第五章 数理统计的基本知识

刘 春 光

暨南大学数学系

2018年4月

目录

目录

总体、个体与样本

数理统计中，称研究问题所涉及对象的全体为总体，总体中的每个成员为个体。从总体中抽出的若干个体称为样本。

例如：研究某工厂生产的某种产品的废品率，则这种产品的全体就是总体，而每件产品都是一个个体。为考察产品情况，从全部产品中抽取一些样品，这些样品就是样本。

总体、个体与样本

实际处理中，我们真正关心的并不一定是总体或个体本身，而真正关心的是总体或个体的某项数量指标。故也将总体理解为那些研究对象的某项数量指标的全体。

例：研究某地区所有人的年收入（设总人数为N），则总体可以认为是所有人年收入的N个数值。

总体、个体与样本

例：用一把尺子测量一件物体的长度。假定n次测量值分别为$X_1, X_2, \cdots, X_n$。

该问题中，我们把测量值$X_1, X_2, \cdots, X_n$看成样本；总体就是一切所有可能的测量值的全体。

总体分布

对一个总体，如果用X表示其数量指标，则我们随机地抽取个体时，X就构成总体上的一个随机变量。

X的分布称为总体分布。总体的特性是由总体分布来刻画的。因此，常把总体和总体分布视为同义语。

总体分布

如果总体包含的个体数量是有限的，则称该总体为有限总体。否则称该总体为无限总体。

有限总体的分布是离散型的，且分布通常与总体所含个体数量有关系，研究起来比较困难。

故总体所含的个体数量很大时，一般近似视之为无限总体。

样本的二重性

假设$X_1, X_2, \cdots, X_n$是从总体X中取出的样本，

1. 在对这些样本进行观测之前，$X_1, \cdots, X_n$是相互独立的随机变量，均服从总体分布；
2. 一旦对样本进行观测，$X_1, \cdots, X_n$即为确定的一组数值。

从而样本兼有随机变量和确定数值两种属性。

有时为了区分，也将$X_1, X_2, \cdots, X_n$的观测值记为$x_1, x_2, \cdots, x_n$。

样本精确定义

Definition (样本(Sample))

称随机变量$X_1, X_2, \cdots, X_n$构成一个（简单）随机样本，如果这些随机变量

1. 相互独立(independent)；
2. 服从相同的分布(identically distributed)。

它们共同服从的分布称为总体分布；样本个数n称为样本容量。

样本分布函数

Definition (样本分布函数)

设总体X的n个样本观测值$x_1, x_2, \cdots, x_n$的所有取值为$x_{(1)} < x_{(2)} < \cdots < x_{(l)}$，且这些取值的频数分别为$m_1, m_2, \cdots, m_l$，则函数

$$F_n(x) = \sum_{x_{(k)} \leq x} \frac{m_k}{n}$$

称为样本分布函数。

样本分布函数

样本分布函数相当于将样本的频率分布表

观测值	$x_{(1)}$	$x_{(2)}$	$\cdots$	$x_{(l)}$
频率	$\frac{m_1}{n}$	$\frac{m_2}{n}$	$\cdots$	$\frac{m_l}{n}$

视作概率分布表时对应的概率分布函数。

样本分布函数

对任意确定的x，$F_n(x)$是事件$X \leq x$在抽样中的频率，根据伯努利定理可知：
当$n \to \infty$时，$F_n(x)$依概率收敛于总体分布函数$F(x)$，即对任意给定的正数ε，有

$$\lim_{n\to\infty} P\{|F_n(x) - F(x)| < \varepsilon\} = 1.$$

目录

统计量

在实际问题中，总体分布一般是未知的，我们常常事先假定总体分布的类型，再通过取样的方式确定分布中的未知参数。此时这些未知参数常常写成样本的函数。

Definition (统计量(Statistic))

样本的已知函数（不含问题中的未知参数）称为统计量。

统计量

例：研究某城市居民的收入情况，事先假定该城市居民的年收入X服从正态分布$N(\mu,\sigma^2)$，其中μ与σ^2都是未知参数。

在抽取样本$X_1, X_2, \cdots, X_n$的情况下，一般用样本平均值

$$\frac{X_1 + X_2 + \cdots + X_n}{n}$$

近似估计μ，该平均值就是一个统计量。

统计量

作为对比，以下函数含有问题中的未知参数，因此不是统计量

$$\frac{X_1+X_2+\cdots+X_n}{n\sigma},$$

$$\frac{X_1+X_2+\cdots+X_n}{n}-\mu.$$

常用统计量

Definition (样本均值(Sample mean))

