Lecture 11

样本分布的性质

Probability and Statistics
Beihang University

11.1　随机样本与统计量

定义 11.1.1 随机变量 $X_1, \dots, X_n$ 被称为**随机样本**, 如果它们是独立同分布的.

定义 11.1.2 T 是定义在随机样本 $(X_1, \dots, X_n)$ 值域空间上的函数, 那么称随机变量 $Y = T(X_1, \dots, X_n)$ 为**统计量**.

- 样本平均值 $\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i$
- 样本方差 $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2$

引理 11.1.1 对任意实数 $x_1, \dots, x_n$, 记 $\overline{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$. 那么

$$\sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} (x_i - \overline{x}_n)^2 + \sum_{i=1}^{n} (\overline{x}_n - a)^2$$

特别地 $(a = 0)$: $\sum_{i=1}^{n} (x_i - \overline{x}_n)^2 = \sum_{i=1}^{n} x_i^2 - n\left(\overline{x}_n\right)^2$

- 注意到

$$\sum_{i=1}^{n}(x_i-\overline{x}_n)=0$$

因此

$$\begin{aligned}
&\sum_{i=1}^{n}(x_i-a)^2\\
&=\sum_{i=1}^{n}(x_i-\overline{x}_n+\overline{x}_n-a)^2\\
&=\sum_{i=1}^{n}(x_i-\overline{x}_n)^2+2\sum_{i=1}^{n}(x_i-\overline{x}_n)(\overline{x}_n-a)+\sum_{i=1}^{n}(\overline{x}_n-a)^2\\
&=\sum_{i=1}^{n}(x_i-\overline{x}_n)^2+\sum_{i=1}^{n}(\overline{x}_n-a)^2
\end{aligned}$$

定理 11.1.1 假设随机样本 $X_1, ..., X_n$ 来自期望和方差分别为 μ 和 σ^2 的分布. 那么 (1)$E\left(\overline{X}_n\right) = \mu$. (2)$Var\left(\overline{X}_n\right) = \sigma^2/n$. (3)$E\left(S_n^2\right) = \sigma^2$.

- 只证明 (3).

$$
\begin{aligned}
E\left(S_n^2\right) &= \frac{1}{n-1} E\left[\sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2\right] \\
&= \frac{1}{n-1} E\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2\right] \\
&= \frac{1}{n-1} \left[\sum_{i=1}^{n} \left(\sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right] \\
&= \sigma^2
\end{aligned}
$$

11.2 顺序统计量

定义 11.2.1 假设 $X_1, \ldots, X_n$ 为取自某一分布的随机样本. 令

$$X_{(1)} = \min\{X_i : 1 \leqslant i \leqslant n\},$$

$$X_{(2)} = \min\left\{X_i : X_{(1)} \leqslant X_i, 1 \leqslant i \leqslant n\right\},$$

$$\cdots\cdots$$

$$X_{(n)} = \min\left\{X_i : X_{(n-1)} \leqslant X_i, 1 \leqslant i \leqslant n\right\},$$

那么 $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$, 我们称 $X_{(1)}, \ldots, X_{(n)}$ 为顺序统计量, 称 $X_{(i)}$ 为第 i 个顺序统计量.

定理 11.2.1 假设 $X_1, \ldots, X_n$ 为来自某一离散分布的随机样本, 其概率函数为 $f_X(x_j) = p_j$, 其中 $x_1 < x_2 < \cdots$ 为 X 的依递增次序

排列的所有可能取值. 令 $P_0 = 0, P_j = \sum_{k=1}^{j} p_k, j \geqslant 1$. 那么

$$P\left(X_{(i)} \leqslant x_j\right) = \sum_{k=i}^{n} \binom{n}{k} \left(P_j\right)^k \left(1 - P_j\right)^{n-k}, 1 \leqslant i \leqslant n$$

- 给定 j, 令 $Y_k = 1_{\{X_k \leqslant x_j\}}, (1 \leqslant k \leqslant n)$, $Y = \sum_{k=1}^{n} Y_k$, 即随机变量 Y 记录每一组随机样本 $X_1, ..., X_n$ 中样本值不超过 x_j 的样本个数. 由于 $X_1, ..., X_n$ 独立同分布, 每一个 Y_k 都是参数为 $P\left(X_k \leqslant x_j\right) = P_j$ 的 Bernoulli 随机变量, 从而 Y 服从参数为 n 和 P_j 的二项分布. 容易看出 $\left\{X_{(i)} \leqslant x_j\right\} = \{Y \geqslant i\}$.

