量为盖了帽了的 $\hat{\theta}$。注意，θ 是一个具体数值，$\hat{\theta}$ 是一个统计量。

不过猜测肯定有好有坏，我们有这样几种方法来评判估计的好坏：

弱相合性 如果随着样本容量的增大，对于任意可能的真实值 $\theta \in \Theta$，估计量越来越接近真实参数值（依概率收敛），即

$$\forall \epsilon > 0, \quad \lim_{n\to\infty} P(|\hat{\theta} - \theta| < \epsilon) = 1$$

则称该估计量是相合的。这意味着样本容量越大，猜得越准。不相合的估计量还有什么存在的必要呢（悲）。

当然还可以有强相合性，即几乎必然收敛。

无偏性 我这个估计量显然是以均值接近真实参数值为好吧。对于任意可能的参数值 $\theta \in \Theta$

$$E(\hat{\theta}) = \theta$$

则称估计量为无偏的。

有效性 显然对于一个无偏估计量来说，估计量的分布更靠近均值更好吧。如果说对于任意可能的参数 $\theta \in \Theta$，

$$D(\hat{\theta}_1) \le D(\hat{\theta}_2)$$

就可以说 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效。如果过来也成立，可以视为一样有效；如果反过来不成立，就可以说 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 严格有效。

9.1.1 矩估计

样本矩是好东西。前面的大数定律告诉我们，它们会**依概率收敛**到对应阶矩（如果存在）。所以如果我们可以把参数表示成矩的连续可测函数，那我们就赢了。

如果说 $\theta = f(\mu_1, \ldots, \mu_k)$，那么令

$$\hat{\theta} = f(A_1, \ldots, A_k)$$

那么根据

定理 13 (弱连续映射定理)**.** *设 $X_1, X_2, \ldots$ 是一列随机变量，f 是一个连续可测函数。若 $X_n \xrightarrow{P} X$ 则*

$$f(X_n) \xrightarrow{P} f(X)$$

我们有

$$\hat{\theta} \xrightarrow{P} f(\mu_1, \ldots, \mu_k) = \theta$$

相合性就有了。无偏性则不能完全保证，除非 f 是线性函数（注意不是**多线性函数**）：

$$E(\hat{\theta}) = E(f(A_1, \ldots, A_k)) = f(E(A_1), \ldots, E(A_k)) = \theta$$

我们把 $\hat{\theta}$ 叫做 θ 的**矩估计量**，其观察值叫**矩估计值**。

9.1.2 最大似然估计

矩估计只是一个比较粗糙的估计，这里我们将引入一种更为精巧的估计方式。

设总体 X 服从有一个未知参数 θ 的某种离散分布，其概率质量为 $p(x;\theta)$。注意这里有一种区分：变量在分号前，参数在分号后。有人认为这种区分是本质性的，我则认为这个区分基本是名字上的，没那么本质性。设 $\mathbf{X} = (X_1, \ldots, X_n)$ 是来自 X 的样本，则有：

$$P(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{n} p(x_i;\theta)$$

给定一组观察值 $\mathbf{x} = (x_1, \ldots, x_n)$，我们可以定义**似然函数** $L(\mathbf{x};\theta) = \prod_{i=1}^{n} p(x_i;\theta)$。我们给出估计的方式就是，给出让 $L(\mathbf{x};\theta)$ 最大的 $\hat{\theta}$。这种估计就叫**极大似然估计**。

对于连续的情况，也有类似的结论成立。X_i 落在 $[x, x+dx]$ 内的概率质量为 $f(x;\theta)dx$，则 $X_1 \times \cdots \times X_n$ 落在 $[x_1, x_1+dx_1] \times \cdots \times [x_n, x_n+dx_n]$ 内的概率质量就是 $\prod_{i=1}^n f(x_i;\theta)dx_i = \left[\prod_{i=1}^n f(x_i;\theta)\right] d\mathbf{x}$，即联合分布在 $\mathbf{x}$ 处的概率密度为 $\prod_{i=1}^n f(x_i;\theta)$。不过这都是废话，在测度视角下看二者没有本质区别。

要解释它，我们有两种方式，一种是频率的，一种是贝叶斯式的。它们的分歧主要在是否将 θ 视为确定的上。我相信在这个问题上贝叶斯式的解释给出了更好更广泛的理解。不过，我这里并不是认可贝叶斯式的认识论，而只是认为贝叶斯式的计算解释方法有更好的泛用性或者一般性。首先我们需要审视一些词语的含义。似然（Likelihood）和概率（Probability）是一对含义相似，都表示某种可能性，但是含义微妙地对偶的词。贝叶斯公式就是这种对偶性淋漓尽致的体现。

贝叶斯式的观点是这样的：在我们对参数一无所知的时候，可以把它视作随机的而不是固定的。前面那种变量和参数的本质性区分就模糊了，取而代之的是一种对偶性关系。我们把似然函数视为一种条件概率分布。这样的写法（不完全形式严格）可能会给人以启发：参数估计问题中，我们希望最大化的是 $p(\theta|\mathbf{x})$，即猜对参数的“可能性”。这种可能性就叫似然性而不是概率。概率是知道分布后变量取值的可能性，似然性是知道变量取值后分布参数的可能性。

（上面几段段话里暂时让概率取一个更狭义一点的说法，从更广义的观点来看，他们的区分就是两种条件概率的区分（有人称为先后验概率）。我们一直在处理条件概率，不过条件在不同东西上面。所以不妨放宽一点，为了更广的一般性。）

似然函数 $L(\mathbf{x};\theta)$ 正是我们在知道分布和参数时的概率，可以理解为 $p(\mathbf{x}|\theta)$。那在什么时候，我们可以认为最大化 $p(\mathbf{x}|\theta)$ 就和最大化 $p(\mathbf{x}|\theta)$ 是一样的呢？贝叶斯公式告诉我们：$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$。在我们这个问题中，观察值应该被认为是已知的，$p(\mathbf{x})$ 应该被视为常数（当然在贝叶斯视角下常数和变量的位置经常发生对易）。于是答案就很简单了：二者等价当且仅当 $p(\theta)$ 是完全均匀的（或者说我们对 θ 什么都不知道）。

这就是对最大似然估计的一种解释。在具体计算时，我们一般只会遇到可导的函数（如果不是的话，要小心了！前面那些概率密度得 0 的特殊情况都需要谨慎对待，比如说指数分布 < 0 部分，均匀分布区间外的部分。）对于可导的函数，我们就有充足的手段来最大化，求导数找驻点。如果我们嫌

$L(\theta)$ 里的乘积不好算，可以求个对数把它变成求和再最大化 $\ln L(\theta)$。

对于多个参数的估计，问题只不过变成了多变量的优化问题罢了，我们只用算梯度就行了。参数有约束的话的情况用个拉格朗日乘数法就好了（虽然一般不会有）。

在相当一般的正则条件下，最大似然估计是相合的，这里不作证明。在更严格的条件下，最大似然估计还**依分布**收敛于正态分布，这里也不证明。

比矩估计更进一步，最大似然估计有不变性的好性质：设 $\hat{\theta}$ 是 θ 的最大似然估计，g 是任意函数（不必是一对一的）。则 $g(\hat{\theta})$ 是 $g(\theta)$ 的最大似然估计。对于任意可能的参数值 θ，有：

$$L(\mathbf{x};) = L(\mathbf{x}; g^{-1}(g()))$$

当 $\theta = \hat{\theta}$ 时，似然函数达到最大，因此 $g(\hat{\theta})$ 使得 $g(\theta)$ 对应的似然函数也达到最大。这个性质非常实用！例如，如果 $\hat{\sigma}^2$ 是方差的最大似然估计，那么 $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ 就是标准差的最大似然估计。

