第一节　随机样本

一、总体与个体

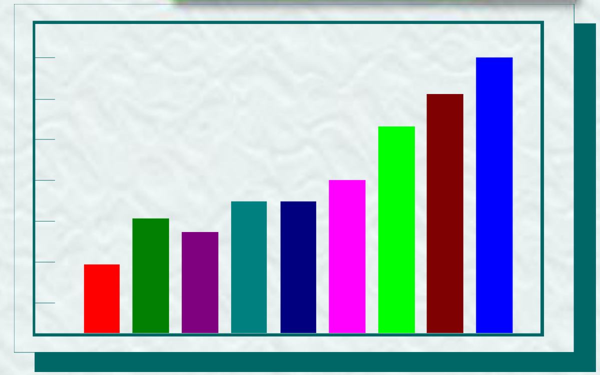
二、随机样本的定义

三、小结

引言

本章转入课程的第二部分

数理统计



数理统计学的内容:

(1)如何收集、整理数据资料;

(2)如何对所得到的数据资料进行　分析、研究, 从而对所研究的对象的性质、特点作出推断.

即:统计推断

例如：生产厂家声称他们生产的灯泡平均寿命不低于6000小时，如何验证厂家说法的真伪？由于灯泡寿命试验是破坏性试验，不可能把整批灯泡逐一检测，只能抽取一部分灯泡进行检验，通过这部分灯泡的寿命数据来推断整批灯泡的平均寿命。以部分数据信息来推断整体未知参数，就是数理统计研究问题的基本方式。

一、总体与个体

1. 总体

试验的全部可能观察值称为总体.

2. 个体

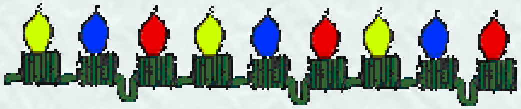
总体中的每个可能观察值称为个体.

实例1 在研究2000名学生的年龄时, 这些学生的年龄的全体就构成一个总体, 每个学生的年龄就是个体.

3. 有限总体和无限总体

实例2 某工厂10月份生产的灯泡寿命所组成的总体中, 个体的总数就是10月份生产的灯泡数, 这是一个有限总体; 而该工厂生产的所有灯泡寿命所组成的总体是一个无限总体, 它包括以往生产和今后生产的灯泡寿命.

当有限总体包含的个体的总数很大时, 可近似地将它看成是无限总体.



- 例：1）了解某校大学生“做过家教（包括正在做家教）”的比例。

  总体是该校大学生全体。这是一个有限总体，每个大学生有许多指标，比如性别，年龄，身高，体重，高考成绩…。现在我们关心的是学生是否“做过家教”这一指标。

- 2）了解某城市的空气质量情况，关注该城市的PM2.5值。总体是城市上空一定范围内的空气，这是一个无限总体，描述空气质量有许多指标，而我们仅关心PM2.5值。
- 3）药厂研究某种药物在人体中的吸收情况。总体是全体国民，这是一个有限总体，但数量非常巨大，我们常把它看成无限总体。

为了采用数理统计方法进行分析，首先要收集数据，数据收集方法一般有两种。

（1）通过调查、记录收集数据。如为了调查大学生是否“做过家教”，可以进行问卷调查；要了解PM2.5值，需要在城市设立若干监测站点，定时收集PM2.5数据。

（2）通过实验收集数据。如为了了解药物吸收情况，首先要进行试验设计，并征集若干志愿者，按试验设计方案将他们分成若干组，监测他们服药后不同时间点身体中药物含量，记录相应的数据。

关于数据的收集（调查数据和实验数据）可以根据数据本身的特点有多种不同的方法和设计，有专门的课程讲授，本课程不作详细介绍。

4. 总体分布

实例3 在2000名大学一年级学生的年龄中, 年龄指标值为“15”，“16”，“17”，“18”，“19”，“20”的依次有9，21，132，1207，588，43 名, 它们在总体中所占比率依次为

$$\frac{9}{2000},\quad \frac{21}{2000},\quad \frac{132}{2000},\quad \frac{1207}{2000},\quad \frac{588}{2000},\quad \frac{43}{2000},$$

即学生年龄的取值有一定的分布.

一般地, 我们所研究的总体, 即研究对象的某项数量指标 X , 其取值在客观上有一定的分布, X 是一个随机变量.