  因此

$$P\left(X_{(i)} \leqslant x_j\right) = P\left(Y \geqslant i\right) = \sum_{k=i}^{n} \binom{n}{k} \left(P_j\right)^k \left(1 - P_j\right)^{n-k}.$$

定理 11.2.2 假设随机样本 $X_1, ..., X_n$ 来自某一连续型分布, 其分布和密度函数分别为 $F_X(x)$ 和 $f_X(x)$. 那么 $X_{(i)}$ 的密度函数为, $1 \leqslant i \leqslant n$,

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} f_X(x) (F_X(x))^{i-1} (1 - F_X(x))^{n-i}.$$

■ (解法 I) 给定 x, 令 $Y_k = 1_{\{X_k \leqslant x\}}, (1 \leqslant k \leqslant n), Y = \sum_{k=1}^n Y_k$, 即随机变量 Y 记录每一组随机样本 $X_1, ..., X_n$ 中样本值不超过 x 的样本个数., 那么它服从参数为 n 和 $P(X \leqslant x) = F_X(x)$ 的二项分布. 从而

$$P\left(X_{(i)} \leqslant x\right) = P(Y \geqslant i) = \sum_{k=i}^{n} \binom{n}{k} (F_X(x))^k (1 - F_X(x))^{n-k},$$

将上式对 x 求导即得到 $f_{X_{(i)}}(x)$, 化简过程作为练习!

(解法 II) Y_k $(1 \leqslant k \leqslant n)$ 的定义同解法 I. 注意

$$f_{X_{(i)}}(x) = \lim_{\varepsilon \to 0+} \frac{F_X(x+\varepsilon) - F_X(x)}{\varepsilon} = \lim_{\varepsilon \to 0+} \frac{P\left(X_{(i)} \in (x, x+\varepsilon]\right)}{\varepsilon}$$

为求得 $f_X(x)$, 考虑 $\forall \varepsilon > 0$, $Z_k = \sum_{j=1}^{k-1} Y_j + \sum_{j=k+1}^{n} Y_j$

$$\begin{aligned}
& P\left(X_{(i)} \in (x, x+\varepsilon]\right) \\
& = \sum_{k=1}^{n} P\left(X_k \in (x, x+\varepsilon], Z_k = i-1\right) \\
& = nP\left(X_1 \in (x, x+\varepsilon], Z_1 = i-1\right) \\
& = nP\left(X_1 \in (x, x+\varepsilon]\right) P\left(Z_1 = i-1\right) \\
& = nP\left(X_1 \in (x, x+\varepsilon]\right) \binom{n-1}{i-1} \left(F_X(x)\right)^{i-1} \left(1 - F_X(x)\right)^{n-i}
\end{aligned}$$

两边同除以 ε 并取极限就有

$$
\begin{aligned}
f_{X_{(i)}}(x) &= n f_X(x) \binom{n-1}{i-1} (F_X(x))^{i-1} (1-F_X(x))^{n-i} \\
&= \frac{n!}{(i-1)!(n-i)!} f_X(x) (F_X(x))^{i-1} (1-F_X(x))^{n-i}
\end{aligned}
$$

在上述两个定理的证明中, $\frac{1}{n}\sum_{k=1}^{n} Y_k$ 常常被称为**经验分布函数**, 我们在大数定律的应用部分已经遇到过该函数.

例题 11.2.1 随机样本 $X_1, ..., X_n$ 服从 $(0,1)$ 上的 (标准) 均匀分布. 试写出其顺序统计量分布.

■ 标准均匀分布的分布函数为 $F_X(x) = x$. 因此 $X_{(i)}, (1 \leqslant i \leqslant n)$ 的密度函数为

$$
f_{X_{(i)}}(x) = \begin{cases} n\binom{n-1}{i-1} x^{i-1}(1-x)^{n-i}, & x \in (0,1) \\ 0, & x \notin (0,1) \end{cases}
$$

作为例题 11.2.1的一个副产品, 注意到, 由于 $f_{X_{(i)}}(x)$ 是密度函数, 我们有

$$\int_0^1 x^{i-1}(1-x)^{n-i} = \frac{1}{n\binom{n-1}{i-1}} = \frac{(i-1)!\,(n-i)!}{n!}, 1 \leqslant i \leqslant n.$$

因此, 作代换 $i \longleftrightarrow \alpha, n-i+1 \longleftrightarrow \beta$ 就有, $\alpha, \beta \in \{1, 2, \dots\}$

$$\begin{aligned}\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} &= \frac{1}{(\alpha+\beta-1)\binom{\alpha+\beta-2}{\alpha-1}} \\ &= \frac{(\alpha-1)!\,(\beta-1)!}{(\alpha+\beta-1)!} = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}.\end{aligned}$$

标准均匀分布的顺序统计量 $X_{(i)}$ 的分布实际上是 Beta 分布的一个特殊情形.