9.2 区间估计

不过我们在做估计的时候，一般不会得到一个估计值就满足了，我们还想知道我们猜对的可能性有多大。不过对于连续分布，单点零测谈不了什么概率，我们必须给出一个估计区间才可以给出概率。这就是区间估计的玩法。我们一般在给定某个概率 p 的情况下，给出一个区间 $(\underline{\theta}, \overline{\theta})$，使得 $P(\underline{\theta} < \theta < \overline{\theta}) \geq p$。事实上这个 $\overline{\theta}$ 和 $\underline{\theta}$ 都应该被视为统计量。不过呢，我们做区间估计的时候，基本都是为了后面做假设检验，所以这里我们一般不是给出 p，而是给出 $\alpha = 1 - p$，即 $p = 1 - \alpha$。这里的 $1 - \alpha$ 被称为**置信水平**。区间 $(\underline{\theta}, \overline{\theta})$ 就叫置信水平 $1 - \alpha$ 下的**置信区间**，$\underline{\theta}, \overline{\theta}$ 分别叫置信水平 $1 - \alpha$ 下的**置信下限**和**置信上限**。

不过实际在做的估计时候，我们会取尽可能窄的区间。在连续情况下会取令 θ 落在置信区间中的概率正好为 $1 - \alpha$ 的区间；在离散情况可能取不到，就只能尽量。并且我们还会让区间的中点尽可能无偏。

不过在估计的时候，我们也会用单侧的区间来估计，即只给出置信上限 $\overline{\theta}$，让 $P(\theta < \overline{\theta}) \geq 1 - \alpha$；或者只给出置信下限 $\underline{\theta}$，让 $P(\theta > \underline{\theta}) \geq 1 - \alpha$。此时的区间 $(-\infty, \overline{\theta})$ 或 $(\underline{\theta}, +\infty)$ 叫做**单侧置信区间**；$\overline{\theta}$ 叫**单侧置信上限**，$\underline{\theta}$ 叫**单侧置信下限**。

在区间估计中，我们想构造一个置信区间，其中上下限都是统计量。但是这就是问题了，如果仅凭统计量，不依赖于和未知参数有关的具体分布，我要怎么给出这个区间？答案是，我们需要把某个未知参数纳入进来，构造一种分布已知的量，即**枢轴量**。

设 $\mathbf{X} = (X_1, \ldots, X_n)$ 是来自总体 $F(x;\theta)$ 的样本，θ 是待估参数（我们也可以把它扩展为参数向量）。

如果一个函数 $G(\mathbf{X};\theta)$ 是依赖于参数 θ 和样本 $\mathbf{X}$ 的函数（和别的参数没有关系，并且必须和这个参数有关系！），并且其分布完全已知且与 θ（和任何其他**未知**参数）无关，则称 $G(\mathbf{X};\theta)$ 为枢轴量。

注意它不是统计量，它依赖于那个待估的未知参数！有了枢轴量，我们就可以很容易地根据它的分布来列出

$$P(\underline{g} < G(\mathbf{X};\theta) < \overline{g}) = 1 - \alpha$$

然后就能给出 $(\underline{g}, \overline{g})$，通过变换得到等价的 $(\underline{\theta}, \overline{\theta})$。

我们下面就来给出一些对正态总体进行区间估计的例子，它们的枢轴量基本都可以通过上一节抽样分布中的定理得出。下面的例子中我都采用置信水平 $1-\alpha$。

9.2.1 单个正态总体 $N(\mu, \sigma^2)$

这里我们假设样本容量是 n。

μ 待估，σ^2 已知 这里我们构造的枢轴量是

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

它的分布已知（与未知参数无关），包含待估参数 μ，包含已知参数 σ，所以可以拿来做枢轴量。所以我们给出的区间：

$$z_{1-\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}$$

上分位数作为临界值起到了很好的作用。对于正态这种对称分布，我们有 $z_{1-\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}}$。换成关于 μ 的区间就是：

$$\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}} < \mu < \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}$$

单侧区间估计也是类似的，我就不再赘述其推导了。在后面其他的例子中，我也不再会给出单侧置信区间了，因为推导是雷同的。

单侧置信上限：

$$\overline{\mu} = \overline{X} + \frac{\sigma}{\sqrt{n}} z_\alpha$$

单侧置信下限

$$\underline{\mu} = \overline{X} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha} = \overline{X} - \frac{\sigma}{\sqrt{n}} z_\alpha$$

μ 待估，σ^2 未知 方差未知的时候，我们就不能用刚才的正态分布做估计了，我们需要 t 分布。这里的枢轴量是

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$$

所以我们给出区间：

$$t_{1-\frac{\alpha}{2}}(n-1) < \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} < t_{\frac{\alpha}{2}}(n-1)$$

t 分布也是对称的，也有 $t_{1-\frac{\alpha}{2}}(n-1) = -t_{\frac{\alpha}{2}}(n-1)$，我们于是做变换得到：

$$\overline{X} - \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1) < t < \overline{X} + \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1)$$

后面由于推导过于雷同，我只给出枢轴量和置信区间。真正消耗智力的内容在上一节，这一节只是应用一下。

σ^2 待估，μ 未知 这里适用卡方分布进行估计。枢轴量

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

置信区间

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

注意我们的这套记号有一个明显的弊端：容易把自由度参数当成因式，不小心消掉 $n-1$。

9.2.2 两个独立正态总体 $N(\mu_1, \sigma_1^2)$，$N(\mu_2, \sigma_2^2)$

这里我们假设样本容量分别是 n_1，n_2。

$\mu_1-\mu_2$ **待估，** σ_1^2,σ_2^2 **已知**　方差已知的情况一直都很好办，我们可以用正态做估计。枢轴量

$$Z=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1)$$

置信区间

$$\overline{X}-\overline{Y}-z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}<\mu_1-\mu_2<\overline{X}-\overline{Y}+z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

$\mu_1-\mu_2$ **待估，** σ_1^2,σ_2^2 **未知相等**　这个时候需要用 t 分布估计，枢轴量

$$T=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)$$

其中样本合并方差 $S_W^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$。

置信区间

$$\overline{X}-\overline{Y}-t_{\frac{\alpha}{2}}S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}<\mu_1-\mu_2<\overline{X}-\overline{Y}+t_{\frac{\alpha}{2}}S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

σ_1^2/σ_2^2 **待估，** μ_1,μ_2 **未知**　这里用到 F 分布进行估计。枢轴量

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

置信区间

$$\frac{S_1^2}{S_2^2}\frac{1}{F_{\frac{\alpha}{2}}(n_1-1,n_2-1)}<\sigma_1^2/\sigma_2^2<\frac{S_1^2}{S_2^2}\frac{1}{F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)}$$

注意有个倒数别搞反了。

10 假设检验

在统计推断中，除了对未知参数进行估计（点估计和区间估计）之外，另一类核心问题是根据样本信息来判断关于总体分布或参数的某个陈述是否成立。我们会提出一个假设，然后通过统计的手段给出证据决定接受还是拒绝这个假设，这类问题就称为假设检验。

我们这里涉及的假设检验都是**显著性检验**。它是这样的一个过程：首先我们给出一个**原假设** H_0，和一个与它**对立**的**备择假设** H_1。也就是说无论怎样，H_0, H_1 两个命题中都恰好有一个命题是对的，另一个是错的。比如说，$H_0 : \mu = \mu_0$，那么 H_1 必须是 $\mu \neq \mu_0$。

你可能觉得原假设和备择假设选取是完全任意的，把哪个当作原假设都可以。但实际上是有区别的。一般来说，原假设是一个我们尝试去推翻的东西。我们推翻一个东西时，可以利用归谬法，即先假设它是正确的，然后从中引出荒诞的结论。在概率的世界里，没有什么是绝对荒诞的，只有事情发生的概率不同。对于一枚均匀的硬币，连扔一千次都是正面确实是可能的，但如果这件事真的发生了，我们就该怀疑这个假设了。换句话说，如果在原假设成立的情况下，这组观察值的出现是一个几乎不可能的小概率事件，我们就该拒绝原假设。对于什么是几乎不可能的小概率事件，我们可以用一个**显著性水平** α 来量化描述：如果某个事情发生的概率小于 α，我们就视之为一个几乎不可能的小概率事件。对于不同的任务，我们对显著性水平有不同的选取。比如在检验某种药物是否有效时，我们显然要比检验两种冰淇凌谁更畅销使用更苛刻的标准。