鉴于此，常用随机变量的记号或用其分布函数表示总体. 如说总体X或总体$F(x)$.

如**实例**3中, 总体就是数集 {15, 16, 17, 18, 19, 20}.

总体分布为

年龄	15	16	17	18	19	20
比率	$\frac{9}{2000}$	$\frac{21}{2000}$	$\frac{132}{2000}$	$\frac{1207}{2000}$	$\frac{588}{2000}$	$\frac{43}{2000}$

二、随机样本的定义

1. 样本的定义

设 X 是具有分布函数 F 的随机变量, 若 X_1, $X_2, \cdots, X_n$ 是具有同一分布函数 F、相互独立的随机变量, 则称 $X_1, X_2, \cdots, X_n$ 为从分布函数 F (或总体 F、或总体 X) 得到的容量为 n 的简单随机样本, 简称样本.

它们的观察值 $x_1, x_2, \cdots, x_n$ 称为样本值, 又称为 X 的 n 个独立的观察值.

2. 简单随机抽样的定义

获得简单随机样本的抽样方法称为简单随机抽样.

根据定义得：若 $X_1, X_2, \cdots, X_n$ 为 F 的一个样本,

则 $X_1, X_2, \cdots, X_n$ 的联合分布函数为

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

又若 X 具有概率密度 f,

则 $X_1, X_2, \cdots, X_n$ 的联合概率密度为

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

例4 设总体 X 服从参数为 $1/\lambda\ (\lambda > 0)$ 的指数分布, $(X_1, X_2, \cdots, X_n)$ 是来自总体的样本, 求样本 $(X_1, X_2, \cdots, X_n)$ 的概率密度.

解 总体 X 的概率密度为 $f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & x \le 0, \end{cases}$

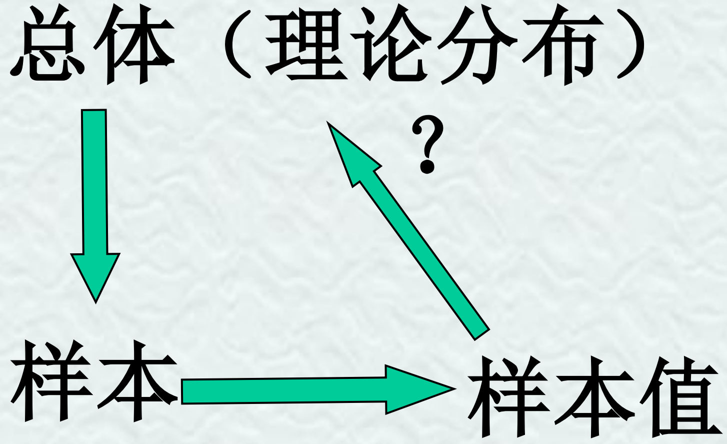
因为 $X_1, X_2, \cdots, X_n$ 相互独立, 且与 X 有相同的分布,

所以 $(X_1, X_2, \cdots, X_n)$ 的概率密度为

$$f_n(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i) = \begin{cases} \lambda^n e^{-\lambda \sum\limits_{i=1}^{n} x_i}, & x_i > 0, \\ 0, & \text{其他}. \end{cases}$$

总体、样本、样本值的关系

事实上我们抽样后得到的资料都是具体的、确定的值. 如我们从某班大学生中抽取10人测量身高，得到10个数，它们是样本取到的值而不是样本. 我们只能观察到随机变量取的值而见不到随机变量.



统计是从手中已有的资料--样本值，去推断总体的情况---总体分布$F(x)$的性质.

样本是联系二者的桥梁

总体分布决定了样本取值的概率规律，也就是样本取到样本值的规律，因而可以由样本值去推断总体.

三、小结

基本概念：个体　总体 $\begin{cases}\text{有限总体}\\\text{无限总体}\end{cases}$　随机样本

说明1　一个总体对应一个随机变量X, 我们将不区分总体和相应的随机变量, 统称为总体X.

说明2　在实际中遇到的总体往往是有限总体, 它对应一个离散型随机变量; 当总体中包含的个体的个数很大时, 在理论上可认为它是一个无限总体.