定义 11.2.2 X 称为服从 Beta 分布, 如果它具有密度函数, $\alpha > 0$, $\beta > 0$,

$$f(t) = \begin{cases} \frac{1}{B(\alpha,\beta)} t^{\alpha-1} (1-t)^{\beta-1}, & t \in (0,1) \\ 0, & t \notin (0,1) \end{cases}$$

记作 $X \sim Beta(\alpha, \beta)$. 其中

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1} (1-t)^{\beta-1} dt$$

是密度函数的归 1 常数, 也被称为 Beta 函数.

引理 11.2.1 Beta 函数与 Gamma 函数满足以下关系, $\alpha > 0$, $\beta > 0$,

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- Gamma 函数的定义为

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$$

令 $x = sy$ 得到

$$\Gamma\left(\alpha\right) = s^{\alpha}\int_{0}^{\infty} y^{\alpha-1}e^{-sy}dy$$

运用这一关系式

$$\begin{aligned}
\Gamma\left(\alpha\right)\Gamma\left(\beta\right) &= \left[\int_{0}^{\infty} x^{\alpha-1}e^{-x}dx\right]\cdot\left[x^{\beta}\int_{0}^{\infty} y^{\beta-1}e^{-xy}dy\right] \\
&= \int_{0}^{\infty}\left[\int_{0}^{\infty} x^{\alpha+\beta-1}e^{-(1+y)x}dx\right]y^{\beta-1}dy \\
&= \Gamma\left(\alpha+\beta\right)\int_{0}^{\infty}\frac{y^{\beta-1}}{\left(1+y\right)^{\alpha+\beta}}dy
\end{aligned}$$

令

$$y = \frac{t}{1-t} = \frac{1}{1-t} - 1, dy = \frac{1}{\left(1-t\right)^{2}}dt$$

就有

$$
\begin{aligned}
&\int_0^{\infty} \frac{y^{\beta-1}}{(1+y)^{\alpha+\beta}} dy \\
&= \Gamma(\alpha+\beta) \int_0^{\infty} \left(\frac{t}{1-t}\right)^{\beta-1} (1-t)^{\alpha+\beta} \frac{1}{(1-t)^2} dt \\
&= \int_0^{\infty} t^{\beta-1} (1-t)^{\alpha-1} dt = B(\beta,\alpha) = B(\alpha,\beta)
\end{aligned}
$$

因此所要证明的等式成立.

习题 11.2.1 随机变量 X_1, X_2 相互独立, $X_i \sim Gamma(\alpha_i, \beta)$, $i = 1, 2$. $\alpha_1 > 0$, $\alpha_2 > 0$, $\beta > 0$. 试运用密度卷积公式直接证明 $X_1 + X_2 \sim Gamma(\alpha_1 + \alpha_2, \beta)$.

11.3 χ^2 分布

例题 11.3.1 $X \sim \mathcal{N}(0,1)$, 试求出 $Y = X^2$ 的分布.

- Y 的密度为, $\forall y > 0$

$$
\begin{aligned}
f_Y(y) &= \frac{\partial}{\partial y} P(Y \leqslant y) \\
&= \frac{\partial}{\partial y} P\left(-Y^{1/2} \leqslant X \leqslant Y^{1/2}\right) \\
&= \frac{1}{\sqrt{2\pi}} \frac{\partial}{\partial y} \int_{-y^{1/2}}^{y^{1/2}} e^{-\frac{1}{2}x^2} dx \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(y^{1/2}\right)^2} \cdot \frac{1}{2y^{1/2}} - \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(-y^{1/2}\right)^2} \cdot \left(-\frac{1}{2y^{1/2}}\right) \\
&= \frac{1}{\sqrt{2\pi}} \cdot y^{-1/2} e^{-y/2}
\end{aligned}
$$

X^2 的分布正是 Gamma 分布的一个子情形. 当 $\alpha = 1/2$ 时, 令 $x = y^2/2$, $dx = ydy$,

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty x^{-1/2} e^{-x} dx = 2^{1/2} \int_0^\infty e^{-y^2/2} dy$$
$$= 2^{1/2} \cdot \frac{(2\pi)^{1/2}}{2} = \pi^{1/2}$$

由此可见例题 11.3.1中 $X^2 \sim Gamma\,(1/2, 1/2)$.