一个常见的误区是，我们拒绝和接受原假设不代表证明其真假。我们实际在做的是尝试去拒绝原假设，而接受原假设只是因为没有足够证据拒绝原假设。

那么这就引出了一个问题：我在是否拒绝原假设这件事上肯定有几率犯错，那我犯错的概率是多大？首先我们需要明确一下什么是犯错。错误有两种，一种是我们本来不该拒绝原假设，但是我们却拒绝了，这叫**第 I 类错误**；另一种是我们本来应该拒绝原假设，但是我们却没拒绝，这叫**第 II 类错误**。在我们给出显著性水平 α 之后，这种检验犯第 I 类错误的概率就是：概率小于 α 的某个事件发生的概率。有点绕，但实际上很直接，这一概率就是 α。显著性检验就是预先给出显著性水平以控制犯第 I 类错误的概率。但是在显著性检验中，我们犯第 II 类错误的概率就没这么简单了，我们有时把它叫做 β，把 $1 - \beta$ 称为检验的**功效**。功效是一个和效应大小、参数、

样本容量等诸多都相关的量。我们一般会先设定一个预期的功效，然后按该功效决定样本容量，以达到预期功效。当我们把原假设和备择假设的位置对调之后，第 I 类错误和第 II 类错误其实就反过来了，α 和 β 也就对调了，所以它们的地位并不随便。实际检验时我们一般会遵循一个“无罪推定”原则，即通常将“无效果”、“无差异”设为原假设。

我们一般把原假设成立时，观测到与当前样本一样极端或更极端结果的概率叫做 **p 值**。显然由定义，$\alpha < p$ 时就无法拒绝原假设，$p > \alpha$ 时就要拒绝原假设。显著性水平越小，我们就需要更强力的证据来拒绝原假设。我们必须在试验前就指定 α，在试验后求出 p 值，如果倒过来无异于先射箭再画靶。这里需要着重强调，这个概率并不是 H_0 为假的概率，而是我前面所说的类似 $P(\mathbf{X} = \mathbf{x}|H_0)$ 这样的条件概率。另一个需要澄清的事情是：p 值小不代表实际上的重要效应，有可能只是我们有充足大量的样本。

有一些关于分布参数的特殊形式假设，它们各自有各自的名字和检验方法。不要害怕，求 p 值这一步的方法就是上一节做区间估计的方法。

对于

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0$$

这类假设，H_1 表示参数的取值可能大于 μ_0 或者小于 μ_0。我们把 H_1 称为**双边备择假设**，我们的检验方法称为**双边假设检验**。实际上我们做的就是双侧区间估计。

对于

$$H_0 : \mu \geq \mu_0, \quad H_1 : \mu < \mu_0$$

这类假设，我们称 H_1 为**左边备择假设**，我们的检验方法称为**左边假设检验**。实际上我们做的就是只给出置信**下**限的单侧区间估计。

对于

$$H_0 : \mu \leq \mu_0, \quad H_1 : \mu > \mu_0$$

这类假设，我们称 H_1 为**右边备择假设**，我们的检验方法称为**右边假设检验**。实际上我们做的就是只给出置信**上**限的单侧区间估计。

上面两种检验合称**单边假设检验**。我们必须在试验之前就给出检验的方向，不能看到数据的偏向性之后再选择单边检验。

具体检验中，我们做区间估计用到的枢轴量在这里就被称为**检验统计量**。我们根据检验统计量 T 的可能取值范围，将其划分成两个互斥的区域：

- **拒绝域** R：如果根据样本计算出的统计量观测值 t 落入这个区域，我们就拒绝原假设。
- **接受域** A：如果 t 落入这个区域，我们就不拒绝原假设。

拒绝域和接受域的边界点就叫**临界点**。单侧检验有一个临界点，双侧检验有两个临界点。

10.1 常见的正态总体均值、方差检验法

许多正态总体均值、方差检验方法和上一章是雷同的。和上一章内容没有本质区别的部分被我整理在了表 1。

10.2 基于成对数据的 t 检验

成对数据是这样的和两组独立的数据是有区别的，成对数据可能有每一对之间的耦合关系。不过我们可以认为不同对的数据是独立的。更形式地说，我们可以认为 $(X_1,Y_1),\dots,(X_n,Y_n)$ 是来自总体 (X,Y) 的样本。我们在比较它们的差异值 $D=X-Y$ 时，和两个独立总体的情况有所不同。

在 X 和 Y 独立的条件下，我们知道：

$$\frac{\overline{D}-\delta}{S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)$$

其中 $\delta=\mu_1-\mu_2$，$S_W^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$。

这一结论的基础是在 X,Y 独立的情况下，

$$D(\overline{D})=D(\overline{X})+D(\overline{Y})=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$$

但在成对数据中，X 和 Y 并没有预设的独立性。于是 X,Y 协方差仍然需要考虑。

$$D(\overline{D})=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}-2Cov(\overline{X},\overline{Y})$$

样本容量比较小的时候，我们对 $\overline{D}$ 的分布基本束手无策。但是当样本容量很大时，由中心极限定理就有 $\overline{D}\approx N(\mu_D,\frac{\sigma_D^2}{n})$。于是这又回到了单正态总体的 t 检验问题了。

我们来推导一下它的双边检验。单边检验留给读者。

表 1: 正态总体均值与方差的假设检验

原假设 H_0	**检验统计量**	**拒绝域**
$\mu \le \mu_0$	$Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$	$Z \ge z_\alpha$
$\mu \ge \mu_0$		$Z \le -z_\alpha$
$\mu = \mu_0$		$\lvert Z\rvert \ge z_{\alpha/2}$
$\mu \le \mu_0$	$T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$	$T \ge t_\alpha(n-1)$
$\mu \ge \mu_0$		$T \le -t_\alpha(n-1)$
$\mu = \mu_0$		$\lvert T\rvert \ge t_{\alpha/2}(n-1)$
$\sigma^2 \le \sigma_0^2$	$\chi^2 = \dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$	$\chi^2 \ge \chi^2_\alpha(n-1)$
$\sigma^2 \ge \sigma_0^2$		$\chi^2 \le \chi^2_{1-\alpha}(n-1)$
$\sigma^2 = \sigma_0^2$		$\chi^2 \ge \chi^2_{\alpha/2}(n-1)$ 或$\chi^2 \le \chi^2_{1-\alpha/2}(n-1)$
$\mu_1 - \mu_2 \le \delta_0$	$Z = \dfrac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$	$Z \ge z_\alpha$
$\mu_1 - \mu_2 \ge \delta_0$		$Z \le -z_\alpha$
$\mu_1 - \mu_2 = \delta_0$		$\lvert Z\rvert \ge z_{\alpha/2}$
$\mu_1 - \mu_2 \le \delta_0$	$T = \dfrac{(\bar{X} - \bar{Y}) - \delta_0}{S_W\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1+n_2-2)$	$T \ge t_\alpha(n_1+n_2-2)$
$\mu_1 - \mu_2 \ge \delta_0$		$T \le -t_\alpha(n_1+n_2-2)$
$\mu_1 - \mu_2 = \delta_0$		$\lvert T\rvert \ge t_{\alpha/2}(n_1+n_2-2)$
$\sigma_1^2 \le \sigma_2^2$	$F = \dfrac{S_1^2}{S_2^2} \sim F(n_1-1, n_2-1)$	$F \ge F_\alpha(n_1-1, n_2-1)$
$\sigma_1^2 \ge \sigma_2^2$		$F \le F_{1-\alpha}(n_1-1, n_2-1)$
$\sigma_1^2 = \sigma_2^2$		$F \ge F_{\alpha/2}(n_1-1, n_2-1)$ 或$F \le F_{1-\alpha/2}(n_1-1, n_2-1)$

检验假设

$$H_0 : \mu_D = 0$$
$$H_1 : \mu_D \neq 0$$

我们的检验统计量为

$$T = \frac{\overline{D} - 0}{\frac{S_D}{\sqrt{n}}} \sim t(n-1)$$

检验统计量的观察值为

$$t = \frac{\overline{d}}{\frac{s_d}{\sqrt{n}}}$$

双边的拒绝域就应该是

$$|t| \geq t_{\frac{\alpha}{2}}(n-1)$$

10.3 分布拟合检验

前面的例子都是分布已知但是参数未知的问题。而分布拟合检验处理的假设检验问题是这样的:

$$H_0 : \text{总体 X 的分布为} F(x)$$

这看起来很困难，但是 Pearson 在 1900 年给出了一个精妙的解答:

定理 14 (Pearson 定理)**.** 设随机变量 X 可能取值的空间为 Ω。一组 k 个集合 $\{A_j\}$ 给出了 Ω 的划分，即 $\bigsqcup_{j=1}^{k} A_j = \Omega$。记 $p_j = P(A_j)$。

设 $X_1, \ldots, X_n$ 与 X 同分布并相互独立, 令随机向量 $\mathbf{O} = (O_1, \ldots, O_k)^T$

$$O_j = \#\{X_i \in A_j\}, \quad j = 1, 2, \ldots, k$$

即这些随机变量落在 A_k 内的个数。

令

$$\chi_n^2 = \sum_{j=1}^{k} \frac{(O_j - np_j)^2}{np_j}$$

则有

$$\chi_n^2 \xrightarrow{d} \chi^2(k-1)$$

证明. 首先做一点观察，我们知道 $\sum_{j=1}^{k} p_j = 1$。根据定义我们就知道，我们可以把它当作 n 次有 k 种可能结果的独立试验，$\mathbf{O} \sim Mult(n, k, p_1, \ldots, p_k)$。

回忆下多项分布的性质，有

$$E(O_j) = np_j, \quad Cov(O_i, O_j) = \delta_{ij} p_i - p_i p_j$$

我们最后想要得到一个卡方分布，也就是一些独立标准正态分布的平方和。我们现在构造的这个统计量也是一个平方和。我们不妨设随机向量 $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$，其中 $Z_j = \frac{O_j - np_j}{\sqrt{np_j}}$，就有

$$\chi_n^2 = \sum_{j=1}^{k} Z_j^2$$

我们要做的就是通过某种变换把它变成一些独立正态变量。

但是问题是：正态压根就没从我们前面的式子里出现过。让它出现就必须要中心极限定理。这里我们用到的是一种多元中心极限定理，可以认为是 Laplace-De Moivre 中心极限定理的推广。多项分布在这里就起到了一个类似二项分布的作用。不过多元的结论和一元的区别最大之处在于：它需要考虑诸变量之间的相关性。在这里相关性肯定是存在的：$\sum_{j=1}^{k} p_k = 1$，$\sum_{j=1}^{k} O_j = n$。这个关系明确地对 $\mathbf{Z}$ 写出来就是

$$\sum_{j=1}^{k} \sqrt{p_j} Z_j = \frac{1}{\sqrt{n}} \sum_{j=1}^{k} (O_j - np_j) = \frac{1}{\sqrt{n}} (\sum_{j=1}^{k} O_j - n \sum_{j=1}^{k} p_j) = 0$$

我们用个紧凑的写法，设 $\mathbf{v} = (\sqrt{p_1}, \ldots, \sqrt{p_k})^T$，就有 $\mathbf{v}^T \mathbf{Z} = 0$。$\mathbf{v}$ 在后面仍然有用。

我不证明这个中心极限定理了，总之有 $n \to \infty$ 时：

$$\frac{O - n\mathbf{p}}{\sqrt{n}} \xrightarrow{d} N_k(\mathbf{0}, \Sigma)$$

其中 $\Sigma = Cov(\mathbf{O})$，$\Sigma_{ij} = \delta_{ij} p_i - p_i p_j$。

于是我们也就知道，$n \to \infty$ 时 $\mathbf{Z}$ 也服从多元正态分布。这个分布显然就应该是 $N_k(\mathbf{0}, \Psi)$，其中 $\Psi = Cov(\mathbf{Z})$，$\Psi_{ij} = \delta_{ij} - \sqrt{p_i p_j}$。仔细观察下这个矩阵，这就是 $I - \mathbf{v}\mathbf{v}^T$，这是个沿 $\mathbf{v}$ 方向的正交投影！我们把这个分布先叫做 $\mathbf{Z}_\infty$。

（注意到这其实不是一个正定的协方差矩阵，和我们之前的讨论有些许出入。但我们仍然可以允许这么做，它是一个“退化”的正态分布，即这个

正态变量其实**几乎必然**分布在一个子空间上，没有那么大的自由度。这种形式的正交投影会把 k 维的向量正交投影到一个 $k-1$ 的子空间里。之前的讨论其实仍然适用（线性变换不变性，**部分**变量的独立性），因为谱定理也能处理特征值为 0 的情况。只不过唯一的问题就是我们写不出概率密度函数了，我们需要把它成一个更低维度的多元正态分布来处理。）

我们梦寐以求的标准性仍然没有到来，但它离得不远了，我们可以通过线性变换做到这一点。由线性变换不变性，我们确实可以把 $\mathbf{Z}_\infty$ 通过变换降维得到标准正态向量。我们取 $\operatorname{Im}\Psi$ 的一组规范正交基，组成一个 $k\times(k-1)$ 的矩阵 U。这就是一个赤裸裸的降维，我们其实给 U 添上一列 $\mathbf{v}$ 就得到 Ψ 了，令

$$\mathbf{W}=U^T\mathbf{Z}_\infty$$

则有

$$\mathbf{W}\sim N_{k-1}(\mathbf{0},U^T\Psi U)=N_{k-1}(\mathbf{0},I_{k-1})$$

这是好事啊。于是 $n\to\infty$ 时就有

$$\mathbf{Z}_\infty^T\mathbf{Z}_\infty=\mathbf{W}^T\mathbf{W}\sim\chi^2(k-1)$$

注意，从这里到最终的结论还差一步，并不显然。我们需要连续映射定理告诉我们 $\mathbf{Z}\xrightarrow{d}\mathbf{Z}_\infty$ 时，

$$\chi_n^2=\mathbf{Z}^T\mathbf{Z}\xrightarrow{d}\mathbf{Z}_\infty^T\mathbf{Z}_\infty$$

这样才算结束了：

$$\chi_n^2\xrightarrow{d}\chi^2(k-1)$$

□

那我们怎么应用它呢？注意到定理中的 X_i 其实就是样本，O_j 的观察值就是频率 f_i。p_j 则是根据假设推断出来的。于是就有

$$\sum_{j=1}^{k}\frac{(f_j-np_j)^2}{np_j}\approx\chi^2(k-1)$$

不过这里要注意中心极限定理的条件。首先样本容量需要足够大（$n\geq 50$），其次 np_j 不能太小（$np_j\geq 5$）。如果不满足的话，我们需要适当合并类别。当卡方检验统计量过大时（超过 $\chi_\alpha^2(k-1)$），我们就该拒绝原假设。

10.3.1 分布族拟合检验

真实情况并非如此简单。我们往往连分布的参数也不知道，这时我们连 p_j 也需要猜了。这时我们一般会在假设下先对 p 做一个最大似然估计 $\hat{p}$，然后再照样去做。注意这时要用原始数据而不是分组数据去估计。

当从数据估计 r 个参数时，卡方统计量的极限分布也会发生变化：因为 $\hat{p_j}$ 这时也掺和进来了相关性。这里给出定理而不证明。

定理 15 (Chernoff-Lehmann 定理)**.** *在适当正则条件下，*

$$\sum_{j=1}^{k} \frac{(O_j - n\hat{p}_j)^2}{n\hat{p}_j} \xrightarrow{d} \chi^2(k-r-1)$$

其中 r 是估计的参数个数。

最后我们可能并不拒绝原假设。这时我们接受假设意味着我们认可总体的分布属于某个分布族，但我们不能认为它就是符合某个特定参数的分布。我们可以采用该分布族作为模型，但是还有其他可能同样兼容的分布族，并且其参数值也不确定。

11 方差分析

方差分析（简称 ANOVA）是一种统计方法，用于检验两个或两个以上总体均值之间的差异是否显著。单因素方差分析是指只考虑一个因素（自变量）对观测结果（因变量）的影响。其基本思想是将观测数据的总变异分解为两部分：一是由所研究的因素引起的变异（组间变异），二是由随机误差引起的变异（组内变异）。通过比较这两部分变异的大小来判断该因素对观测结果的影响是否显著。