第二节　　抽样分布

一、基本概念

二、常见分布

三、小结

一、基本概念

1. 统计量的定义

设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本, $g(X_1, X_2, \cdots, X_n)$ 是 $X_1, X_2, \cdots, X_n$ 的函数, 若 g 中不含未知参数, 则称 $g(X_1, X_2, \cdots, X_n)$ 是一个统计量.

设 $x_1, x_2, \cdots, x_n$ 是相应于样本 $X_1, X_2, \cdots, X_n$ 的样本值, 则称 $g(x_1, x_2, \cdots, x_n)$ 是 $g(X_1, X_2, \cdots, X_n)$ 的观察值.

实例1 设 X_1, X_2, X_3 是来自总体 $N(\mu, \sigma^2)$ 的一个样本, 其中 μ 为已知, σ^2 为未知, 判断下列各式哪些是统计量, 哪些不是?

$T_1 = X_1,$ $\quad T_2 = X_1 + X_2 e^{X_3},$

$T_3 = \frac{1}{3}(X_1 + X_2 + X_3),$

$T_4 = \max(X_1, X_2, X_3),$ $\quad T_5 = X_1 + X_2 - 2\mu,$

是

$T_6 = \frac{1}{\sigma^2}(X_1^2 + X_2^2 + X_3^2).$

不是

2. 几个常用统计量的定义

设 $X_1, X_2, \cdots, X_n$ 是来自总体的一个样本，

$x_1, x_2, \cdots, x_n$ 是这一样本的观察值.

(1)样本平均值 $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i;$$

其观察值 $$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

(2)样本方差

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

其观察值

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right).$$

(3)样本标准差

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

其观察值 $$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

(4) 样本 k 阶(原点)矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, k = 1, 2, \cdots;$

其观察值 $\alpha_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, k = 1, 2, \cdots.$

(5)样本 k 阶中心矩

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k, k = 2, 3, \cdots;$$

其观察值 $b_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k, k = 2, 3, \cdots.$

由以上定义得下述结论:

若总体 X 的 k 阶矩 $E(X^k)$ 记成 μ_k 存在,

则当 $n \to \infty$ 时, $A_k \xrightarrow{P} \mu_k, k = 1, 2, \cdots.$

证明　因为 $X_1, X_2, \cdots, X_n$ 独立且与 X 同分布,

所以 $X_1^k, X_2^k, \cdots, X_n^k$ 独立且与 X^k 同分布,

故有　$E(X_1^k) = E(X_2^k) = \cdots = E(X_n^k) = \mu_k.$

再根据第五章辛钦定理知

辛钦定理

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} \mu_k\ , \quad k = 1, 2, \cdots;$$

由第五章关于依概率收敛的序列的性质知

$$\boxed{g(A_1, A_2, \cdots, A_k) \xrightarrow{P} g(\mu_1, \mu_2, \cdots, \mu_k),}$$

其中 g 是连续函数.

以上结论是下一章所要介绍的矩估计法的理论根据.

3. 经验分布函数

总体分布函数 $F(x)$ 相应的统计量称为经验分布函数.

经验分布函数的做法如下:

设 $X_1, X_2, \cdots, X_n$ 是总体 F 的一个样本,

用 $S(x)\ (-\infty < x < +\infty)$ 表示 $X_1, X_2, \cdots, X_n$ 中不大于 x 的随机变量的个数,

定义经验分布函数 $F_n(x)$ 为

$$F_n(x) = \frac{1}{n} S(x), \quad (-\infty < x < +\infty)$$

对于一个样本值, $F_n(x)$ 的观察值容易求得.

($F_n(x)$ 的观察值仍以 $F_n(x)$ 表示.)

实例2 设总体 F 具有一个样本值 1, 2, 3,

则经验分布函数 $F_3(x)$ 的观察值为

$$F_3(x)=\begin{cases}0, & x<1,\\ \dfrac{1}{3}, & 1\le x<2,\\ \dfrac{2}{3}, & 2\le x<3,\\ 1, & x\ge 3.\end{cases}$$

实例3 设总体 F 具有一个样本值 1, 1, 2, 则经验分布函数 $F_3(x)$ 的观察值为

$$F_3(x)=\begin{cases}0, & x<1,\\ \dfrac{2}{3}, & 1\le x<2,\\ 1, & x\ge 2.\end{cases}$$