对样本$X_1, X_2, \cdots, X_n$，称

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

为样本均值。

常用统计量

Definition (Sample variance, sample standard deviation)

对样本$X_1, X_2, \cdots, X_n$，称

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

为<u>样本方差</u>；称

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

为<u>样本标准差</u>。

常用统计量

样本方差的性质：

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right).$$

常用统计量

Definition (Sample raw moment, sample central moment)

对样本$X_1, X_2, \cdots, X_n$及正整数k，称

$$A_k := \frac{1}{n}\sum_{i=1}^{n} X_i^k = \frac{X_1^k + X_2^k + \cdots + X_n^k}{n}$$

为<u>k阶样本原点矩</u>；对$k \geq 2$，称

$$M_k := \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^k$$

为<u>k阶样本中心矩</u>。

补充：其它常用统计量

Example

其它常用的统计量还有：

1. 样本最大值：$\max\{X_1, X_2, \cdots, X_n\}$；
2. 样本第k大值：$X_{(n-k+1)}$；
3. 样本最小值：$\min\{X_1, X_2, \cdots, X_n\}$；
4. 样本第k小值：$X_{(k)}$，

其中$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$为样本的顺序重排。

目录

统计学的三大分布

以下三个来自正态分布的抽样分布

χ^2分布，t分布，F分布

称为统计学的三大分布。

χ^2分布

定义：具有以下概率密度的分布称为具有n个自由度的χ^2-分布（简记为$\chi^2(n)$）：

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

一般用记号χ_n^2表示服从具有n个自由度的χ^2-分布的随机变量。

χ^2分布

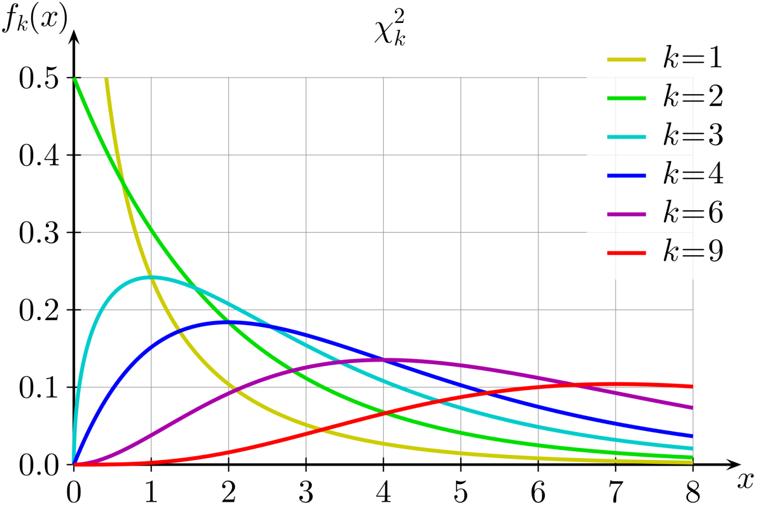

Figure: χ^2分布的密度函数

χ^2分布

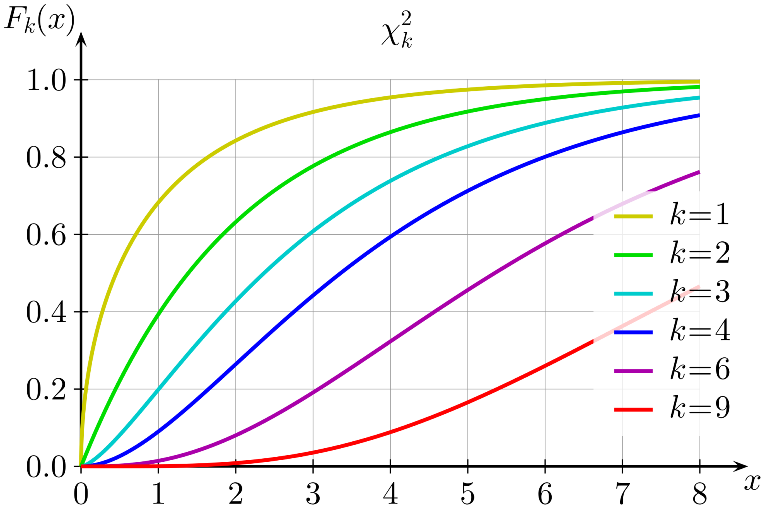

Figure: χ^2分布的分布函数

χ^2分布

χ^2分布的性质：

1. 若X服从标准正态分布，则X^2 服从1个自由度的χ^2分布，即
$$X^2 \sim \chi^2(1).$$
2. 可加性：设$Y_1 \sim \chi^2(m)$, $Y_2 \sim \chi^2(n)$，且两者相互独立，则
$$Y_1 + Y_2 \sim \chi^2(m+n).$$