11.1 单因素试验的方差分析

我们的模型是这样的：假设某因素有 r 个水平，每个水平下进行了 n_i 次独立试验，结果为随机变量 X_{ij}。单因素方差分析的模型为：

$$X_{ij} = \mu_i + \epsilon_{ij} = \mu + \delta_i + \epsilon_{ij}, \quad i = 1, 2, \ldots, r; \quad j = 1, 2, \ldots, n_i$$

其中：

- X_{ij} 服从正态分布 $N(\mu_i, \sigma^2)$，且相互独立。
- μ 为**总平均**。
- δ_i 为第 i 个水平的**效应**，满足 $\sum_{i=1}^{r} n_i \delta_i = 0$。
- ϵ_{ij} 为随机误差，服从正态分布 $N(0, \sigma^2)$。注意这里预设了方差齐性。

它的假设检验也就是这样的：

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_r = \mu$$
$$H_1 \ \exists \mu_i \neq \mu$$

或者等价地：

$$H_0 : \delta_1 = \delta_2 = \cdots = \delta_r = 0$$
$$H_1 \ \exists \delta_i \neq 0$$

这一检验的统计量构造是有挑战性的。为了行文方便，我们先给出一些基本参数和统计量的定义。

$$\text{总样本量：}\ n=\sum_{i=1}^{r} n_i$$

$$\text{总样本均值：}\ \overline{X}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} X_{ij}$$

$$\text{组内样本均值：}\ \overline{X}_i=\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$$

其中关键的一些量是一些平方和：

- **总离差平方和** $S_T=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\overline{X})^2$，反映全部观测值的总变异。
- **组间离差平方和** $S_A=\sum_{i=1}^{r} n_i(\overline{X}_i-\overline{X})^2$，反映因素水平不同引起的变异。
- **组内离差平方和** $S_E=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)^2$，反映随机误差引起的变异。
- 三者关系：$S_T=S_A+S_E$。

我们在下面简单地展开验证下平方分解。

$$\begin{aligned}
S_T &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\overline{X})^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i}[(X_{ij}-\overline{X}_i)+(\overline{X}_i-\overline{X})]^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)^2+\sum_{i=1}^{r}\sum_{j=1}^{n_i}(\overline{X}_i-\overline{X})^2+2\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)(\overline{X}_i-\overline{X}) \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)^2+\sum_{i=1}^{r} n_i(\overline{X}_i-\overline{X})^2 \\
&= S_E+S_A
\end{aligned}$$

有一些朴素的计算化简：

$$\begin{aligned}
S_T &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\overline{X})^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij}^2 - 2\overline{X}\sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij} + \sum_{i=1}^{r}\sum_{j=1}^{n_i}\overline{X}^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij}^2 - 2n\overline{X}^2 + n\overline{X}^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij}^2 - n\overline{X}^2
\end{aligned}$$

以及

$$\begin{aligned}
S_A &= \sum_{i=1}^{r} n_i(\overline{X}_i-\overline{X})^2 \\
&= \sum_{i=1}^{r} n_i\overline{X}_i^2 - 2\overline{X}\sum_{i=1}^{r} n_i\overline{X}_i^2 + \sum_{i=1}^{r} n_i\overline{X}^2 \\
&= \sum_{i=1}^{r} n_i\overline{X}_i^2 - 2n\overline{X}^2 + n\overline{X}^2 \\
&= \sum_{i=1}^{r} n_i\overline{X}_i^2 - n\overline{X}^2
\end{aligned}$$

记总和 $T = n\overline{X}$，组内总和 $T_i = n_i\overline{X}_i$，也可以写成

$$\begin{aligned}
S_T &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij}^2 - \frac{T^2}{n} \\
S_A &= \sum_{i=1}^{r}\frac{T_i^2}{n_i} - \frac{T^2}{n} \\
S_E &= S_T - S_A
\end{aligned}$$

它们有这样的统计特性：

定理 16. 对于组间离差平方和 S_A，组内离差平方和 S_E，总样本均值 $\overline{X}$，有：

- S_A 与 S_E 与 $\overline{X}$ 相互独立

- $\frac{S_E}{\sigma^2} \sim \chi^2(n-r)$
- **在原假设下**，$\frac{S_A}{\sigma^2} \sim \chi^2(r-1)$

证明. 这个命题看起来很像我们对样本方差和样本均值的论述，其证明也是类似的，重点在于构造正交变换。

首先我们给出样本的随机向量 $\mathbf{X} = (X_{11}, \ldots, X_{1n_1}, \ldots, X_{r1}, \ldots, X_{rn_r})^T$。

我们构造的这个正交变换 Q，要让我们得到的 $\mathbf{Z} = Q\mathbf{X}$ 是这样的：

- 第一个分量 Y_1 对应总体均值
- 接下来 $r-1$ 个分量对应组间变异
- 剩下 $n-r$ 个分量对应组内变异

这个构造基本上就是把这些表达式写出来，然后归一化。具体来说，Q 的第一行为

$$\mathbf{q}_1 = (\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$$

接下来的 $r-1$ 行（组间对比方向）：

第 $k+1$ 行（$k = 1, \ldots, r-1$）：

$$\mathbf{q}_{k+1} = (c_{k1}, \ldots, c_{k1}, c_{k2}, \ldots, c_{k2}, \cdots, c_{kr}, \ldots, c_{kr})$$

其中：

$$c_{ki} = \begin{cases} \dfrac{1}{\sqrt{n_1 + \cdots + n_k}} \cdot \sqrt{\dfrac{n_i}{n}} & i \le k \\ -\sqrt{\dfrac{n_1 + \cdots + n_k}{n_{k+1}}} \cdot \sqrt{\dfrac{n_i}{n}} & i = k+1 \\ 0 & i > k+1 \end{cases}$$

剩下的 $n-r$ 行（组内变异方向）：

对于第 i 组（$i = 1, \ldots, r$），在该组内部构造 $n_i - 1$ 个正交向量。记第 i 组的观测位置为 $p_i + 1$ 到 $p_i + n_i$，其中 $p_i = \sum_{j=1}^{i-1} n_j$。

对于 $a = 1, \ldots, n_i - 1$，定义向量 $\mathbf{q}_{r+p_i+a}$ 的第 j 个分量为：

$$(\mathbf{q}_{r+p_i+a})_j = \begin{cases} \dfrac{1}{\sqrt{a(a+1)}} & p_i + 1 \le j \le p_i + a \\ -\dfrac{r}{\sqrt{a(a+1)}} & j = p_i + a + 1 \\ 0 & \text{其他} \end{cases}$$

这样就构造完成了，请自己检验下吧。

于是我们就可以把 $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \sigma^2 I_n)$ 变换成 $\mathbf{Z} \sim N_n(Q\boldsymbol{\mu}, Q\sigma^2 I_n Q^T) = N_n(Q\boldsymbol{\mu}, \sigma^2 I_n)$。它的各分量是独立的!

$\overline{X}$ 就是 $\mathbf{Z}$ 第一个分量，S_A 就是 $\mathbf{Z}$ 后面的 $r-1$ 个分量的平方和，S_E 就是剩下 $n-r$ 各分量的平方和。三者相互独立就证明完成了。剩下就是证明两个卡方分布了。

注意我们对 Q 的定义，**在原假设下**（也就是说一切 X_{ij} 都是独立同分布的），$\boldsymbol{\mu} = \mu \mathbf{1}_n$，由我们对 Q 的定义，$Q\boldsymbol{\mu} = \mu(\sqrt{n}, 0, \ldots, 0)^T$。所以我们取走 $\mathbf{Z}$ 的后面 $r-1$ 和 $n-r$ 个分量，它们就服从独立同方差的正态分布 $N_{n-1}(\mathbf{0}, \sigma^2 I_{n-1})$。

于是这两个卡方分布也很显然了。不过事实上，在原假设不为真的情况下，$\frac{S_E}{\sigma^2} \sim \chi^2(n-r)$ 也是成立的。这只需要一点小小的分析：

$$\begin{aligned}\frac{S_E}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2 \\ &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(\frac{\epsilon_{ij}}{\sigma}\right)^2\end{aligned}$$

我们知道 $\frac{\epsilon_{ij}}{\sigma} \sim N(0,1)$，并且所有随机误差相互独立。所以平方和加在一起就是 $\frac{S_E}{\sigma^2} \sim \chi^2(n-r)$。 □