一般地,

设 $x_1, x_2, \cdots, x_n$ 是总体 F 的一个容量为 n 样本值,

先将 $x_1, x_2, \cdots, x_n$ 按自小到大的次序排列,

并重新编号, $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$,

则经验分布函数 $F_n(x)$ 的观察值为

$$F_n(x) = \begin{cases} 0, & x < x_{(1)}, \\ \dfrac{k}{n}, & x_{(k)} \le x < x_{(k+1)}, \\ 1, & x \ge x_{(n)}. \end{cases}$$

格里汶科定理

格里汶科

对于任一实数 x, 当 $n \to \infty$ 时, $F_n(x)$ 以概率1一致收敛于分布函数 $F(x)$, 即

$$P\left\{\lim_{n\to\infty} \sup_{-\infty<x<+\infty} \left|F_n(x) - F(x)\right| = 0\right\} = 1.$$

对于任一实数 x 当 n 充分大时, 经验分布函数的任一个观察值 $F_n(x)$ 与总体分布函数 $F(x)$ 只有微小的差别, 从而在实际上可当作 $F(x)$ 来使用.

二、常见分布

统计量的分布称为抽样分布.

1. χ^2分布

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $N(0,1)$ 的样本,

则称统计量 $\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$ 服从自由度为 n 的 χ^2 分布, 记为 $\chi^2 \sim \chi^2(n)$.

自由度:

指 $\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$ 中右端包含独立变量的个数.

$\chi^2(n)$分布的概率密度为

$$f(y)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}y^{\frac{n}{2}-1}\mathrm{e}^{-\frac{y}{2}}, & y>0\\ 0 & \text{其他.}\end{cases}$$

其中伽玛函数 $\Gamma(x)$通过积分

$$\Gamma(x)=\int_0^{\infty}e^{-t}t^{x-1}dt,\quad x>0$$

来定义.

χ^2 分布的性质

性质1 (χ^2 分布的可加性)

设 $\chi_1^2 \sim \chi^2(n_1)$, $\chi_2^2 \sim \chi^2(n_2)$, 并且 χ_1^2, χ_2^2 独立, 则 $\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2)$.

(此性质可以推广到多个随机变量的情形.)

设 $\chi_i^2 \sim \chi^2(n_i)$, 并且 χ_i^2 $(i = 1, 2, \cdots, m)$ 相互独立, 则 $\sum_{i=1}^{m} \chi_i^2 \sim \chi^2(n_1 + n_2 + \cdots + n_m)$.

性质2

(χ^2分布的数学期望和方差)

若 $\chi^2 \sim \chi^2(n)$, 则 $E(\chi^2) = n, \quad D(\chi^2) = 2n.$

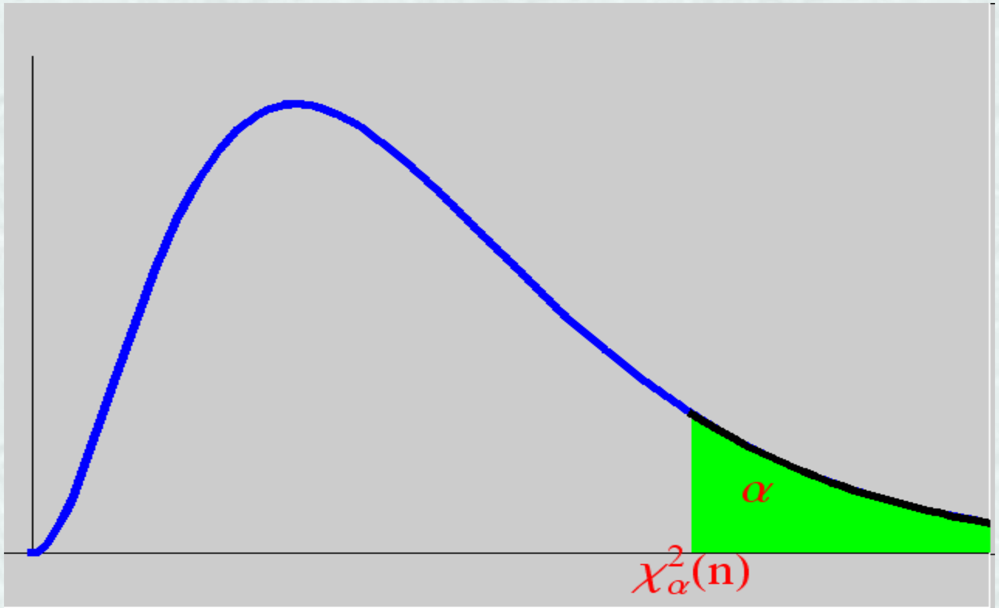
χ^2 分布的分位点

对于给定的正数α, $0<\alpha<1$, 称满足条件

$$P\{\chi^2>\chi_\alpha^2(n)\}=\int_{\chi_\alpha^2(n)}^{\infty}f(y)\mathrm{d}y=\alpha$$

