概 率 论 与 数 理 统 计

Probability and Statistics

南京大学 张绍群 & 徐科

更新： November 20, 2025

Part I

统计 (Statistics)

Part IX – Ch09: 统计的基本概念

Ch09: 统计的基本概念

Basic Concepts of Statistics

November 20, 2025

提纲

- 总体 vs 个体
- 统计量
  - 样本均值、样本方差、样本矩
  - 次序统计量
- 常用的三个统计分布
  - Beta 分布、Dirichlet 分布、Gamma 分布
- 三大抽样分布
  - χ^2 分布、t 分布、F 分布

引言

之前的课程属于概率论的范畴. 随机变量及其概率分布全面地描述了随机现象的统计性质. 在概率论的许多问题中, 随机分布被假定为已知的, 而一切的计算及推理都基于已知的分布函数进行. 但在实际问题中, 情况往往并非如此.

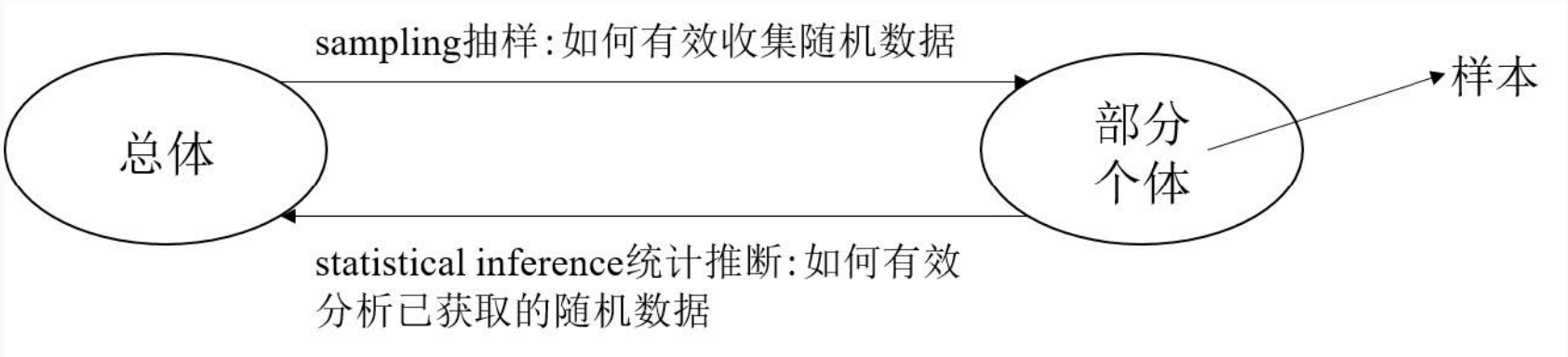
例如调查一批产品的不合格率, 某个省份的人均收入等诸如此类的问题, 是无法事先知道其分布函数的. 这类问题属于统计学的范畴. 一般认为, 统计学是一门研究如何有效地收集和分析所获数据的统计规律的学科. 统计学的研究内容包括: 抽样调查、参数估计、假设检验等.

基本概念

从总体中随机抽取一些个体, 表示为 $X_1, X_2, \ldots, X_n$, 称 $X_1, X_2, \ldots, X_n$ 为取自总体 X 的随机样本, 其样本容量为 n

- 总体: 研究对象的全体, 用随机变量 X 表示 (分布未知)
- 抽样: 抽取样本的过程
- 样本值: 观察样本得到的数值, 例如: $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ 为样本观察值或样本值
- 样本的二重性:
  - 就一次观察而言, 样本值是确定的数值
  - 不同的抽样下, 样本值会发生变化, 某次 sampling 可看作随机变量

基本概念 – 示图


sampling抽样：如何有效收集随机数据
总体
部分
个体
样本
statistical inference统计推断：如何有效
分析已获取的随机数据

简单随机样本

定义 0.1 (简单随机样本) 样本 $X_1, X_2, \ldots, X_n$ 是总体 X 的 简单随机样本, 简称 样本, 如果 $X_1, X_2, \ldots, X_n$ 满足

- 代表性: X_i 与 X 同分布
- 独立性: $X_1, X_2, \ldots, X_n$ 之间相互独立

Remarks: 课程后面所考虑的样本均为简单随机样本.

样本的分布

定义 0.2 (因为独立性) 总体 X 的联合分布函数为 $F(x), x_1, x_2, \ldots, x_n$ 为取自该总体的容量为 n 的样本, 则样本的联合分布函数为

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

若总体 X 的概率密度为 $f(x)$, 则样本 $x_1, x_2, \ldots, x_n$ 的联合概率密度为

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i) \, .$$

若总体 X 的分布列 $P(X = x_i)$, 则样本 $x_1, x_2, \ldots, x_n$ 的联合分布列为

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i) \, .$$

统计量

样本来自于总体, 因此样本中含有总体各方面的信息. 为将这些分散在样本中的有关总体的信息集中起来以反映总体的各种特征, 需要对样本进行加工, 较有效的加工方法是构建样本的函数, 不同样本函数反映总体的不同特征.

定义 0.3 设 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的一个样本, 若 $T = g(X_1, X_2, \ldots, X_n)$ 是一个连续、且不含任意参数的函数. 称 T 是一个统计量.

统计量具有以下性质:

- 因为 $X_1, X_2, \ldots, X_n$ 是随机变量, 所以 $g(X_1, X_2, \ldots, X_n)$ 是随机变量;
- $g(x_1, x_2, \ldots, x_n)$ 为 $g(X_1, X_2, \ldots, X_n)$ 的一次观察值.

Remarks: g 是定义在随机变量 $(X_1, X_2, \ldots, X_n)$ 上的.

统计量 – 样本均值及样本方差

定义 0.4 设 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的样本, 其算术平均值称为样本均值, 一般用 $\bar{X}$ 表示, 即

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \,.$$

样本关于它本身的样本均值 $\bar{X}$ 的平均偏差平方和称为 样本方差, 即

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

其算术根 $S_n = \sqrt{S_n^2}$ 称为 样本标准差. 当 n 不大时, 常用

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

作为样本方差 (也称 无偏方差).

无偏方差解释

直觉原因：样本均值靠得太近 $\rightarrow$ 方差被系统性压缩

真实方差围绕总体均值 μ，偏差为 $(X_i-\mu)^2$；但样本方差使用的是围绕样本均值的偏差 $(X_i-\bar{X})^2$。由于 $\bar{X}$ 是根据样本计算出来的，它会被样本“拉过去”，使偏差自然变小，导致方差被低估。严格可证样本方差 S_n^2 的期望是 $\mathbb{E}[S_n^2]=\frac{n-1}{n}\sigma^2$，因此低估比例恰好为 $\frac{n-1}{n}$。

自由度解释：偏差只能自由变化 $n-1$ 个

偏差满足 $\sum_{i=1}^{n}(X_i-\bar{X})=0$，意味着只要前 $n-1$ 个偏差确定，最后一个偏差就被强制决定。有效自由度因此只有 $n-1$ 个；但若仍用 n 去平均，就会把“$n-1$ 个自由信息”分摊到 n，从而压缩方差。修正方法就是除以 $n-1$，即无偏方差 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$。

样本均值及样本方差的性质

- 设总体 X 的期望 $\mathbb{E}[X] = \mu$ 以及方差 $\mathbb{VAR}(X) = \sigma^2$, 则有

$$\mathbb{E}[\bar{x}] = \mu, \quad \mathbb{VAR}(\bar{x}) = \sigma^2/n, \quad \bar{x} \xrightarrow{d} \mathcal{N}(\mu, \sigma^2/n)$$

- 样本方差 S_n^2 与总体方差 σ^2 之间存在偏差, 即

$$\mathbb{E}[S_n^2] = \frac{n-1}{n}\sigma^2$$

- 无偏方差 S^2 与总体方差 σ^2 相等

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

Remarks: 在实际中, S^2 比 S_n^2 更常用, 因此以后讲样本方差通常是指 S^2. (可证)

样本均值及样本方差的性质 – Proof

$$
\begin{aligned}
\mathbb{E}[S_n^2] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i^2-2X_i\bar{X}+\bar{X}^2)\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2\right] + \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}-2X_i\bar{X}\right] + \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\bar{X}^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2\right] - \frac{2}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}(X_jX_i)\right] + \mathbb{E}\left[\bar{X}^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2\right] - \frac{2}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_iX_i)\right] - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}(X_jX_i) \\
&\quad + \frac{1}{n^2}\left[\sum_{i=1}^{n}\mathbb{E}(X_i^2)\right] + \frac{1}{n^2}\left[\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}(X_jX_i)\right] \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}
$$

样本均值及样本方差的性质 – Proof

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

统计量：例 0.1

例 0.1 在一批产品中随机检查了 10 箱, 发现每箱中的不合格品数为

$$4, 5, 6, 0, 3, 1, 4, 2, 1, 4$$

试计算其样本均值、样本方差和样本标准差.