于是我们就能构造一个 F 检测统计量：

$$F = \frac{S_A/(r-1)}{S_E/(n-r)}$$

这是一个单侧检验，因为假设不成立时，S_A 会偏大。于是它的的拒绝域就是

$$F = \frac{S_A/(r-1)}{S_E/(n-r)} \geq F_\alpha(r-1, n-r)$$

我们可以把分析结果排在一个**方差分析表**中。

11.2 未知参数估计

在对 σ^2 做估计的时候，我们应该求助于 S_E，因为 S_A 与 σ^2 的关系会因原假设是否成立而改变。

由前面的式子，知道 $E(\frac{S_E}{n-r}) = \sigma^2$，它可以作为 σ^2 的无偏估计。我们对 μ 和 μ_i 也有无偏估计 $\overline{X}$，$\overline{X}_i$。

方差来源	平方和	自由度	均方	F 值
因素 A	S_A	$r-1$	$\overline{S}_A = \frac{S_A}{r-1}$	$F = \frac{\overline{S_A}}{\overline{S_E}}$
误差 E	S_E	$n-r$	$\overline{S}_E = \frac{S_E}{r-1}$	
综合 T	S_T	$n-1$		

表 2: 方差分析表

当我们拒绝原假设时，问题会变得复杂。这时均值是不完全相同的，即效应 δ_i 不全为 0。因为 $\delta_i = \mu_i - \mu$，所以其无偏估计为 $\overline{X}_i - \overline{X}$。

我们还想知道，拒绝原假设时 S_A 是怎样的。我们知道它会偏大，但是定量地说：

$$\begin{aligned}
E(S_A) &= E\left(\sum_{i=1}^{r} n_i \overline{X}_i^2 - n\overline{X}^2\right) \\
&= \sum_{i=1}^{r} n_i E(\overline{X}_i^2) - nE(\overline{X}^2) \\
&= \sum_{i=1}^{r} n_i [D(\overline{X}_i) + (E(\overline{X}_i))^2] - n[D(\overline{X}) + (E(\overline{X}))^2] \\
&= \sum_{i=1}^{r} n_i \left(\frac{\sigma^2}{n_i} + \mu_i^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \\
&= r\sigma^2 + \sum_{i=1}^{r} n_i(\mu + \delta_i)^2 - \sigma^2 + n\mu^2 \\
&= (r-1)\sigma^2 + \sum_{i=1}^{r} n_i\mu^2 + 2\mu\sum_{i=1}^{r} n_i\delta_i + \sum_{i=1}^{r} n_i\delta_i^2 - n\mu^2 \\
&= (r-1)\sigma^2 + \sum_{i=1}^{r} n_i\delta_i^2
\end{aligned}$$

在拒绝原假设时，我们不光想了解两个总体 $N(\mu_i, \sigma^2)$ 和 $N(\mu_j, \sigma^2)$ 的均值之差的点估计，我们还想要区间估计。根据前面的讨论，它们是独立的。我们这时应该用 t 分布去估计，而不是正态分布，因为方差是未知的，我们只是用 S_E 估计了方差!

这个 t 分布则应该通过 $\overline{X}_i - \overline{X}_j$ 和 S_E 来构造，它们是独立的。应用

前面的结论，就有：

$$\frac{(\overline{X}_i - \overline{X_j}) - (\mu_i - \mu_j)}{\sqrt{S_E\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim t(n_i + n_j - 2)$$

注意 S_E 是一个平方和，是类似方差而不是标准差的量，虽然它记号上很像标准差。

11.3 等重复双因素试验的方差分析

加入第二个因素之后，事情会一下子变得复杂起来。两个因素不仅可能各自独立地对结果产生效应，它们还可能交互地对结果产生效应。同时如果试验重复次数不等，那么因素的分解还会更加困难，和顺序相关（会产生四类平方和的计算方式）。所以这里我们先研究等重复的情况。

我们的模型会变成这样：假设因素 A 有 r 个水平，因素 B 有 r 个水平，每种水平的组合下进行了 t 次独立试验，结果为随机变量 X_{ijk}。双因素方差分析的模型为：

$$X_{ijk} = \mu_{ij} + \epsilon_{ijk}, \quad i = 1, 2, \ldots, r; \quad j = 1, 2, \ldots, s; \quad k = 1, 2, \ldots, t$$

或者

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

其中：

- X_{ijk} 服从正态分布 $N(\mu_{ij}, \sigma^2)$，且相互独立。
- μ 为**总平均**。
- $\mu_{i\cdot}$ 为**行平均**，即对因素 A 的第 i 个水平求平均。
- $\mu_{\cdot j}$ 为**列平均**，即对因素 B 的第 j 个水平求平均。
- α_i 为因素 A 第 i 个水平的效应，即 $\mu_{i\cdot} - \mu$，满足 $\sum_{i=1}^{r} \alpha_i = 0$。
- β_j 为因素 B 第 j 个水平的效应，即 $\mu_{\cdot j} - \mu$，满足 $\sum_{j=1}^{s} \beta_j = 0$。
- γ_{ij} 为因素 A 第 i 个水平与因素 B 第 j 个水平的**交互效应**，即 $\mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu$，满足 $\sum_{i=1}^{r} \gamma_{ij} = 0$，$\sum_{j=1}^{s} \gamma_{ij} = 0$。
- ϵ_{ijk} 为随机误差，服从正态分布 $N(0, \sigma^2)$。注意这里预设了方差齐性。

这时其实就有三个假设待我们检验：

$$
\begin{aligned}
&H_{01}:\alpha_1=\cdots=\alpha_r=0\\
&H_{02}:\beta_1=\cdots=\beta_s=0\\
&H_{03}:\gamma_{11}=\cdots=\gamma_{rs}=0
\end{aligned}
$$

也就是因素 A 有没有差女生效应，因素 B 有没有产生效应，因素 A 和因素 B 有没有产生交互效应。

类比单因素的思路，我们可以将总偏差平方和（又叫**总变差**）进行分解，分解为**误差平方和** S_E，因素 A、因素 B 的**效应平方和** S_A, S_B，和 A 与 B 的**交互效应平方和** $S_{A\times B}$。

我们先引入一些统计量的记号，然后再来研究这一分解。

$$
\begin{aligned}
\overline{X}&=\frac{1}{rst}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}X_{ijk}\\
\overline{X}_{i\cdot}&=\frac{1}{st}\sum_{j=1}^{s}\sum_{k=1}^{t}X_{ijk}\\
\overline{X}_{\cdot j}&=\frac{1}{rt}\sum_{i=1}^{r}\sum_{k=1}^{t}X_{ijk}\\
\overline{X}_{ij}&=\frac{1}{t}=\sum_{k=1}^{t}X_{ijk}
\end{aligned}
$$

和单因素类似，总变差不过多了一次求和，我们这里给出它的定义和分

解:

$$
\begin{aligned}
S_T &= \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(X_{ijk}-\overline{X})^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}[(X_{ijk}-\overline{X}_{ij})+(\overline{X}_{i\cdot}-\overline{X})+(\overline{X}_{\cdot j}-\overline{X})+(\overline{X}_{ij}-\overline{X}_{i\cdot}-\overline{X}_{\cdot j}+\overline{X})]^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(X_{ijk}-\overline{X}_{ij})^2 \\
&\quad + st\sum_{i=1}^{r}(\overline{X}_{i\cdot}-\overline{X})^2 \\
&\quad + rt\sum_{j=1}^{s}(\overline{X}_{\cdot j}-\overline{X})^2 \\
&\quad + t\sum_{i=1}^{r}\sum_{j=1}^{s}(\overline{X}_{ij}-\overline{X}_{i\cdot}-\overline{X}_{\cdot j}+\overline{X})^2
\end{aligned}
$$

这四部分分解就是:

$$
\begin{aligned}
S_E &= \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(X_{ijk}-\overline{X}_{ij})^2 \\
S_A &= st\sum_{i=1}^{r}(\overline{X}_{i\cdot}-\overline{X})^2 \\
S_B &= rt\sum_{j=1}^{s}(\overline{X}_{\cdot j}-\overline{X})^2 \\
S_{A\times B} &= t\sum_{i=1}^{r}\sum_{j=1}^{s}(\overline{X}_{ij}-\overline{X}_{i\cdot}-\overline{X}_{\cdot j}+\overline{X})^2 \\
S_T &= S_E+S_A+S_B+S_{A\times B}
\end{aligned}
$$