定义 11.3.1 假设 $m > 0$. 若 $Y \sim Gamma\,(m/2, 1/2)$, 即其密度函数为

$$f(x|m) = \begin{cases} \frac{2^{-m/2}}{\Gamma(m/2)} x^{m/2-1} e^{-x/2}, & x > 0 \\ 0, & x \leqslant 0 \end{cases}$$

那么称 Y 具有自由度为 m 的 χ^2 分布, 记为 $Y \sim \chi^2(m)$.

因此若 $X \sim \mathcal{N}(0, 1)$, 那么 $X^2 \sim \chi^2(1) = Gamma\,(1/2, 1/2)$.

定理 11.3.1 随机变量 $X_1, ..., X_n$ 独立同分布, $X_i \sim \mathcal{N}(0,1)$, $1 \leqslant i \leqslant n$. 那么 $X_1^2 + ... + X_n^2 \sim \chi^2(n)$.

■ 随机变量 $X_1, ..., X_n$ 独立同分布, 因此 $X_1^2, ..., X_n^2$ 独立同分布, 并且 $X_i^2 \sim \chi^2(1) = Gamma(1/2, 1/2)$, 由 $Gamma$ 分布的性质可知

$$X_1^2 + ... + X_n^2 \sim Gamma(n/2, 1/2) = \chi^2(n).$$

例题 11.3.2 假设 $X_1, ..., X_n$ 为来自 $\mathcal{N}(\mu, \sigma^2)$ 的随机样本, 那么

$$\sum_{k=1}^{n} \left(\frac{X_k - \mu}{\sigma}\right)^2 \sim \chi^2(n)$$

11.4 样本均值与方差的联合分布

同前面一样, 记样本均值 $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

定理 11.4.1 假设 $X_1, \dots, X_n$ 为来自 $\mathcal{N}\left(\mu, \sigma^2\right)$ 的随机样本, 那么

(1) $\overline{X}_n$ 与 $\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$ 相互独立;

(2) $\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$;

(3) $\frac{1}{\sigma^2}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2 \sim \chi^2(n-1)$.

- 结论 (2) 可以由正太分布的性质直接得到. 现证明 (1).

// 情形 I：$X_k \sim \mathcal{N}(0, 1), \forall k$

根据 Gram-Schmidt 正交化, 存在正交矩阵 A, 使得其第一行为

$$u^T = \left(n^{-1/2}, \dots, n^{-1/2}\right)$$

令 $X = (X_1, \ldots, X_n)^T$, $Y = AX$. 由于正交变换不改变正态分布之间的独立性以及协方差矩阵, 因此 $Y_1, \ldots, Y_n$ 仍然是相互独立的标准正太分布, 并且

$$Y_1 = u^T \cdot X = n^{1/2}\overline{X}_n,$$

由正交变换的性质我们还有

$$\sum_{i=1}^n Y_i^2 = Y^T Y = (AX)^T AX = X^T A^T AX = \sum_{i=1}^n X_i^2$$

从而

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

由于 Y_1 与 $Y_2, \ldots, Y_n$ 相互独立, 因此 Y_1 与 $\sum_{i=2}^{n} Y_i^2$ 相互独立, 也就是 $n^{1/2}\overline{X}_n$ 与 $\sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2$ 相互独立. 由于 $Y_2, \ldots, Y_n$ 是相互独立的标准正太分布, $\sum_{i=2}^{n} Y_i^2 \sim \chi^2 (n-1)$, 从而 (3) 成立.

情形 II: $X_k \sim \mathcal{N}\left(\mu, \sigma^2\right), \forall k$

令 $Z_k = (X_k - \mu)/\sigma$, 那么 $Z_1, \ldots, Z_n$ 是相互独立的标准正太分布. 注意

$$\overline{Z}_n = \frac{\overline{X}_n - \mu}{\sigma}, \sum_{i=1}^{n} \left(Z_i - \overline{Z}_n\right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2$$

借由情形 I 便得到所要证明的结论.

11.5 t 分布

定义 11.5.1 假设 Y 与 Z 相互独立. $Y \sim \chi^2(n), Z \sim \mathcal{N}(0,1)$. 称

$$X = \frac{Z}{\left(\frac{Y}{n}\right)^{1/2}}$$

的分布为自由度为 n 的 t 分布, 记为 $X \sim t(n)$. 其分布函数常常写作 $T_n(x)$.