解答：例 0.1

题目: 在一批产品中随机检查了 10 箱, 发现每箱中的不合格品数为

$$4, 5, 6, 0, 3, 1, 4, 2, 1, 4$$

试计算其样本均值、样本方差和样本标准差.

解答:

- 根据样本均值、样本方差和样本标准差的定义直接计算, 即

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{4 + 5 + \cdots + 4}{10} = 3.$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{9}[(4-3)^2 + (5-3)^2 + \cdots + (4-3)^2] = 3.78.$$

$$s = \sqrt{s^2} = 1.94.$$

统计量：例 0.2

例 0.2 设总体 $X \sim \mathcal{N}(20, 3)$, 从总体中抽取两独立样本, 容量分别为 10 和 15. 求这两个样本均值之差的绝对值大于 0.3 的概率.

解答：例 0.2

题目: 设总体 $X \sim \mathcal{N}(20, 3)$, 从总体中抽取两独立样本, 容量分别为 10 和 15. 求这两个样本均值之差的绝对值大于 0.3 的概率.

解答:

- 设 $x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}$ 和 $x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)}$ 分别为来自总体 $X \sim \mathcal{N}(20, 3)$ 的两个独立样本. 根据正态分布的性质有

$$\bar{x^{(1)}} = \frac{1}{10}\sum_{i=1}^{10} x_i^{(1)} \sim \mathcal{N}(20, 3/10), \quad \bar{x^{(2)}} = \frac{1}{15}\sum_{i=1}^{15} x_i^{(2)} \sim \mathcal{N}(20, 1/5)$$

- 进一步根据正态分布的性质有 $\bar{x^{(1)}} - \bar{x^{(2)}} \sim \mathcal{N}(0, 1/2)$, 于是可得

$$P(|\bar{x^{(1)}} - \bar{x^{(2)}}| > 0.3) = 2 - 2\Phi(0.3/\sqrt{1/2}).$$

统计量 – 样本矩

样本均值和样本方差的更一般推广是样本矩, 这是一类常见的统计量.

定义 0.5 设 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的样本, k 为正整数, 则统计量

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

称为 样本 k 阶原点矩. 特别的, 样本一阶原点矩就是样本均值. 统计量

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

称为 样本 k 阶中心矩. 特别的, 样本二阶中心矩就是样本方差.

次序统计量

除了样本矩以外, 另外一类常见的统计量是次序统计量. 它在实际应用及理论中都有广泛的应用.

定义 0.6 设 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的样本, $X_{(i)}$ 称为该样本的第 i 个次序统计量, 它的取值是将样本观测值从小到大排序后得到的第 i 个观测值. 其中

$$X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$$

称为该样本的 最小次序统计量.

$$X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$$

称为该样本的最大次序统计量. $R_n = X_{(n)} - X_{(1)}$ 称为样本极差.

单个次序统计量的分布

定理 0.1 设总体 X 的密度函数为 $f(x)$, 分布函数为 $F(x)$, $X_1, X_2, \ldots, X_n$ 是来自总体 X 的样本, 则第 k 个次序统计量 $X_{(k)}$ 的分布函数和密度函数分别为

$$F_k(x) = \sum_{r=k}^{n} \binom{n}{r} [F(x)]^r [1 - F(x)]^{n-r} ,$$

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

Remarks:

- 次序统计量 $X_{(k)}$ 表示 $X_1, X_2, \ldots, X_n$ 中有 k 个变量小于等于 $X_{(k)}$
- f_k 理解为 $X_{(k)}$ 在 x 附近的小区间 $(x, x + dx)$ 内的事件
- 令 $k = 1$ 和 $k = n$, 分别得到最小次序统计量和最大次序统计量的分布函数和密度函数

证明: 单个次序统计量的分布（应用莱布尼茨定理）

证明 根据题意有第 k 次序统计量 $X_{(k)}$ 的分布函数为

$$\begin{aligned} F_k(x) &= \Pr[X_{(k)} \leqslant x] = \Pr[X_1, X_2, \cdots, X_n \text{ 中至少有 } k \text{ 个随机变量 } \leqslant x] \\ &= \sum_{r=k}^{n} \Pr[X_1, X_2, \cdots, X_n \text{ 中恰有 } r \text{ 个随机变量 } \leqslant x,\ n-r \text{ 个随机变量 } > x] \\ &= \sum_{r=k}^{n} \binom{n}{r} [F(x)]^r [1-F(x)]^{n-r}. \end{aligned}$$

利用恒等式

$$\sum_{r=k}^{n} \binom{n}{r} p^r (1-p)^{n-r} = \frac{n!}{(k-1)!(n-k)!} \int_0^p t^{k-1}(1-t)^{n-k} dt \quad (r \in [n],\ p \in [0,1])$$

由此可知

$$F_k(x) = \frac{n!}{(k-1)!(n-k)!} \int_0^{F(x)} t^{k-1}(1-t)^{n-k} dt,$$

根据积分函数求导完成证明.

证明: 单个次序统计量的分布（应用莱布尼茨定理）

令

$$G(p)=\sum_{r=k}^{n}\binom{n}{r}p^r(1-p)^{n-r},\qquad 0\le p\le 1.$$

显然 $G(0)=0$。对 $G(p)$ 求导，有

$$\begin{aligned}
G'(p)&=\sum_{r=k}^{n}\binom{n}{r}\left[rp^{r-1}(1-p)^{n-r}-(n-r)p^r(1-p)^{n-r-1}\right]\\
&=\sum_{r=k}^{n}n\binom{n-1}{r-1}p^{r-1}(1-p)^{n-r}-\sum_{r=k}^{n}n\binom{n-1}{r}p^r(1-p)^{n-r-1}\\
&=n\sum_{j=k-1}^{n-1}\binom{n-1}{j}p^j(1-p)^{n-1-j}-n\sum_{j=k}^{n-1}\binom{n-1}{j}p^j(1-p)^{n-1-j}\\
&=n\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k}=\frac{n!}{(k-1)!(n-k)!}p^{k-1}(1-p)^{n-k}.
\end{aligned}$$

因此

$$G(p)=G(0)+\int_0^p G'(t)\,dt=\frac{n!}{(k-1)!(n-k)!}\int_0^p t^{k-1}(1-t)^{n-k}\,dt,$$

即

$$\sum_{r=k}^{n}\binom{n}{r}p^r(1-p)^{n-r}=\frac{n!}{(k-1)!(n-k)!}\int_0^p t^{k-1}(1-t)^{n-k}\,dt.$$

小结

通过前面的学习, 我们知道样本统计量也是一个随机变量. 以样本均数为例, 假设我们现在想了解南京大学男性学生的身高情况, 按照同样的方法重复 50 次抽样, 每次抽 100 人, 每组样本都可以计算一个样本均数, 假设分别为: 1.76, 1.72, 1.69, 1.77, . . . , 1.75, 样本均数会随着抽样的不同而随机变动.

进一步, 我们可以将样本统计量作为随机变量研究其概率分布 (又称“抽样分布”), 从而得到其分布的性质或计算特定情况下的概率.

本课程主要研究的抽样分布通常是样本均值的分布、样本方差的分布、样本标准差的分布.

Beta (贝塔) 分布

定义 0.7 (Beta 函数) 对任意给定 $\alpha_1 > 0$ 和 $\alpha_2 > 0$, 定义 Beta 函数为

$$\text{Beta}(\alpha_1, \alpha_2) = \int_0^1 x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1}\,\mathrm{d}x,$$

简记为 $B(\alpha_1, \alpha_2)$, 被称为第一类欧拉积分函数.

定义 0.8 给定 $\alpha_1 > 0$ 和 $\alpha_2 > 0$, 若随机变量 X 的概率密度函数为

$$f(x) = \begin{cases} \frac{1}{B(\alpha_1,\alpha_2)} x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1}, & x \in (0,1) \\ 0, & \text{其它} \end{cases}$$

称 X 服从参数为 α_1 和 α_2 的 Beta 分布, 记为 $X \sim B(\alpha_1, \alpha_2)$.