平方和也可以被写成：

$$
\begin{aligned}
S_E &= \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t} X_{ijk}^2 - rst\overline{X}^2 \\
S_A &= st\sum_{i=1}^{r}\overline{X}_{i\cdot}^2 - rst\overline{X}^2 \\
S_B &= rt\sum_{j=1}^{s}\overline{X}_{\cdot j}^2 - rst\overline{X}^2 \\
S_{A\times B} &= \left(t\sum_{i=1}^{r}\sum_{j=1}^{s}\overline{X}_{ij} - rst\overline{X}^2\right) - S_A - S_B \\
S_E &= S_T - S_A - S_B - S_{A\times B}
\end{aligned}
$$

它们的统计特性和单因素类似：

- $\overline{X}$,S_A,S_B,$S_{A\times B}$,S_E 相互独立。
- $\frac{S_E}{\sigma^2} \sim \chi^2(rs(t-1))$。
- $E(S_E) = rs(t-1)\sigma^2$。
- H_{01} 成立时，$\frac{S_A}{\sigma^2} \sim \chi^2(r-1)$。更进一步，$\frac{S_A/(r-1)}{S_E/[rs(t-1)]} \sim F(r-1, rs(t-1))$。
- $E(S_A) = (r-1)\sigma^2 + st\sum_{i=1}^{r}\alpha_i^2$。
- H_{02} 成立时，$\frac{S_B}{\sigma^2} \sim \chi^2(s-1)$。更进一步，$\frac{S_B/(s-1)}{S_E/[rs(t-1)]} \sim F(s-1, rs(t-1))$。
- $E(S_B) = (s-1)\sigma^2 + rt\sum_{j=1}^{s}\beta_j^2$。
- H_{03} 成立时，$\frac{S_{A\times B}}{\sigma^2} \sim \chi^2((r-1)(s-1))$。更进一步，$\frac{S_{A\times B}/[(r-1)(s-1)]}{S_E/[rs(t-1)]} \sim F((r-1)(s-1), rs(t-1))$。
- $E(S_{A\times B}) = (r-1)(s-1)\sigma^2 + t\sum_{i=1}^{r}\sum_{j=1}^{s}\gamma_{ij}^2$。

于是像单因素试验一样，我们也可以对这三个假设用 F 检验了。我们一样可以列出方差分析表 3。

方差来源	平方和	自由度	均方	F 值
因素 A	S_A	$r-1$	$\overline{S}_A = \frac{S_A}{r-1}$	$F_A = \frac{\overline{S}_A}{\overline{S}_E}$
因素 B	S_B	$s-1$	$\overline{S}_B = \frac{S_B}{s-1}$	$F_B = \frac{\overline{S}_B}{\overline{S}_E}$
交互作用 $A \times B$	$S_{A\times B}$	$(r-1)(s-1)$	$\overline{S}_{A\times B} = \frac{S_{A\times B}}{(r-1)(s-1)}$	$F_{A\times B} = \frac{\overline{S}_{A\times B}}{\overline{S}_E}$
误差 E	S_E	$rs(t-1)$	$\overline{S}_E = \frac{S_E}{rs(t-1)}$	
总和 T	S_T	$rst-1$		

表 3: 双因素等重复试验方差分析表

12 线性回归分析

回归分析是一种统计方法，用于研究因变量（被解释变量）与一个或多个自变量（解释变量）之间的关系。它旨在建立一个数学模型，描述变量间的依赖关系，并用于预测、解释或控制。这里的关系不完全是确定性的，还会加上一定的随机误差。一般的回归模型可以表示为

$$Y = f(x_1, \ldots, x_n) + \epsilon$$

Y 是因变量，f 是我们的回归函数模型，x_i 是自变量，ϵ 是随机误差项（它的均值应该是 0，并且它常常是正态的）。这里注意一种本性上的区分，我们的自变量是普通的变量，而因变量是随机变量。对于自变量每个确定的取值，都有一个因变量的分布与之对应。

先看回单变量的形式。对于自变量的一组不完全相同的取值 $x_1, \ldots, x_n$，设 $Y_1, \ldots, Y_n$ 是在这些取值处对 Y 的独立观察结果。那么我们称

$$(x_1, Y_1), \ldots, (x_n, Y_n)$$

为一个**样本**。注意它们独立但不一定同分布，这就是回归分析的不同之处。

对应的样本值就记为

$$(x_1, y_1), \ldots, (x_n, y_n)$$

我们先从最简单的回归模型入手，即

$$Y = a + bx + \epsilon$$

这种一元线性函数。

12.1 一元线性回归

在给定的线性模型和假设下，我们要怎么估计参数 a 与 b 的值？我们会采取最大似然估计。

按我们的模型：

$$Y_i \sim a + bx_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad \epsilon_i$$

或者说

$$Y_i \sim N(a + bx_i, \sigma^2), \quad Y_i$$

那么我们就有似然函数：

$$\begin{aligned} L(a,b) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - a - bx_i)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2\right) \end{aligned}$$

取个对数就知道最大化似然函数 $L(a,b)$ 等价于最小化

$$Q(a,b) = \sum_{i=1}^{n}(y_i - a - bx_i)^2$$

下面就是最小二乘法的推导！

最小化就需要我们对其取个偏导：

$$\begin{aligned} \frac{\partial Q}{\partial a} &= -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \\ \frac{\partial Q}{\partial b} &= -2\sum_{i=1}^{n}(y_i - a - bx_i)x_i = 0 \end{aligned}$$

整理一下就是

$$\begin{aligned} na + \left(\sum_{i=1}^{n} x_i\right) b &= \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right) a + \left(\sum_{i=1}^{n} x_i^2\right) b &= \sum_{i=1}^{n} x_i y_i \end{aligned}$$

这被称为**正则方程**。

我们用矩阵给出更紧凑的表示：

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} a \\ b \end{bmatrix}$$

我们线性模型的估计值是：

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}$$

我们的目标是最小化平方和：

$$Q(\boldsymbol{\beta}) = (\hat{\mathbf{y}} - \mathbf{y})^T(\hat{\mathbf{y}} - \mathbf{y})$$

用这种形式重述正则方程就是

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = 2\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} - \mathbf{y}) = \mathbf{0}$$

也就是：

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}$$

问题的关键就在 $\mathbf{X}^T\mathbf{X}$ 的可逆性上。事实上，如果 x_i 不全相同，它肯定可逆。这个方阵是：

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}, \quad \mathbf{X}^T\mathbf{y} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

（本节求和号过多，有时我会在不造成歧义的情况下省略上下限。）

令人震惊的事实是：

$$\det(\mathbf{X}^T\mathbf{X}) = n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 = n\sum_{i=1}^{n}(x_i - \overline{x})^2$$

于是求个逆就是：

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{adj(\mathbf{X}^T\mathbf{X})}{\det(\mathbf{X}^T\mathbf{X})} = \begin{bmatrix} \frac{\frac{1}{n}\sum x_i^2}{\sum(x_i-\overline{x})^2} & -\frac{\overline{x}}{\sum(x_i-\overline{x})^2} \\ -\frac{\overline{x}}{\sum(x_i-\overline{x})^2} & \frac{1}{\sum(x_i-\overline{x})^2} \end{bmatrix}$$

解出来就是：

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y}) = \begin{bmatrix} \frac{\frac{1}{n}(\sum x_i^2)(\sum y_i) - \overline{x}\sum x_i y_i}{\sum(x_i-\overline{x})^2} \\ \frac{\sum x_i y_i - \overline{x}\sum y_i}{\sum(x_i-\overline{x})^2} \end{bmatrix}$$

这不是那种最方便记忆的版本。对分母没什么可说的了，不过我们还可以调整一下分子。

$$\begin{aligned} \hat{a} &= \frac{1}{n}\frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{\sum(x_i - \overline{x})^2} \\ \hat{b} &= \frac{\sum x_i y_i - \overline{x}\sum y_i - \overline{y}\sum x_i + n\overline{x}\overline{y}}{\sum(x_i - \overline{x})^2} \\ &= \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} \\ \overline{y} &= a + b\overline{x} \end{aligned}$$