从定义中看出, t 分布是关于 0 点对称的, 即对任意 $c > 0$, 总有 $T_n(-c) = 1 - T_n(c)$. 事实上, 正态分布一样, t 分布密度点点大于 0, 分布函数 T_n 严格增加, 因而分为函数正是逆函数 T_n^{-1}. 若 γ 给定, $c > 0$ 使得下式成立

$$P(-c < X < c) = \gamma$$

那么 $\gamma = 2T_n(c) - 1, c = T_n^{-1}((\gamma+1)/2)$, 即 c 是对应于 $(\gamma+1)/2$ 的分位值.

定理 11.5.1 假设 $X_1, \ldots, X_n$ 为来自 $\mathcal{N}(\mu, \sigma^2)$ 的随机样本, $\overline{X}_n$ 为样本平均, 令

$$\sigma' = \left[\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2}{n-1}\right]^{1/2}$$

那么

$$\frac{1}{\sigma'} n^{1/2}\left(\overline{X}_n - \mu\right) \sim t(n-1).$$

- 令

$$Y = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2, Z = n^{1/2}\frac{\overline{X}_n - \mu}{\sigma}$$

由定理 11.4.1, Y 与 Z 相互独立, $Y \sim \chi^2(n-1)$, $Z \sim \mathcal{N}(0,1)$. 因

此根据定义

$$\frac{n^{1/2}\left(\overline{X}_n-\mu\right)}{\sigma'}=\frac{Z}{\left(\frac{Y}{n-1}\right)^{1/2}}\sim t\left(n-1\right)$$

证明过程中, 尽管 Y 和 Z 都依赖于 σ, 但是

$$\frac{\overline{X}_n-\mu}{\sigma'/\sqrt{n}}$$

确是与无关的.

定理 11.5.2 假设 $X_1,\ldots,X_m$ 为来自 $X\sim\mathcal{N}\left(\mu,\sigma^2\right)$ 的随机样本, $Y_1,\ldots,Y_n$ 为来自 $Y\sim\mathcal{N}\left(\mu,\sigma^2\right)$ 的随机样本. 两个总体 X,Y 的分布是相互独立的. 令 $\overline{X}_m=\frac{1}{m}\sum_{i=1}^{m}X_i,\overline{Y}_n=\frac{1}{n}\sum_{j=1}^{n}Y_j,$

$S_X^2=\sum_{i=1}^m\left(X_i-\overline{X}_m\right)^2, S_Y^2=\sum_{j=1}^n\left(Y_j-\overline{Y}_n\right)^2$ 那么对任意参数值 $\left(\mu,\sigma^2\right)$ 都有

$$U=\frac{(m+n-2)^{1/2}\left(\overline{X}_m-\overline{Y}_n\right)}{\left(\frac{1}{m}+\frac{1}{n}\right)^{1/2}\left(S_X^2+S_Y^2\right)^{1/2}}\sim t\left(m+n-2\right)$$

- 令

$$Z=\frac{\overline{X}_m-\overline{Y}_n}{\left(\frac{1}{m}+\frac{1}{n}\right)^{1/2}\sigma}, W=\frac{S_X^2+S_Y^2}{\sigma^2}$$

那么

$$U=\frac{Z}{\left(\frac{W}{m+n-2}\right)^{1/2}}$$

根据假设, $X_1,\ldots,X_m$ 与 $Y_1,\ldots,Y_n$ 相互独立, 因此 $\overline{X}_m$ 与 S_Y^2 相互独立, $\overline{Y}_n$ 与 S_X^2 相互独立, 又知 $\overline{X}_m$ 与 S_X^2 相互独立, $\overline{Y}_n$ 与 S_Y^2

相互独立. 从而 Z 与 W 相互独立. 容易看出 $Z \sim \mathcal{N}(0,1)$, $W \sim \chi^2(m+n-2)$, 因此 $U \sim t(m+n-2)$.

11.6 F 分布

定义 11.6.1 假设 m, n 为正整数, Y 与 W 相互独立. $Y \sim \chi^2(m)$, $W \sim \chi^2(n)$. 称

$$X = \frac{Y/m}{W/n}$$

的分布为自由度为 m 和 n 的 F 分布, 记为 $X \sim F(m, n)$.

定理 11.6.1 (1) 若 $X \sim F(m, n)$. 那么 $1/X \sim F(n, m)$. (2) 若 $X \sim t(n)$. 那么 $X^2 \sim F(1, n)$.

■ (1) 直接由 F 分布的定义得到. (2) 由 t 分布以及 F 分布的定义得到.