Beta (贝塔) 分布的数字特征

定理 0.2 若随机变量 $X \sim B(\alpha_1, \alpha_2)$, 则有

$$\mathbb{E}[X] = \frac{\alpha_1}{\alpha_1 + \alpha_2}, \quad \mathbb{VAR}(X) = \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)}.$$

由期望定义得到

$$\mathbb{E}[X] = \int_0^1 x^{\alpha_1}(1-x)^{\alpha_2 - 1}\, dx = B(\alpha_1 + 1, \alpha_2),$$

故

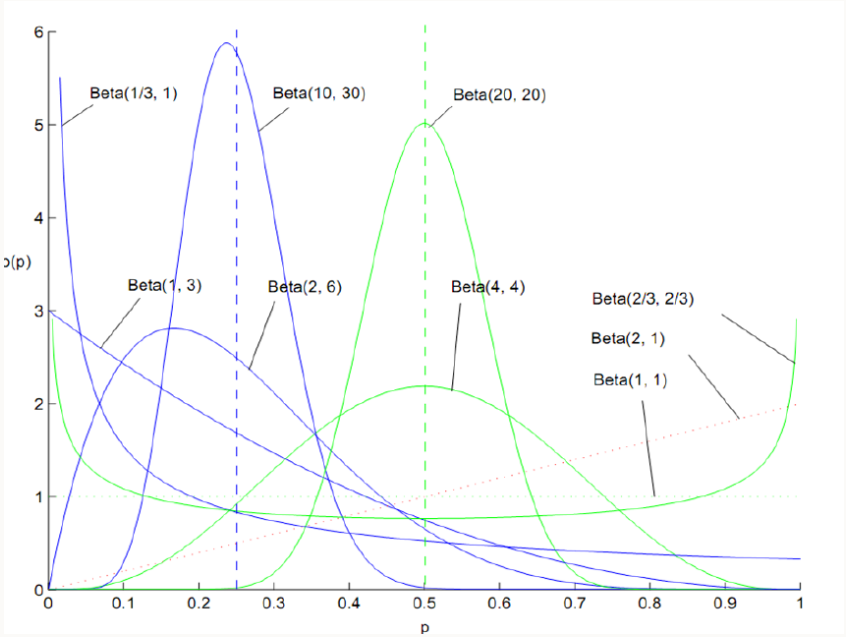
$$\mathbb{E}[X] = \frac{B(\alpha_1 + 1, \alpha_2)}{B(\alpha_1, \alpha_2)}.$$

利用 $B(a,b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ 以及 $\Gamma(a+1) = a\Gamma(a)$, 可得

$$\mathbb{E}[X] = \frac{\alpha_1}{\alpha_1 + \alpha_2}.$$

Beta (贝塔) 分布的性质

Beta 分布的概率密度我们把它画成图, 会发现它是个百变星君, 它可以是凹的、凸的、单调上升的、单调下降的; 可以是曲线也可以是直线, 而均匀分布也是特殊的分布. 由于分布能够拟合如此之多的形状, 因此它在统计数据拟合和贝叶斯分析中被广泛使用.



Dirichlet (狄利克雷) 分布

定义 0.9 给定 $\alpha_1, \ldots, \alpha_k \in (0, +\infty)$, 若多元随机向量 $X = (X_1, \ldots, X_k)$ 的概率密度函数为

$$f(x_1, x_2, \ldots, x_k) = \begin{cases} \frac{x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \ldots x_k^{\alpha_k - 1}}{B(\alpha_1, \alpha_2, \ldots, \alpha_k)}, & \sum_{i=1}^{k} x_i = 1 \text{且} x_i > 0 (i \in [k]) \\ 0, & \text{其它} \end{cases}$$

称 X 服从参数为 $\alpha_1, \ldots, \alpha_k$ 的 Dirichlet 分布, 记为 $X \sim \text{Dir}(\alpha_1, \ldots, \alpha_k)$.

Dirichlet 分布是 Beta 分布的一种高维推广. 当 $k = 2$ 时, Dirichlet 分布退化为 Beta 分布.

Dirichlet (狄利克雷) 分布的数字特征

定理 0.3 若随机变量 $X \sim \mathrm{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_k)$, 设 $\tilde{\alpha} = \alpha_1 + \alpha_2 + \cdots + \alpha_k$ 和 $\tilde{\alpha}_i = \alpha_i / \tilde{\alpha}$ 则有

$$\mathbb{E}[X_i] = \tilde{\alpha}_i, \quad \mathbb{VAR}(X_i, X_j) = \begin{cases} \frac{\tilde{\alpha}_i(1-\tilde{\alpha}_i)}{\tilde{\alpha}+1}, & i = j \\ -\frac{\tilde{\alpha}_i \tilde{\alpha}_j}{\tilde{\alpha}+1}, & i \neq j \end{cases}$$

Remarks:

- $k = 2$ 时，$\mathbb{E}[X_i]$ 和 $\mathbb{VAR}(X_i, X_j)$ 就是 Beta 对应计算值
- 因为 Dirichlet 分布描述的是一个总和必须等于 1 的概率向量。所以如果某个 X_i 增大，那其他分量必须减少一点做补偿。因此不同分量之间天然是负相关的

Gamma (伽马) 分布

定义 0.10 (Γ 函数) 对任意给定 $\alpha > 0$, 定义 Γ 函数为

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} \, \mathrm{d}x,$$

被称为第二类欧拉积分函数.

定义 0.11 若随机变量 X 的概率密度函数为

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

其中 $\alpha > 0$ 且 $\lambda > 0$, 称 X 服从参数为 α 和 λ 的 Γ 分布, 记为 $X \sim \Gamma(\alpha, \lambda)$.

Γ (伽马) 分布的数字特征

定理 0.4 若随机变量 $X \sim \Gamma(\alpha, \lambda)$, 则有

$$\mathbb{E}[X] = \alpha/\lambda\,, \quad \mathbb{VAR}(X) = \alpha/\lambda^2\,.$$

定理 0.5 (Gamma 分布的可加性) 若随机变量 $X \sim \Gamma(\alpha_1, \lambda)$ 和 $Y \sim \Gamma(\alpha_2, \lambda)$, 且 X 和 Y 相互独立, 则有 $X + Y \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$.

另外, 对比指数分布和伽马分布的密度函数形式, 易知 $\Gamma(1, \lambda) = e(\lambda)$.

Remarks:

- 定理0.5说明，如果 X 表示等到第 α_1 次事件的时间，Y 表示等到第 α_2 次事件的时间，那么 $X + Y$ 就是等到第 $\alpha_1 + \alpha_2$ 次事件的时间。—这就是 Poisson 过程的本质

数字特征的分布: 图像、数字特征、例子

- Beta 分布: 数字特征

$$\mathbb{E}[X] = \frac{\alpha_1}{\alpha_1 + \alpha_2}, \quad \mathbb{VAR}(X) = \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)}$$

- Dirichlet 分布: 数字特征

$$\mathbb{E}[X_i] = \tilde{\alpha}_i, \quad \mathbb{VAR}(X_i, X_j) = \begin{cases} \frac{\tilde{\alpha}_i(1-\tilde{\alpha}_i)}{\tilde{\alpha}+1}, & i = j \\ -\frac{\tilde{\alpha}_i \tilde{\alpha}_j}{\tilde{\alpha}+1}, & i \neq j \end{cases}$$

- Gamma 分布: 数字特征

$$\mathbb{E}[X] = \alpha/\lambda, \quad \mathbb{VAR}(X) = \alpha/\lambda^2$$

Beta 分布：例 0.3

例 0.3 设总体分布 $U(0,1)$, $x_1, x_2, \ldots, x_n$ 为样本, 试求第 k 个次序统计量 $x_{(k)}$ 的密度函数.

解答：例 0.3

题目: 设总体分布 $U(0,1), x_1, x_2, \ldots, x_n$ 为样本, 试求第 k 个次序统计量 $x_{(k)}$ 的密度函数.

解答:

- 由题易知总体 X 的分布函数 $F(x) = x, x \in (0,1)$ 和密度函数 $f(x) = 1, x \in (0,1)$, 进一步根据定理 0.1, 可得

$$\begin{aligned} p_k(x) &= \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k} \\ &= \frac{1}{B(k, n-k+1)} x^{k-1}(1-x)^{n-k}, \qquad 0 < x < 1. \end{aligned}$$

- 即贝塔分布 $B(k, n-k+1)$.

解答：例 0.3

本页求证

$$\int_0^1 C_{n-1}^{k-1} x^{k-1}(1-x)^{n-k}dx = \frac{1}{n} \quad \text{for any} \quad k\,.$$

Proof: 分布积分

$$\begin{aligned}\int_0^1 C_{n-1}^{k-1} x^{k-1}(1-x)^{n-k}dx &= C_{n-1}^{k-1}\left[\frac{x^k(1-x)^{n-k}}{k}\ |_0^1 - \int_0^1 -\frac{x^k}{k}(n-k)(1-x)^{n-k-1}\,\mathrm{d}x\right] \\ &= \int_0^1 C_{n-1}^{k} x^{k}(1-x)^{n-k-1}\,\mathrm{d}x \\ &= \int_0^1 x^{n-1}\,\mathrm{d}x \\ &= \frac{1}{n}\end{aligned}$$

因此，有

$$B(k, n-k+1) = \int_0^1 x^{k-1}(1-x)^{n-k}dx = \frac{1}{n}\frac{1}{C_{n-1}^{k-1}}\,.$$

Gamma 分布：例 0.4

例 0.4 若随机变量 $X \sim \mathcal{N}(0,1)$, 则有 $X^2 \sim \Gamma(1/2, 1/2)$.