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

对于不同的α, n,
可以通过查表求
得上α 分位点的值.



例1 设 $Z \sim \chi^2(n)$, $\chi^2(n)$ 的上 α 分位点满足

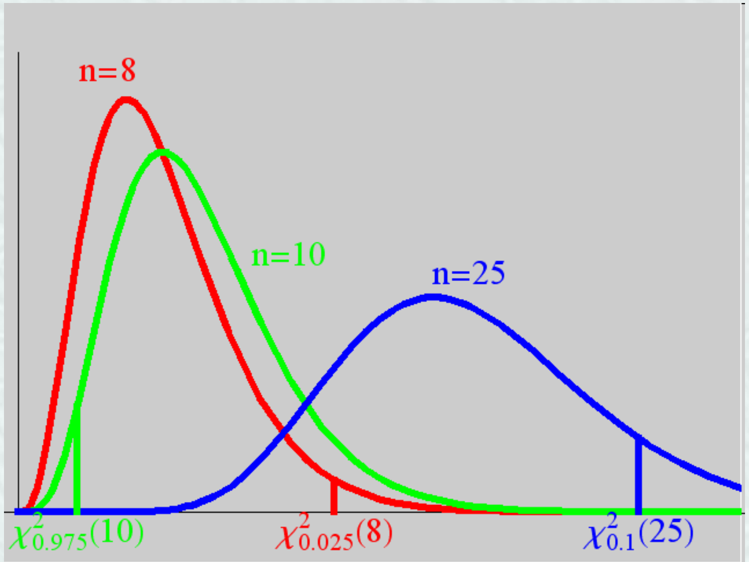
$$P\{Z > \chi_\alpha^2(n)\} = \int_{\chi_\alpha^2(n)}^{+\infty} \chi^2(y;n)\mathrm{d}y = \alpha,$$

求 $\chi_\alpha^2(n)$ 的值, 可通过查表完成.

$\chi_{0.025}^2(8) = 17.534,$ 附表4-1

$\chi_{0.975}^2(10) = 3.247,$ 附表4-2

$\chi_{0.1}^2(25) = 34.381.$ 附表4-3



附表5只详列到 n=40 为止.

费舍尔(R.A.Fisher)证明:

当 n 充分大时, $\chi_{\alpha}^{2}(n) \approx \frac{1}{2}(z_{\alpha}+\sqrt{2n-1})^{2}$.

其中 z_{α} 是标准正态分布的上 α 分位点.

利用上面公式,

可以求得 $n>40$ 时, 上 α 分位点的近似值.

例如 $\chi_{0.05}^{2}(50) \approx \frac{1}{2}(1.645+\sqrt{99})^{2}=67.221.$

而查详表可得 $\chi_{0.05}^{2}(50)=67.505$.

例：设$X_i\ (i=1,2....15)$是总体X~N(0,0.04)的一个样本，求$P\{\sum_{i=1}^{15}X_i^2>1.223\}$

解：因为X~N(0,0.04)，$X_i\sim N(0,0.04),(i=1,2....15)$

所以，$\frac{X_i}{0.2}=Y_i\sim N(0,1),(i=1,2....15)$ $\sum_{i=1}^{15}Y_i^2\sim\chi^2(15)$

于是，$P\{\sum_{i=1}^{15}X_i^2>1.223\}$

$$=P\{\frac{\sum_{i=1}^{15}X_i^2}{0.2^2}>\frac{1.223}{0.2^2}\}=P\{\sum_{i=1}^{15}Y_i^2>30.57\}$$

由于 $\sum_{i=1}^{15} Y_i^2 \sim \chi^2(15)$ 查表得， $\chi^2_{0.01}(15) = 30.57$

所以 $P\{\sum_{i=1}^{15} X_i^2 > 1.223\}$

$$= P\{\sum_{i=1}^{15} Y_i^2 > 30.57\}$$

$$=0.01$$

2. t 分布

设 $X \sim N(0,1)$, $Y \sim \chi^2(n)$, 且 X, Y 独立,

则称随机变量 $t = \dfrac{X}{\sqrt{Y/n}}$ 服从自由度为 n 的 t

分布, 记为 $t \sim t(n)$.

t 分布又称学生氏(Student)分布.