χ^2分布

定理：设$X_1, X_2, \cdots, X_n$相互独立，都服从标准正态分布，则

$$\chi^2 := \sum_{i=1}^{n} X_i^2 = X_1^2 + X_2^2 + \cdots + X_n^2$$

服从n个自由度的χ^2分布，即

$$\chi^2 \sim \chi^2(n).$$

χ^2分布

χ^2分布的数字特征：

$$E(\chi_n^2) = n, \quad D(\chi_n^2) = 2n.$$

χ^2分布

对给定的$\alpha \in (0,1)$，称满足条件

$$P\{\chi_n^2 \geq \chi_n^2(\alpha)\} = \alpha$$

的点$\chi_n^2(\alpha)$为$\chi^2(n)$分布的上α分位点。

t分布

定义：具有以下概率密度的分布称为具有n个自由度的t分布（简记为$t(n)$）：

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

性质：t分布的概率密度为偶函数。

t分布

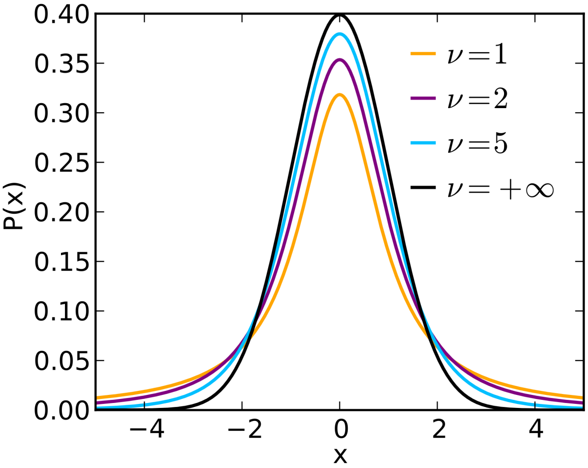

Figure: t分布的密度函数

t分布

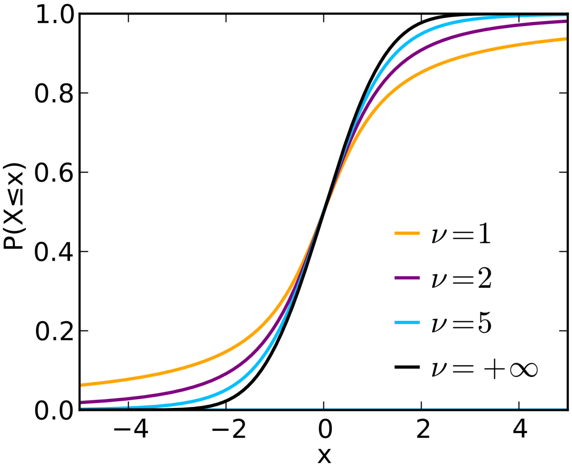

Figure: t分布的分布函数

t分布

定理：设两个随机变量ξ,η相互独立，并且

$$\xi \sim N(0,1), \qquad \eta \sim \chi^2(n).$$

则

$$T := \frac{\xi}{\sqrt{\eta/n}}$$

服从具有n个自由度的t分布。

t分布

设$T \sim t(n)$。对给定的$\alpha \in (0,1)$，称满足条件

$$P\{T > t_n(\alpha)\} = \alpha$$

的点$t_n(\alpha)$为$t(n)$分布的上α分位点。

性质：

$$t_n(1-\alpha) = -t_n(\alpha).$$

F分布

定义：具有以下概率密度的分布称为第一个自由度（分子自由度）为m、第二个自由度（分母自由度）为n的F分布（简记为$F(m,n)$）：

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

F分布

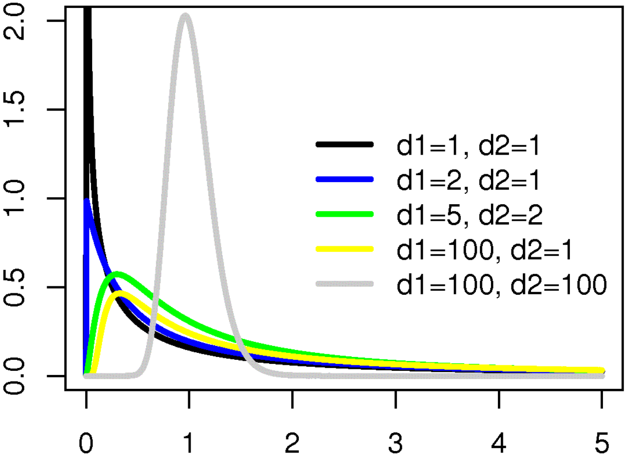

Figure: F分布的密度函数

F分布

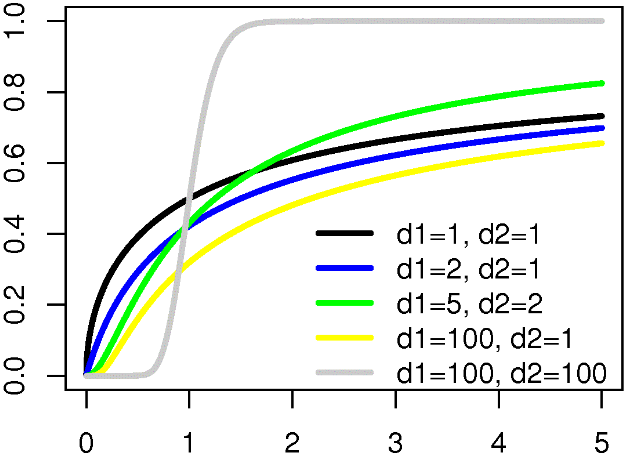

Figure: F分布的分布函数

F分布

定理：设两个随机变量ξ_1,ξ_2相互独立，并且