解答：例 0.4

题目: 若随机变量 $X \sim \mathcal{N}(0,1)$, 则有 $X^2 \sim \Gamma(1/2, 1/2)$.

解答:

- 由题易知总体 $Y = X^2$ 的分布函数为: 当 $y \leq 0$ 时, 有 $F_Y(y) = 0$; 当 $y > 0$ 时, 有

$$F_Y(y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \,\mathrm{d}x,$$

  由此可得概率密度函数为 $f_Y(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-\frac{y}{2}}$, 从而得到 $X^2 \sim \Gamma(1/2, 1/2)$.
- 利用高斯积分

$$\Gamma\left(\tfrac{1}{2}\right) = \int_0^\infty x^{-1/2} e^{-x}\, dx = \int_0^\infty \frac{1}{t} e^{-t^2} (2t\, dt) = 2\int_0^\infty e^{-t^2} dt,$$

$$\int_{-\infty}^{\infty} e^{-t^2} dt = \sqrt{\pi} \quad \Longrightarrow \quad \Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi}.$$

三大 (运算) 抽样分布

中心极限定理揭示了样本均值的分布, 即不管总体是什么分布, 任意一个总体的样本平均值都会围绕在总体的平均值周围, 并且呈正态分布. 在数理统计中, 用于描述抽样分布的分布函数, 除了正态分布外, 最重要的三个分布分别是:

- χ^2 (卡方) 分布
- t 分布
- F 分布

这三个以标准正态分布而构造的统计量在统计推断中有广泛的应用, 不仅是因为这三个统计量有明确的构造背景, 而且其抽样分布的密度函数都有显性表达式, 它们被称为统计中的 “三大抽样分布”.

χ^2 (卡方) 分布及其密度函数

定义 0.12 若 $X_1, X_2, \ldots, X_n$ 是来自总体 $X \sim \mathcal{N}(0,1)$ 的一个样本, 称 $Y = X_1^2 + X_2^2 + \cdots + X_n^2$ 为服从自由度为 n 的 χ^2 分布, 记 $Y \sim \chi^2(n)$.

根据 $X_1^2 \sim \Gamma(1/2, 1/2)$ 和 Γ 函数的可加性可得 $Y \sim \Gamma(n/2, 1/2)$. 于是有随机变量 Y 的概率密度函数为

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

χ^2 分布的性质

- 若随机变量 $X \sim \chi^2(n)$, 则 $\mathbb{E}[X] = n$ 和 $\mathbb{VAR}(X) = 2n$;
- 若随机变量 $X \sim \chi^2(m)$ 和 $Y \sim \chi^2(n)$ 相互独立, 则 $X + Y \sim \chi^2(m + n)$;
- 推广命题：若随机变量 $X \sim \mathcal{N}(0, 1)$, 则

$$\mathbb{E}[X^k] = \begin{cases} (k-1)!!, & k \text{ 为偶数} \\ 0, & k \text{ 为奇数} \end{cases}$$

其中,

$$\begin{cases} (2k)!! = 2k \cdot (2k-2) \cdots 2 \,, \\ (2k+1)!! = (2k+1) \cdot (2k-1) \cdots 1 \,, \end{cases}$$

χ^2 分布：例 0.5

例 0.5 若 X_1, X_2, X_3, X_4 是来自总体 $X \sim \mathcal{N}(0, 4)$ 的样本, 以及

$$Y = a(X_1 - 2X_2)^2 + b(3X_3 - 4X_4)^2$$

求 a, b 取何值时, Y 服从 χ^2 分布, 并求其自由度.

解答：例 0.5

题目: 若 X_1, X_2, X_3, X_4 是来自总体 $X \sim \mathcal{N}(0, 4)$ 的样本, 以及

$$Y = a(X_1 - 2X_2)^2 + b(3X_3 - 4X_4)^2$$

求 a, b 取何值时, Y 服从 χ^2 分布, 并求其自由度.

解答:

- 根据正态分布的性质有 $X_1 - 2X_2 \sim \mathcal{N}(0, 20)$ 和 $3X_3 - 4X_4 \sim \mathcal{N}(0, 100)$, 因此

$$\frac{X_1 - 2X_2}{2\sqrt{5}} \sim \mathcal{N}(0, 1), \qquad \frac{3X_3 - 4X_4}{10} \sim \mathcal{N}(0, 1),$$

  所以当 $a = 1/20, b = 1/100$ 时有 $Y \sim \chi^2(2)$ 成立.

χ^2 分布：例 0.6

例 0.6 相互独立的随机变量 $X_1, X_2, \ldots, X_n$ 满足 $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, 求 $\sum_{i=1}^{n}(X_i - \mu_i)^2/\sigma_i^2$ 的分布.

解答：例 0.6

题目: 相互独立的随机变量 $X_1, X_2, \ldots, X_n$ 满足 $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, 求 $\sum_{i=1}^n (X_i - \mu_i)^2/\sigma_i^2$ 的分布.

解答:

- 令 $Y_i = (X_i - \mu_i)/\sigma_i$, 由题意知 $Y_1, Y_2, \ldots, Y_n$ 是独立同分布的随机变量, 其共同分布为 $\mathcal{N}(0, 1)$, 于是由定义 0.12可知

$$Y = \sum_{i=1}^n Y_i^2 = \sum_{i=1}^n (X_i - \mu_i)^2/\sigma_i^2 \sim \chi^2(n)$$

  即 $\sum_{i=1}^n (X_i - \mu_i)^2/\sigma_i^2$ 服从自由度为 n 的 χ^2 分布.

分布可加性小结

- 若随机变量 $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ 和 $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ 相互独立, 那么

$$X \pm Y \sim \mathcal{N}(\mu_1 \pm \mu_2, \sigma_1^2 + \sigma_2^2)$$

- 若随机变量 $X \sim B(n_1, p)$ 和 $Y \sim B(n_2, p)$ 相互独立, 那么

$$X + Y \sim B(n_1 + n_2, p)$$

- 若随机变量 $X \sim P(\lambda_1)$ 和 $Y \sim P(\lambda_2)$ 相互独立, 那么

$$X + Y \sim P(\lambda_1 + \lambda_2)$$

- 若随机变量 $X \sim \Gamma(\alpha_1, \lambda)$ 和 $Y \sim \Gamma(\alpha_2, \lambda)$ 相互独立, 那么

$$X + Y \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$$

- 若随机变量 $X \sim \chi^2(m)$ 和 $Y \sim \chi^2(n)$ 相互独立, 则 $X+Y \sim \chi^2(m+n)$.

t 分布 (student distribution) 及其密度函数

定义 0.13 随机变量 $X \sim \mathcal{N}(0,1)$ 和 $Y \sim \chi^2(n)$ 相互独立, 则随机变量

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

服从自由度为 n 的 t 分布, 记 $T \sim t(n)$.

随机变量 $T \sim t(n)$ 的概率密度为（具有对称性）

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \qquad x \in (-\infty, +\infty).$$

当 $n \to \infty$ 时, 随机变量 $T \sim t(n)$ 的概率密度为

$$f(x) \to \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

因此当 n 足够大时,$f(x)$ 可被近似为 $\mathcal{N}(0,1)$ 的密度函数.

t 分布：例 0.7

例 0.7 设 $X_1, X_2, \ldots, X_9$ 和 $Y_1, Y_2, \ldots, Y_9$ 是分别来自于总体 $\mathcal{N}(0, 9)$ 的两个独立样本, 求 $(X_1 + X_2 + \cdots + X_9)/\sqrt{Y_1^2 + Y_2^2 + \cdots + Y_9^2}$ 的分布.

解答：例 0.7

题目: 设 $X_1, X_2, \ldots, X_9$ 和 $Y_1, Y_2, \ldots, Y_9$ 是分别来自于总体 $\mathcal{N}(0, 9)$ 的两个独立样本, 求 $(X_1 + X_2 + \cdots + X_9)/\sqrt{Y_1^2 + Y_2^2 + \cdots + Y_9^2}$ 的分布.