$t(n)$ 分布的概率密度函数为

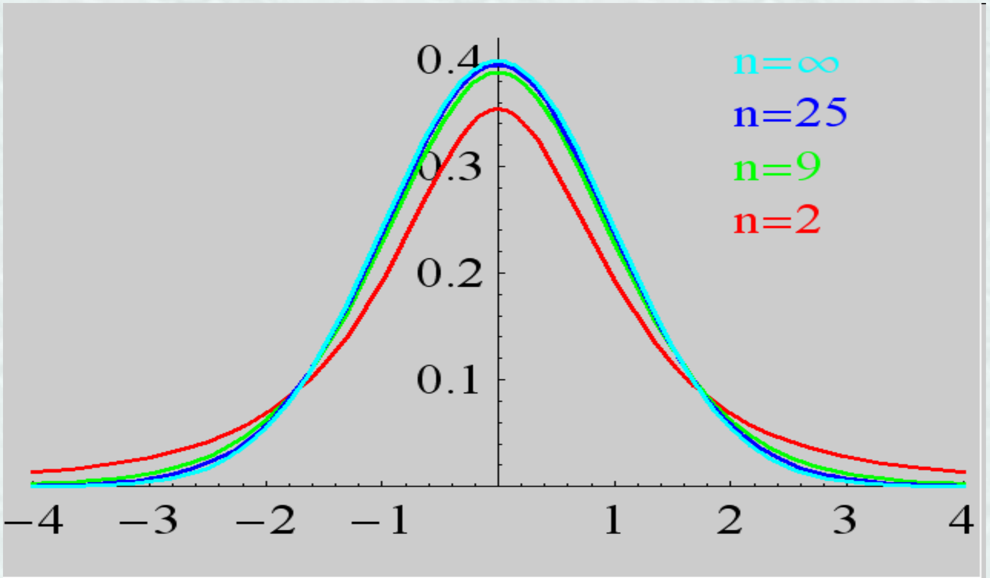
$$h(t) = \frac{\Gamma\left(\dfrac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\dfrac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < t < +\infty$$

t 分布的概率密度曲线如图

显然图形是关于

$t = 0$对称的.

当 n 充分大时, 其图形类似于标准正态变量概率密度的图形.



因为$\lim\limits_{n\to\infty} h(t) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$,

所以当 n 足够大时 t 分布近似于 $N(0,1)$ 分布,

但对于较小的n, t分布与$N(0,1)$分布相差很大.