$$\xi_i \sim \chi^2(n_i), \qquad i = 1, 2.$$

则

$$F := \frac{\xi_1/n_1}{\xi_2/n_2} \sim F(n_1, n_2).$$

F分布

设$F \sim F(m,n)$。对给定的$\alpha \in (0,1)$，称满足条件

$$P\{F > F_{m,n}(\alpha)\} = \alpha$$

的点$F_{m,n}(\alpha)$为$F(m,n)$分布的上α分位点。

F分布

F分布的性质：

1. 若$F \sim F(m,n)$，则
$$\frac{1}{F} \sim F(n,m).$$
2. $F_{m,n}(1-\alpha) = \dfrac{1}{F_{n,m}(\alpha)}.$
3. 若$X \sim t(n)$，则
$$X^2 \sim F(1,n).$$

目录

单个正态总体的统计量的分布

Theorem (样本均值的分布)

设$X_1, X_2, \cdots, X_n$是取自正态总体$N(\mu, \sigma^2)$ 的样本。则

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}),$$

从而有

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

单个正态总体的统计量的分布

Theorem

设$X_1, X_2, \cdots, X_n$是取自正态总体$N(\mu, \sigma^2)$的样本。则

$$\chi^2 := \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \sim \chi^2(n).$$

单个正态总体的统计量的分布

练习：设X_1, X_2, X_3为来自正态总体$N(0, \sigma^2)$的简单随机样本，则统计量$S = \dfrac{X_1 - X_2}{\sqrt{2}|X_3|}$服从的分布为

(A) $F(1, 1)$ (B) $F(2, 1)$ (C) $t(1)$ (D) $t(2)$

单个正态总体的统计量的分布

Theorem

设$X_1, X_2, \cdots, X_n$是取自正态总体$N(\mu, \sigma^2)$的样本。则$\overline{X}$与S^2相互独立，且有

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

证明思路*：令$Y_i = \dfrac{X_i - \mu}{\sigma}$, $i = 1, 2, \cdots, n$, 则$Y_1, Y_2, \cdots, Y_n$相互独立，且都服从标准正态分布。进一步还有