解答:

- 由题意知 $X_1 + X_2 + \cdots + X_9 \sim \mathcal{N}(0, 81)$, 标准化后可得 $X = \frac{X_1 + X_2 + \cdots + X_9}{9} \sim \mathcal{N}(0, 1)$; 同时, 根据定义 0.12可知, $Y = \left(\frac{Y_1}{3}\right)^2 + \left(\frac{Y_2}{3}\right)^2 + \cdots + \left(\frac{Y_9}{3}\right)^2 \sim \chi^2(9)$, 显然 X 与 Y 独立, 根据定义 0.13可知

$$\frac{\frac{X_1 + X_2 + \cdots + X_9}{9}}{\sqrt{\left(\left(\frac{Y_1}{3}\right)^2 + \left(\frac{Y_2}{3}\right)^2 + \cdots + \left(\frac{Y_9}{3}\right)^2\right)/9}} = \frac{(X_1 + X_2 + \cdots + X_9)}{\sqrt{Y_1^2 + Y_2^2 + \cdots + Y_9^2}} \sim t(9).$$

  即 $(X_1 + X_2 + \cdots + X_9)/\sqrt{Y_1^2 + Y_2^2 + \cdots + Y_9^2}$ 服从自由度为 9 的 t 分布.

F 分布及其密度函数

定义 0.14 随机变量 $X \sim \chi^2(m)$ 和 $Y \sim \chi^2(n)$ 相互独立, 则随机变量

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

服从自由度为 (m, n) 的 F 分布, 记 $Z \sim F(m, n)$.

随机变量 $Z \sim F(m, n)$ 的概率密度为

$$f(z) = \begin{cases} \dfrac{\Gamma\left(\frac{m+n}{2}\right)\left(\frac{m}{n}\right)^{\frac{m}{2}} z^{\frac{m}{2}-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)\left(1+\frac{mx}{n}\right)^{\frac{m+n}{2}}}, & z > 0 \\ 0, & z \leq 0 \end{cases}$$

若随机变量 $Z \sim F(m, n)$, 则 $\frac{1}{Z} \sim F(n, m)$.

F 分布：例 0.8

例 0.8 设 $X_1, X_2, \ldots, X_{2n}$ 是来自于总体 $\mathcal{N}(0, \sigma^2)$ 的样本, 求

$$(X_1^2 + X_3^2 + \cdots + X_{2n-1}^2)/(X_2^2 + X_4^2 + \cdots + X_{2n}^2)$$

的分布.

解答：例 0.8

题目: 设 $X_1, X_2, \ldots, X_{2n}$ 是来自于总体 $\mathcal{N}(0, \sigma^2)$ 的样本, 求

$$(X_1^2 + X_3^2 + \cdots + X_{2n-1}^2)/(X_2^2 + X_4^2 + \cdots + X_{2n}^2)$$

的分布.

解答:

- 由题意知 $\frac{X_i}{\sigma} \sim \mathcal{N}(0,1)$, 根据定义0.12可知, $A = \left(\frac{X_1}{\sigma}\right)^2 + \left(\frac{X_3}{\sigma}\right)^2 + \cdots + \left(\frac{X_{2n-1}}{\sigma}\right)^2 \sim \chi^2(n)$, $B = \left(\frac{X_2}{\sigma}\right)^2 + \left(\frac{X_4}{\sigma}\right)^2 + \cdots + \left(\frac{X_{2n}}{\sigma}\right)^2 \sim \chi^2(n)$, 显然 A 与 B 独立, 根据0.14可知

$$\frac{A/n}{B/n} = \frac{[(X_1^2 + X_3^2 + \cdots + X_{2n-1}^2)/n^2]/n}{[(X_2^2 + X_4^2 + \cdots + X_{2n}^2)/n^2]/n} \sim F(n,n).$$

  即 $(X_1^2 + X_3^2 + \cdots + X_{2n-1}^2)/(X_2^2 + X_4^2 + \cdots + X_{2n}^2)$ 服从自由度为 (n,n) 的 F 分布.

三大抽样分布小结

若 $X_1, X_2, \ldots, X_n$ 和 $Y_1, Y_2, \ldots, Y_n$ 是来自标准正态分布的两个相互独立的样本, 因此三个统计量的构造及抽样分布如下表所示.

统计量的构造	抽样分布密度函数	期望	方差
$\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$	$f_Y(y) = \frac{(\frac{1}{n})^{\frac{n}{2}}}{\Gamma(n/2)} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}, (y > 0)$	n	$2n$
$F = \frac{(Y_1^2+Y_2^2+\cdots+Y_m^2)/m}{(X_1^2+X_2^2+\cdots+X_n^2)/n}$	$f_Y(y) = \frac{\Gamma\left(\frac{m+n}{2}\right)\left(\frac{m}{n}\right)^{\frac{m}{2}} y^{\frac{m}{2}-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)\left(1+\frac{mx}{n}\right)^{\frac{m+n}{2}}}, (y > 0)$	$\frac{n}{n-2}, (n > 2)$	$\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, (n > 4)$
$t = \frac{Y_1}{\sqrt{(X_1^2+X_2^2+\cdots+X_n^2)/n}}$	$f_Y(y) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} \left(1 + \frac{y^2}{n}\right)^{-\frac{n+1}{2}}$	$0, (n > 1)$	$\frac{n}{n-2}, (n > 2)$

Sampling Distribution of Gaussian

November 20, 2025

提纲

- 正态分布的抽样分布
- 分位数
  - 正态分布、χ^2 分布、t 分布、F 分布
- 如何考察抽样分布?

正态分布的抽样分布

定理 0.6 设 $X_1, X_2, \ldots, X_n$ 是来自总体 $\mathcal{N}(\mu, \sigma^2)$ 的样本, 则有

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \sigma^2/n), \qquad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

定理 0.7 设 $X_1, X_2, \ldots, X_n$ 是来自总体 $\mathcal{N}(\mu, \sigma^2)$ 的样本, 其样本均值和无偏样本方差分别为

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

则有

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)\,.$$

正态分布的抽样分布

定理 0.8 设 $X_1, X_2, \ldots, X_n$ 是来自总体 $\mathcal{N}(\mu, \sigma^2)$ 的样本, 其样本均值和无偏样本方差分别为

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

则有 $\bar{X}$ 和 S^2 相互独立, 且

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

Proof: 正态分布的抽样分布定理 0.8

定义基于正交矩阵 $\mathbf{A}$ 的一线性变换

$$\mathbf{Y} = \mathbf{A}\mathbf{X} = \begin{pmatrix} 1/\sqrt{n} & 1/\sqrt{n} & 1/\sqrt{n} & \dots & 1/\sqrt{n} \\ \frac{1}{\sqrt{2\cdot 1}} & \frac{-1}{\sqrt{2\cdot 1}} & 0 & \dots & 0 \\ \frac{1}{\sqrt{3\cdot 2}} & \frac{1}{\sqrt{3\cdot 2}} & \frac{-2}{\sqrt{3\cdot 2}} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{n\cdot(n-1)}} & \frac{1}{\sqrt{n\cdot(n-1)}} & \frac{1}{\sqrt{n\cdot(n-1)}} & \dots & \frac{1-n}{\sqrt{n\cdot(n-1)}} \end{pmatrix} \mathbf{X}\,,$$

其中, $\mathbf{Y}$ 和 $\mathbf{X}$ 分别为列向量. 由此, 可得 $Y_1 = \sqrt{n}\bar{X}$, 进而 $\mathbb{E}(Y_1) = \sqrt{n}\mu$ 和 $\mathbb{E}(Y_i) = 0$. 后者可以根据多变量的概率密度函数求得. 进而, 我们有

$$\begin{aligned} \frac{(n-1)S^2}{\sigma^2} &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \frac{1}{n}\sum_{j=1}^n X_j)^2}{\sigma^2} \\ &= \frac{\sum_{i=1}^n X_i^2 - (\sqrt{n}\bar{X})^2}{\sigma^2} = \frac{\sum_{i=1}^n Y_i^2 - Y_1^2}{\sigma^2} \\ &= \sum_{i=2}^n \left(\frac{Y_i}{\sigma}\right)^2 \sim \chi^2(n-1)\,. \end{aligned}$$

正态分布的抽样分布

定理 0.9 设 $X_1, X_2, \ldots, X_m$ 和 $Y_1, Y_2, \ldots, Y_n$ 分别来自总体 $\mathcal{N}(\mu_X, \sigma^2)$ 和 $\mathcal{N}(\mu_Y, \sigma^2)$ 的两个独立样本, 令其样本均值分别为 $\bar{X}$ 和 $\bar{Y}$, 无偏样本方差分别为 S_X^2 和 S_Y^2, 则有

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}}\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2).$$

定理 0.10 设 $X_1, X_2, \ldots, X_m$ 和 $Y_1, Y_2, \ldots, Y_n$ 分别来自总体 $\mathcal{N}(\mu_X, \sigma_X^2)$ 和 $\mathcal{N}(\mu_Y, \sigma_Y^2)$ 的两个独立样本, 令其无偏样本方差分别为 S_X^2 和 S_Y^2, 则有

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(m-1, n-1).$$

Proof: 正态分布的抽样分布定理0.9和0.10

1.