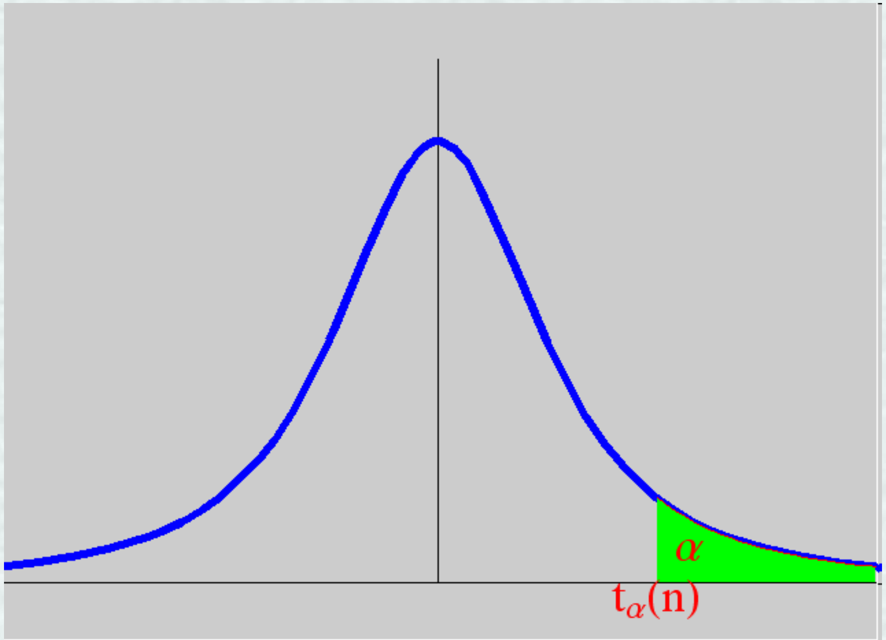
t 分布的分位点

对于给定的 α，$0<\alpha<1$，称满足条件

$$P\{t>t_{\alpha}(n)\}=\int_{t_{\alpha}(n)}^{\infty}h(t)\mathrm{d}t=\alpha$$

的点 $t_{\alpha}(n)$ 为 $t(n)$ 分布的上 α 分位点.

可以通过查表求得上α分位点的值.



由分布的对称性知

$$t_{1-\alpha}(n)=-t_{\alpha}(n).$$

当 $n>45$ 时，$t_{\alpha}(n)\approx z_{\alpha}$.

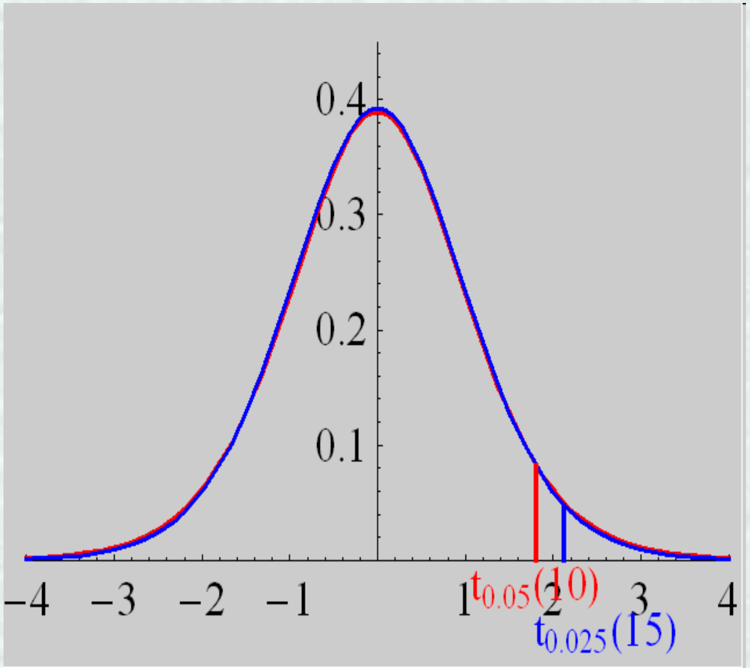
例2　设 $T \sim t(n)$, $t(n)$ 的上 α 分位点满足

$$P\{T > t_\alpha(n)\} = \int_{t_\alpha(n)}^{+\infty} t(y;n)\mathrm{d}y = \alpha,$$

求 $t_\alpha(n)$ 的值, 可通过查表完成.