于是我们就得到了 Y 关于 x 的**回归方程**：$\hat{y} = \hat{a} + \hat{b}x$。

由于 $\hat{b}$ 的式子更好记忆，另一种写法是 $\hat{y} = \overline{y} - \hat{b}(x - \overline{x})$。

我们引入点记号，让它们更好记一点。实际上这些式子等价形式很多。

$$
\begin{aligned}
S_{xx} &= \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2 \\
S_{yy} &= \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2 \\
S_{xy} &= \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} y_i)
\end{aligned}
$$

它们叫**离差平方和**与**交叉离差平方和**。注意不要把这些量和样本方差和标准差搞混，它们差一个 $n-1$ 因子。我们可以把 $\hat{b}$ 写成 $\frac{S_{xx}}{S_{xy}}$。

以上就是**最小二乘法**，适用于用线性函数 $a + bx$ 最小化残差平方和的所有情境。这里正态分布的最大似然估计正好与它不谋而合。

接下来介绍几个平方和。**回归平方和**反映因变量变异中能被自变量线性解释的部分：

$$SSR = \sum(\hat{y} - \overline{y})^2$$

残差平方和反映模型无法解释的随机误差部分：

$$SSE = \sum(\hat{y} - y)^2$$

总平方和反映因变量的总变异程度：

$$SST = \sum(y - \overline{y})^2$$

我们可以得到这样的平方和分解：

$$SST = SSR + SSE$$

其证明仍然和前面的平方和分解一样，是直接的展开。

为了衡量回归模型的拟合优度，定义**决定系数** R^2：

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

它表示因变量变异中被自变量线性解释的比例。

在我们的线性回归中，可以得到：

$$
\begin{aligned}
SST &= S_{yy} \\
SSE &= S_{yy} - \frac{S_{xy}^2}{S_{xx}} \\
SSR &= \frac{S_{xy}^2}{S_{xx}} \\
R^2 &= \frac{S_{xy}^2}{S_{xx}S_{yy}}
\end{aligned}
$$

R^2 正好和样本相关系数平方一致。

刚才我们给出的都是估计值的计算，我们只要把小 y 换成大 Y，就变成了估计量。

我们最感兴趣的肯定是两个参数的估计量，不过让我们先从最基础的来吧。$\overline{Y} = \frac{1}{n}\sum Y_i$ 是一些独立正态分布的线性组合，所以它也是正态的。

它的均值是：

$$
\begin{aligned}
E(\overline{Y}) &= \frac{1}{n}\sum E(Y_i) \\
&= \frac{1}{n}\sum (a + bx_i) = a + b\overline{x}
\end{aligned}
$$

方差是：

$$
D(\overline{Y}) = \frac{1}{n^2}\sum D(Y_i) = \frac{\sigma^2}{n}
$$

于是 $\overline{Y} \sim N(a + b\overline{x}, \frac{\sigma^2}{n})$。

前面的离差平方和 S_{xy} 和 S_{yy} 作为统计量的形式是：

$$
\begin{aligned}
S_{xY} &= \sum (x_i - \overline{x})(Y_i - \overline{Y}) \\
&= \sum (x_i - \overline{x})Y_i - \overline{Y}\sum (x_i - \overline{x}) \\
&= \sum (x_i - \overline{x})Y_i \\
S_{YY} &= \sum (Y_i - \overline{Y})^2 \\
&= \sum Y_i^2 - n\overline{Y}^2
\end{aligned}
$$

能看出 S_{xY} 是正态的，而 S_{YY} 看起来有点像卡方但是没那么卡方。我们先来研究 S_{xY}：

$$\begin{aligned} E(S_{xY}) &= \sum (x_i - \overline{x}) E(Y_i) \\ &= \sum (x_i - \overline{x})(a + bx_i) \\ &= a \sum (x_i - \overline{x}) + b(\sum x_i^2 - \overline{x} \sum x_i) \\ &= b(\sum x_i^2 - n\overline{x}^2) \\ &= bS_{xx} \end{aligned}$$

$$\begin{aligned} D(S_{xY}) &= \sum (x_i - \overline{x})^2 D(Y_i) \\ &= \sum (x_i - \overline{x})^2 \sigma^2 = \sigma^2 S_{xx} \end{aligned}$$

于是 $S_{xY} \sim N(bS_{xx}, \sigma^2 S_{xx})$。

好！这下斜率的估计量 $\hat{b} = \frac{S_{xY}}{S_{xx}}$ 也清楚了，$\hat{b} \sim N(b, \frac{\sigma^2}{S_{xx}})$。一个不错的无偏估计！

于是截距的估计量就是

$$\hat{a} = \overline{Y} - \hat{b}\overline{x} \sim N\left(a, \left(\frac{1}{n} + \frac{\overline{x}}{S_{xx}}\right)\sigma^2\right)$$

顺带着，如果给我某个自变量的新 x_0 值，我也能给出因变量对应的新观察值 Y_0 的估计量：

$$\hat{Y_0} = \hat{a} + \hat{b}x_0 = \overline{Y} + \hat{b}(x_0 - \overline{x}) \sim N\left(a + bx_0, \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)\sigma^2\right)$$

与其说是估计，不如说是预测。我们把 $\hat{Y_0}$ 叫做观察值 Y_0 的**点预测**。

不过写了这么多，我们暂时还是没给出对方差 σ^2 的估计。我们有理由认为它就藏在残差平方和之中。

事实上，这里又有一个重要的结论：

定理 17. *对于样本均值响应 $\overline{Y}$，斜率估计量 $\hat{b}$，残差平方和 SSE，有*

- *$\overline{Y}$，$\hat{b}$，SSE 相互独立*
- *$\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$*

证明. 我已经有点烦了。这种结论已经出现过三次了……感性地说，因为估计了两个参数，我们损失了两个自由度；理性地说，我还是可以构造一个正交变换，变换前是 $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$，变换后结果的第一个分量是 $\overline{Y}$，第二个分量是 $\hat{b}$，剩下分量的平方和是 SSE。

于是正交矩阵 Q 的第一行献给

$$\left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$$

第二行献给

$$\left(\frac{x_1 - \overline{x}}{\sqrt{S_{xx}}}, \ldots, \frac{x_n - \overline{x}}{\sqrt{S_{xx}}}\right)$$

剩下的随便 Gram-Schimdt 规范正交化出来。

这样我们就把 $Q\mathbf{Y}$ 的前两个分量变成 $\sqrt{n}\overline{Y}$ 和 $\hat{b}\sqrt{S_{xx}}$，剩下的 $n-2$ 个分量变成独立同方差的正态了。 □

这个结论还可以用到预测值上面。如果预测值和之前的观察值相互独立，那么预测值与上面三个量相互独立。证明只需要在上面多加一个分量，在此略过。

对 σ^2 的估计这就有了，$\hat{\sigma^2} = \frac{SSE}{n-2} = \frac{1}{n-2}(S_{YY} - \hat{b}S_{xY})$ 是个无偏估计。

有了方差的估计，我们就可以给出前面那些量的区间估计了，t 分布再次出现。

对于斜率 b，我们有

$$\hat{b} \sim N(b, \frac{\sigma^2}{S_{xx}})$$

$$\frac{\hat{b} - b}{\sigma}\sqrt{S_{xx}} \sim N(0, 1)$$

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$$

所以说

$$\frac{\hat{b} - b}{\sqrt{\frac{SEE}{n-1}}}\sqrt{S_{xx}} \sim t(n-2)$$

那么其估计和检验方法就很清楚了。

对于我们的点预测 $\hat{Y}_0$，我们有

$$\hat{Y}_0 \sim N\left(a + bx_0, \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)\sigma^2\right)$$

将其标准化得到：

$$\frac{\hat{Y}_0 - Y_0}{\sigma\sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}}} \sim N(0,1)$$

配合上

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$$

就有

$$\frac{\hat{Y}_0 - Y_0}{\sqrt{\frac{SSE}{n-2}}\sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}}} \sim t(n-2)$$

于是我们就可以给出区间预测了。