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma^2}{m}\right), \quad \bar{Y} \sim N\left(\mu_Y, \frac{\sigma^2}{n}\right) \quad \Rightarrow \quad Z = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim N(0,1).$$

$$\frac{(m-1)S_X^2}{\sigma^2} \sim \chi^2(m-1), \quad \frac{(n-1)S_Y^2}{\sigma^2} \sim \chi^2(n-1) \quad \Rightarrow \quad \frac{(m-1)S_X^2 + (n-1)S_Y^2}{\sigma^2} \sim \chi^2(m+n-2).$$

$$S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}, \qquad W = \frac{(m+n-2)S_p^2}{\sigma^2} \sim \chi^2(m+n-2).$$

$$T = \frac{Z}{\sqrt{W/(m+n-2)}} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2).$$

2.

$$\frac{(m-1)S_X^2}{\sigma_X^2} \sim \chi^2(m-1), \qquad \frac{(n-1)S_Y^2}{\sigma_Y^2} \sim \chi^2(n-1),$$

$$\frac{S_X^2}{\sigma_X^2} = \frac{\chi^2(m-1)}{m-1}, \qquad \frac{S_Y^2}{\sigma_Y^2} = \frac{\chi^2(n-1)}{n-1}.$$

$$\Rightarrow \quad \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} = \frac{\chi^2(m-1)/(m-1)}{\chi^2(n-1)/(n-1)} \sim F(m-1, n-1).$$

抽样分布：例 0.9

例 0.9 设随机变量 $T \sim t(n)$, 求 $Y = T^2$ 的分布.

解答：例 0.9

题目: 设随机变量 $T \sim t(n)$, 求 $Y = T^2$ 的分布.

解答:

- 根据 t 分布的定义 0.13可知, 随机变量 $T = \frac{X}{\sqrt{Z/n}}$ 服从自由度为 n 的 t 分布, 其中随机变量 $X \sim \mathcal{N}(0,1)$ 和 $Z \sim \chi^2(n)$ 相互独立. 因此

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

- 易知 $X^2 \sim \chi^2(1)$, $Z/n \sim \chi^2(n)/n$, 且 X^2 与 Z/n 相互独立, 因此 $Y = T^2 \sim F(1,n)$.

抽样分布：例 0.10

例 0.10 设 $X_1, X_2, \ldots, X_5$ 是来自 $\mathcal{N}(0,1)$ 的样本, 令 $Y = c_1(X_1 + X_3)^2 + c_2(X_2 + X_4 + X_5)^2$. 求常数 c_1, c_2 使 Y 服从 χ^2 分布.

解答：例 0.10

题目: 设 $X_1, X_2, \ldots, X_5$ 是来自 $\mathcal{N}(0,1)$ 的样本, 令 $Y = c_1(X_1+X_3)^2+c_2(X_2+X_4+X_5)^2$. 求常数 c_1, c_2 使 Y 服从 χ^2 分布.

解答:

- 随机变量 $\frac{X_1+X_3}{\sqrt{2}} \sim \mathcal{N}(0,1)$, 又 $\frac{X_2+X_4+X_5}{\sqrt{3}} \sim \mathcal{N}(0,1)$ 且相互独立, 根据 χ^2 分布的定义可得
$$\left(\frac{X_1+X_3}{\sqrt{2}}\right)^2 + \left(\frac{X_2+X_4+X_5}{\sqrt{3}}\right)^2 \sim \chi^2(2),$$
- 即 $c_1 = \frac{1}{2}, c_2 = \frac{1}{3}, Y \sim \chi^2(2)$.

抽样分布：例 0.11

例 0.11 设 X_1, X_2 是来自总体 $\mathcal{N}(0, \sigma^2)$ 的样本, 求 $\frac{(X_1+X_2)^2}{(X_1-X_2)^2}$ 的分布.

解答：例 0.11

题目: 设 X_1, X_2 是来自总体 $\mathcal{N}(0, \sigma^2)$ 的样本, 求 $\frac{(X_1+X_2)^2}{(X_1-X_2)^2}$ 的分布.

解答:

- 随机变量 $\frac{X_1+X_2}{\sqrt{2}\sigma} \sim \mathcal{N}(0,1)$, 又 $\frac{X_1-X_2}{\sqrt{2}\sigma} \sim \mathcal{N}(0,1)$ 且相互独立, 根据 χ^2 分布的定义可得

$$\left(\frac{X_1+X_2}{\sqrt{2}\sigma}\right)^2 \sim \chi^2(1), \quad \left(\frac{X_1-X_2}{\sqrt{2}\sigma}\right)^2 \sim \chi^2(1),$$

- 根据 F 分布的定义可得

$$\frac{(X_1+X_2)^2}{(X_1-X_2)^2} = \frac{\left(\frac{X_1+X_2}{\sqrt{2}\sigma}\right)^2}{\left(\frac{X_1-X_2}{\sqrt{2}\sigma}\right)^2} \sim F(1,1).$$

抽样分布小结

设 $X_1, X_2, \ldots, X_m$ 和 $Y_1, Y_2, \ldots, Y_n$ 分别来自总体 $\mathcal{N}(\mu_X, \sigma^2)$ 和 $\mathcal{N}(\mu_Y, \sigma^2)$ 的两个独立样本, 抽样分布如下表所示.

统计量	抽样分布
单个样本均值	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu_X, \sigma^2/n)$
单个样本方差	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
两个样本之差	$\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{(m-1)S_X^2+(n-1)S_Y^2}{m+n-2}}\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim t(m+n-2)$
两个样本的方差之比	$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(m-1, n-1)$

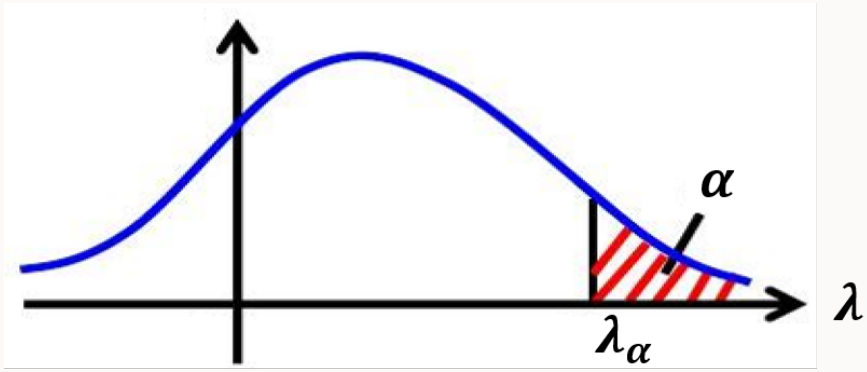
分位数 (点)

很多概率统计问题最后都归结为求解满足概率不等式 $1-F(x)>\alpha$ 的最大 x, 其解可用上侧分位数 λ_α 表示. 为此人们对常用的分布 (如正态分布、t 分布、χ^2 分布等) 编制了各种分位数表供实际使用.

定义 0.15 对给定 $\alpha\in(0,1)$ 和随机变量 X, 称满足

$$P(X>\lambda_\alpha)=\alpha$$

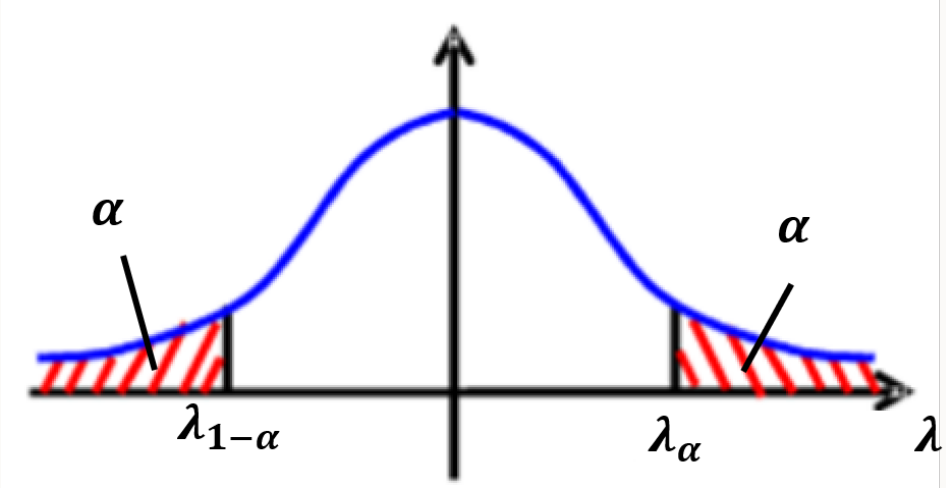
的实数 λ_α 为上侧 α 分位数 (点).