单个正态总体的统计量的分布

证明思路*（续1）：

1. $\overline{X} = \mu + \sigma \overline{Y}$;
2. $\dfrac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2$。

令随机向量

$$Y = (Y_1, Y_2, \cdots, Y_n)^T,$$

则$Y \sim N(0, I_n)$，其中I_n为n阶单位矩阵。

单个正态总体的统计量的分布

证明思路*（续2）：任取正交矩阵C，使得其第一行为

$$\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \cdots, \frac{1}{\sqrt{n}}\right),$$

并令$Z = CY$，其中

$$Z = (Z_1, Z_2, \cdots, Z_n)^T.$$

则

$$Z \sim N(0, I_n).$$

单个正态总体的统计量的分布

证明思路*（续3）：由此得$Z_1, Z_2, \cdots, Z_n$相互独立，且都服从标准正态分布。故Z_1与$\sum_{j=2}^{n} Z_j^2$相互独立，且$\sum_{j=2}^{n} Z_j^2 \sim \chi^2(n-1)$。注意到

1. $\overline{X} = \mu + \sigma \overline{Y} = \mu + \dfrac{1}{\sqrt{n}} \sigma Z_1$；
2. $\dfrac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 = \sum_{j=2}^{n} Z_j^2$，

得定理结论成立。

单个正态总体的统计量的分布

Theorem

设$X_1, X_2, \cdots, X_n$是取自正态总体$N(\mu, \sigma^2)$ 的样本。则

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

单个正态总体的统计量的分布

例1 设总体X服从正态分布$N(\mu,\sigma^2)$，从总体中抽取容量为9的样本，求样本均值$\overline{X}$与总体均值μ之差的绝对值小于2的概率，如果

1. 已知总体方差$\sigma^2 = 16$;
2. 总体方差未知，但知样本方差$S^2 = 18.45$。

单个正态总体的统计量的分布

例2 设总体X服从正态分布$N(\mu, 2^2)$，从总体中抽取容量为16的样本$X_1, X_2, \cdots, X_{16}$。

1. 已知总体均值$\mu = 0$，求$\sum_{i=1}^{16} X_i^2 < 128$的概率；
2. 总体均值未知，求$\sum_{i=1}^{16} (X_i - \overline{X})^2 < 100$的概率。

两个正态总体的统计量的分布

从总体X中抽取样本$X_1, X_2, \cdots, X_{n_1}$，从总体$Y$中抽取样本$Y_1, Y_2, \cdots, Y_{n_2}$，分别记这两组样本的均值为

$$\overline{X} := \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \quad \overline{Y} := \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j;$$

样本方差为

$$S_1^2 := \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\overline{X})^2, \quad S_2^2 := \frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_j-\overline{Y})^2.$$

两个正态总体的统计量的分布

Theorem

设总体$X \sim N(\mu_1, \sigma_1^2)$，总体$Y \sim N(\mu_2, \sigma_2^2)$。则

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

Corollary

设总体$X \sim N(\mu_1, \sigma^2)$，总体$Y \sim N(\mu_2, \sigma^2)$。则

$$U := \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1).$$

两个正态总体的统计量的分布

Theorem

设总体$X \sim N(\mu_1, \sigma^2)$，总体$Y \sim N(\mu_2, \sigma^2)$。则

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

其中

$$S = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}.$$

两个正态总体的统计量的分布

Theorem

设总体$X \sim N(\mu_1, \sigma_1^2)$，总体$Y \sim N(\mu_2, \sigma_2^2)$。则

$$F := \frac{\sum_{i=1}^{n_1}(X_i - \mu_1)^2/(n_1\sigma_1^2)}{\sum_{j=1}^{n_2}(Y_j - \mu_2)^2/(n_2\sigma_2^2)} \sim F(n_1, n_2).$$

Theorem

设总体$X \sim N(\mu_1, \sigma_1^2)$，总体$Y \sim N(\mu_2, \sigma_2^2)$。则

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

两个正态总体的统计量的分布

例3 设总体$X \sim N(20, 5^2)$，总体$Y \sim N(10, 2^2)$，从总体X和Y中分别抽取容量为$n_1 = 10$与$n_2 = 8$的样本，求：

1. 样本均值$\overline{X} - \overline{Y}$的差大于6的概率；
2. 样本方差比$\dfrac{S_1^2}{S_2^2}$小于23的概率。