对称分布的分位数

定理 0.11 随机变量 X 的概率密度函数关于 y 轴对称, 则有

$$\lambda_{1-\alpha} = -\lambda_{\alpha}.$$



分位数：例 0.12

例 0.12 设 T_A, T_B 表示某厂生产的两种轴承 A, B 的寿命, 已知 $\lambda_{0.5}^A = 1000\text{h}$, $\lambda_{0.5}^B = 1500\text{h}$, 请说明 A, B 两种轴承中哪个质量更好?

解答：例 0.12

题目: 设 T_A, T_B 表示某厂生产的两种轴承 A, B 的寿命, 已知 $\lambda_{0.5}^A$ =1000h, $\lambda_{0.5}^B$ =1500h, 请说明 A, B 两种轴承中哪个质量更好?

解答:

- 根据上侧分位数的定义可知, $\lambda_{0.5}^A$ =1000h 表示 A 轴承中约有 50% 的寿命超过 1000h, $\lambda_{0.5}^B$ =1500h 表示 B 轴承中约有 50% 的寿命超过 1500h, 从上侧 $\alpha = 0.5$ 分位数上说明了后者的质量比前者更高一点.
- 特别的, 称上侧 $\alpha = 0.5$ 的分位数 (点)$\lambda_{0.5}$ 为中位数, 中位数和均值一样都是随机变量的特征数.

正态分布的分位数

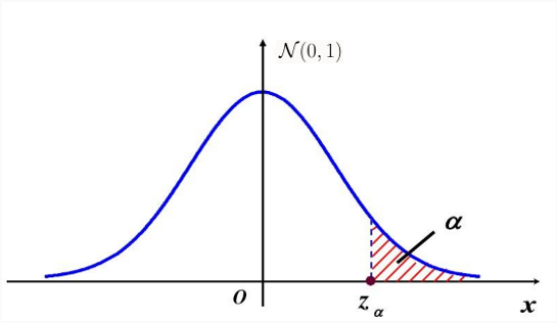
定义 0.16 对正态分布 $X \sim \mathcal{N}(0,1)$, 给定 $\alpha \in (0,1)$, 满足

$$P(X > \mu_\alpha) = \int_{\mu_\alpha}^{\infty} f(x)\,\mathrm{d}x = \alpha$$

的点 μ_α 称为正态分布上侧 α 分位点.

正态分布的分位数的性质:

- 由对称性可知: $\mu_{1-\alpha} = -\mu_\alpha$
- 由正态分布的密度函数可知: $\Phi(\mu_\alpha) = 1 - \alpha$

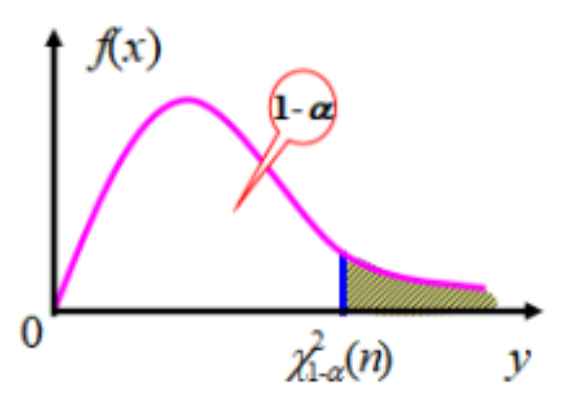


χ^2 分布的分位数

定义 0.17 对 χ^2 分布 $X \sim \chi^2(n)$, 给 $\alpha \in (0,1)$, 满足 $P(X \geq \chi^2_\alpha(n)) = \alpha$ 的点 $\chi^2_\alpha(n)$ 称为 $\chi^2(n)$ 分布上侧 α 分位点.

χ^2 分布的分位数的性质:

- 当 $n \to \infty$ 时有 $\chi^2_\alpha(n) \approx \frac{1}{2}\left(\mu_\alpha + \sqrt{2n-1}\right)^2$, 其中 μ_α 为正态分布上侧 α 分位点.



附录：χ^2 分布的上侧分位数表

附表三　χ^2 分布上侧分位数表　$\left(P\{\chi^2(n) > \chi_\alpha^2(n)\} = \alpha\right)$

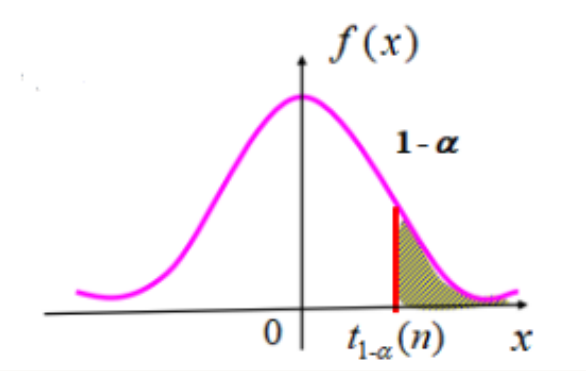
n \ α	0.995	0.99	0.975	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.025	0.01	0.005
1	0.00004	0.00016	0.001	0.004	0.016	0.102	0.455	1.323	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	0.575	1.386	2.773	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	1.213	2.366	4.108	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	1.923	3.357	5.385	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	2.675	4.351	6.626	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	3.455	5.348	7.841	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	5.071	7.344	10.219	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	11.389	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	6.737	9.342	12.549	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	7.584	10.341	13.701	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	8.438	11.340	14.845	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	9.299	12.340	15.984	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	10.165	13.339	17.117	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	11.037	14.339	18.245	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	11.912	15.338	19.369	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	12.792	16.338	20.489	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	13.675	17.338	21.605	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	14.562	18.338	22.718	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	15.452	19.337	23.828	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	16.344	20.337	24.935	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	17.240	21.337	26.039	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	18.137	22.337	27.141	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	19.037	23.337	28.241	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	19.939	24.337	29.339	34.382	37.652	40.646	44.314	46.928

t 分布的分位数

定义 0.18 对 t 分布 $X \sim t(n)$, 给 $\alpha \in (0,1)$, 满足 $P(X \geq t_\alpha(n)) = \alpha$ 的点 $t_\alpha(n)$ 称为 $t(n)$ 分布上侧 α 分位点.

t 分布的分位数的性质:

- 由对称性可知: $t_{1-\alpha} = -t_\alpha$.



附录: t 分布的上侧分位数表

附表四　t 分布上侧分位数表　$\left(P\{t(n)>t_\alpha(n)\}=\alpha\right)$

n \ α	0.20	0.15	0.10	0.05	0.025	0.01	0.005
1	1.376	1.963	3.078	6.314	12.706	31.821	63.656
2	1.061	1.386	1.886	2.92	4.303	6.965	9.925
3	0.978	1.25	1.638	2.353	3.182	4.541	5.841
4	0.941	1.19	1.533	2.132	2.776	3.747	4.604
5	0.92	1.156	1.476	2.015	2.571	3.365	4.032
6	0.906	1.134	1.44	1.943	2.447	3.143	3.707
7	0.896	1.119	1.415	1.895	2.365	2.998	3.499
8	0.889	1.108	1.397	1.86	2.306	2.896	3.355
9	0.883	1.1	1.383	1.833	2.262	2.821	3.25
10	0.879	1.093	1.372	1.812	2.228	2.764	3.169
11	0.876	1.088	1.363	1.796	2.201	2.718	3.106
12	0.873	1.083	1.356	1.782	2.179	2.681	3.055
13	0.87	1.079	1.35	1.771	2.16	2.65	3.012
14	0.868	1.076	1.345	1.761	2.145	2.624	2.977
15	0.866	1.074	1.341	1.753	2.131	2.602	2.947
16	0.865	1.071	1.337	1.746	2.12	2.583	2.921
17	0.863	1.069	1.333	1.74	2.11	2.567	2.898
18	0.862	1.067	1.33	1.734	2.101	2.552	2.878
19	0.861	1.066	1.328	1.729	2.093	2.539	2.861
20	0.86	1.064	1.325	1.725	2.086	2.528	2.845

附表四　t 分布上侧分位数表　$\left(P\{t(n)>t_\alpha(n)\}=\alpha\right)$

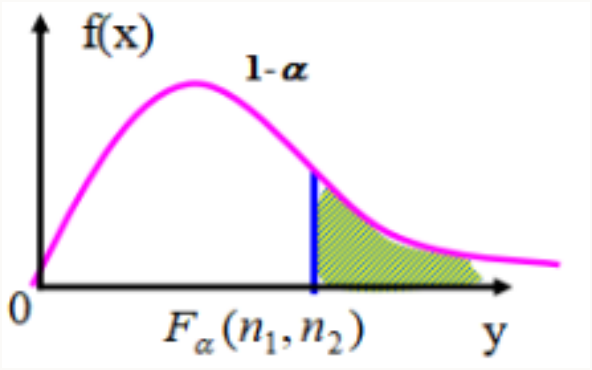
n \ α	0.20	0.15	0.10	0.05	0.025	0.01	0.005
21	0.859	1.063	1.323	1.721	2.08	2.518	2.831
22	0.858	1.061	1.321	1.717	2.074	2.508	2.819
23	0.858	1.06	1.319	1.714	2.069	2.5	2.807
24	0.857	1.059	1.318	1.711	2.064	2.492	2.797
25	0.856	1.058	1.316	1.708	2.06	2.485	2.787
26	0.856	1.058	1.315	1.706	2.056	2.479	2.779
27	0.855	1.057	1.314	1.703	2.052	2.473	2.771
28	0.855	1.056	1.313	1.701	2.048	2.467	2.763
29	0.854	1.055	1.311	1.699	2.045	2.462	2.756
30	0.854	1.055	1.31	1.697	2.042	2.457	2.75
31	0.8535	1.0541	1.3095	1.6955	2.0395	2.453	2.7441
32	0.8531	1.0536	1.3086	1.6939	2.037	2.449	2.7385
33	0.8527	1.0531	1.3078	1.6924	2.0345	2.445	2.7333
34	0.8524	1.0526	1.307	1.6909	2.0323	2.441	2.7284
35	0.8521	1.0521	1.3062	1.6896	2.0301	2.438	2.7239
36	0.8518	1.0516	1.3055	1.6883	2.0281	2.434	2.7195
37	0.8515	1.0512	1.3049	1.6871	2.0262	2.431	2.7155
38	0.8512	1.0508	1.3042	1.686	2.0244	2.428	2.7116
39	0.851	1.0504	1.3037	1.6849	2.0227	2.426	2.7079
40	0.8507	1.0501	1.303	1.684	2.021	2.423	2.704
60	0.8477	1.0455	1.296	1.671	2.000	2.390	2.660
120	0.8446	1.0409	1.289	1.658	1.98	2.358	2.617
∞	0.8416	1.0364	1.282	1.645	1.96	2.326	2.576

F 分布的分位数

定义 0.19 对 F 分布 $X \sim F(m,n)$, 给 $\alpha \in (0,1)$, 满足 $P(X \geq F_\alpha(m,n)) = \alpha$ 的点 $F_\alpha(m,n)$ 称为 $F(m,n)$ 分布上侧 α 分位点.

F 分布的分位数的性质:

- $F_{1-\alpha}(m,n) = 1/F_\alpha(n,m)$.



Proof: F 分布分位数性质

$$X = \frac{U/m}{V/n} \sim F(m,n), \quad Y = \frac{1}{X} \sim F(n,m).$$

$$P(Y \geq F_\alpha(n,m)) = \alpha$$

$$P\left(\tfrac{1}{X} \geq F_\alpha(n,m)\right) = \alpha \Longleftrightarrow P\left(X \leq \tfrac{1}{F_\alpha(n,m)}\right) = \alpha.$$

$$P(X \leq c) = \alpha \Longleftrightarrow P(X \geq c) = 1-\alpha, \quad c = \tfrac{1}{F_\alpha(n,m)}.$$

$$F_{1-\alpha}(m,n) = \frac{1}{F_\alpha(n,m)}.$$

正态分布的抽样分布：例 0.13

例 0.13 设 $X_1, X_2, \ldots, X_{10}$ 是来自总体 $\mathcal{N}(\mu, 1/4)$ 的样本. 问

(1) 若 $\mu = 0$, 求 $P(\sum_{i=1}^{10} X_i^2 \geq 4)$

(2) 若 μ 未知, 求 $P(\sum_{i=1}^{10}(X_i - \bar{X})^2 \geq 3.45)$

解答：例 0.13

题目: 设 $X_1, X_2, \ldots, X_{10}$ 是来自总体 $\mathcal{N}(\mu, 1/4)$ 的样本, i) 若 $\mu = 0$, 求 $P(\sum_{i=1}^{10} X_i^2 \geq 4)$; ii) 若 μ 未知, 求 $P(\sum_{i=1}^{10}(X_i - \bar{X})^2 \geq 3.45)$.

解答:

- 若 $\mu = 0$, 根据 χ^2 分布的定义可知 $\left(\frac{X_1}{1/2}\right)^2 + \left(\frac{X_2}{1/2}\right)^2 + \cdots + \left(\frac{X_{10}}{1/2}\right)^2 \sim \chi^2(10)$. 则有 $P\left(\sum_{i=1}^{10} X_i^2 \geq 4\right) = P\left(\frac{\sum_{i=1}^{10} X_i^2}{1/4} \geq \frac{4}{1/4}\right)$. 通过查询 χ^2 分布的上侧分位数表可知, 自由度为 10 的 χ^2 分布上侧 0.1 分位点恰好约为 16, 即 $P(\sum_{i=1}^{10} X_i^2 \geq 4) \approx 0.1$.
- 根据
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)\,,$$
则有 $\frac{\sum_{i=1}^{10}(X_i - \bar{X})^2}{1/4} \sim \chi^2(9)$. 则有
$$P\left(\sum_{i=1}^{10}(X_i - \bar{X})^2 \geq 3.45\right) = P\left(\frac{\sum_{i=1}^{10}(X_i - \bar{X})^2}{1/4} \geq \frac{3.45}{1/4} \approx 13.8\right)$$
通过查询 χ^2 分布的上侧分位数表可知, 自由度为 9 的 χ^2 分布.

正态分布的抽样分布：例 0.14

例 0.14 设 $X_1, X_2, \ldots, X_{25}$ 是来自总体 $\mathcal{N}(12, \sigma^2)$ 的样本, 求

- i) 若 $\sigma = 2$, 求 $P(\sum_{i=1}^{25} X_i/25 \geq 12.5)$;
- ii) 若 σ 未知, 但知道无偏方差为 $S^2 = 5.57$, 求 $P(\sum_{i=1}^{25} X_i/25 \geq 12.95)$.

解答：例 0.14

题目: 如上所述.

解答:

- 若 $\sigma = 2$, 根据定理0.6可知 $\bar{X} = \frac{1}{25}\sum_{i=1}^{25} X_i \sim \mathcal{N}(12, 4/25)$, 则有

$$P\left(\frac{\sum_{i=1}^{25} X_i}{25} \geq 12.5\right) = P(\bar{X} \geq 12.5) = P\left(\frac{\bar{X} - 12}{2/5} \geq \frac{12.5 - 12}{2/5}\right)$$
$$= 1 - \Phi(1.25) = 0.1056.$$

- 根据定理 0.7可知 $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$, 又根据题目有无偏方差为 $S^2 = 5.57$, 即

$$P\left(\frac{\sum_{i=1}^{25} X_i}{25} \geq 12.95\right) = P\left(\frac{\sum_{i=1}^{25} X_i/25 - 12}{5.57/\sqrt{25}} \geq \frac{12.95 - 12}{5.57/\sqrt{25}} \approx 0.85\right)$$

  通过查询 t 分布的上侧分位数表可知, 自由度为 24 的 t 分布上侧 0.2 分位点恰好约为 0.85, 即 $P(\sum_{i=1}^{25} X_i/25 \geq 12.95) \approx 0.2$.

如何考察抽样分布？

抽样分布中具有极其复杂的公式, 哪一些知识点是需要我们记忆的, 哪一些是比较重要的?

- 利用中心极限定理, 将变量归一化为 $\mathcal{N}(0,1)$
  - $Y \sim \chi^2(n)$, which operation of Y obeys $\mathcal{N}(0,1)$.
  - 设总体分布 $U(0,1)$, $x_1, x_2, \ldots, x_n$ 为样本, 试求 $x_{(k)}$.
- 本节课所提及的六种分布 (包含三大抽样分布), 其来源是怎样的?
  - $X_1, X_2 \sim \mathcal{N}(0,1)$, which distribution does $X_1^2 + X_2^2$ obey?
  - $X_1, X_2 \sim \mathcal{N}(0,1)$, if $aX_1^2 + bX_2^2$ obeys χ^2, find a, b.
  - If $T \sim t(n)$, compute $Y = T^2$.
- 六种分布 (包含三大抽样分布) 的数字特征 (期望、方差)、基本性质
- 查表 (三大抽样分布、正态分布) 判断分位数及概率
  - 设 $X_1, X_2, \ldots, X_{10}$ 是来自总体 $\mathcal{N}(0, 1/4)$ 的样本. 求 $P(\sum_{i=1}^{10} X_i^2 \geq 4)$.