$t_{0.05}(10) = 1.8125,$　附表3-1

$t_{0.025}(15) = 2.1315.$　附表3-2



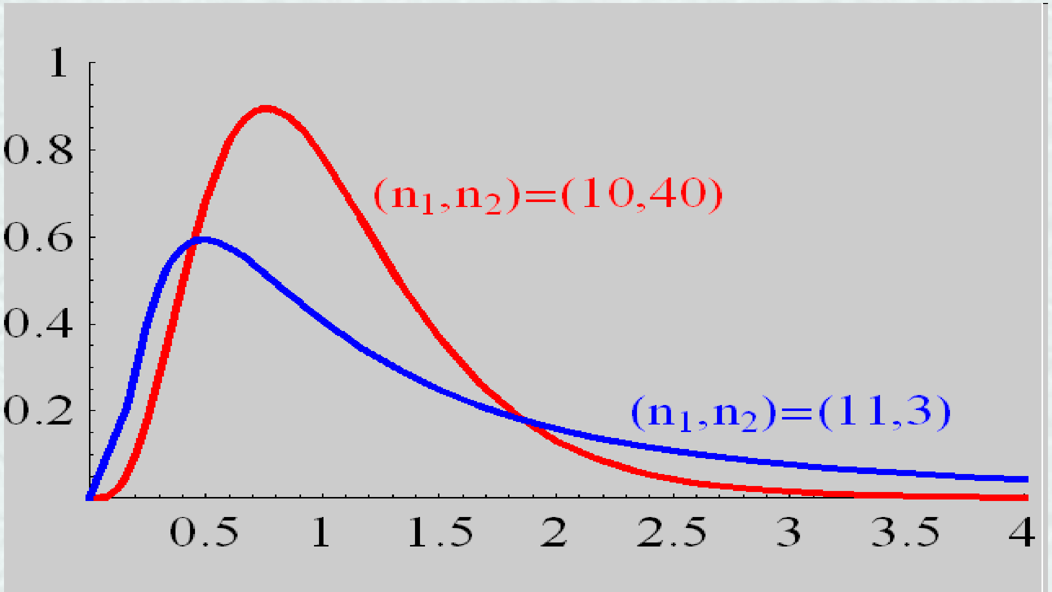
3. F分布

设$U \sim \chi^2(n_1)$, $V \sim \chi^2(n_2)$, 且U, V独立, 则称随机变量$F = \dfrac{U/n_1}{V/n_2}$服从自由度为(n_1, n_2)的F分布, 记为$F \sim F(n_1, n_2)$.

$F(n_1,n_2)$分布的概率密度为

$$\psi(y)=\begin{cases}\dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)\left(\dfrac{n_1}{n_2}\right)^{\frac{n_1}{2}} y^{\frac{n_1}{2}-1}}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)\left[1+\left(\dfrac{n_1 y}{n_2}\right)\right]^{\frac{n_1+n_2}{2}}}, & y>0,\\ 0, & \text{其他}.\end{cases}$$

F分布的概率密度曲线如图



根据定义可知,

若$F \sim F(n_1, n_2)$,

则$\dfrac{1}{F} \sim F(n_2, n_1)$.

F 分布的分位点

对于给定的α, $0<\alpha<1$, 称满足条件

$$P\{F > F_\alpha(n_1, n_2)\} = \int_{F_\alpha(n_1, n_2)}^{+\infty} \psi(y)\mathrm{d}y = \alpha$$

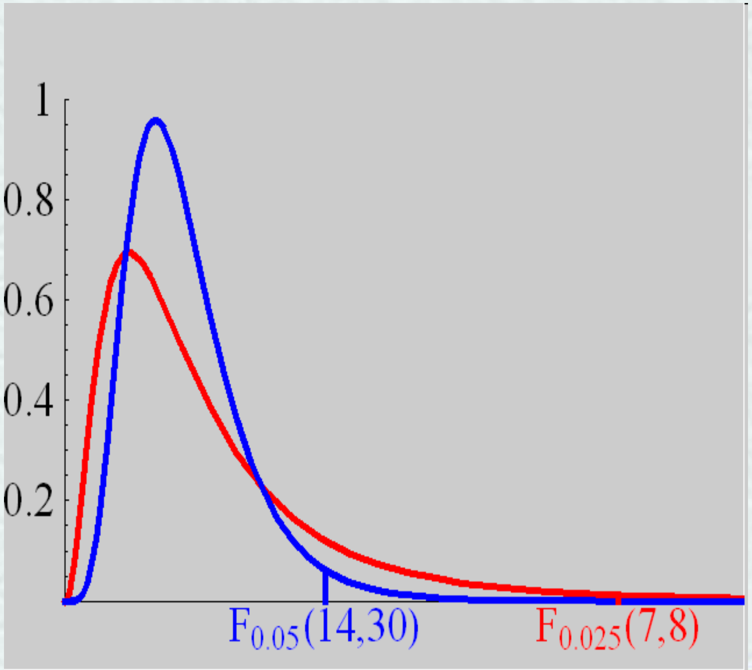
的点$F_\alpha(n_1, n_2)$为$F(n_1, n_2)$分布的上α分位点.

例3　设 $F(n_1, n_2)$分布的上α分位点满足

$$P\{F > F_\alpha(n_1, n_2)\} = \int_{F_\alpha(n_1, n_2)}^{+\infty} \psi(y)\mathrm{d}y = \alpha,$$

求 $F_\alpha(n_1, n_2)$的值, 可通过查表完成.

$$F_{0.025}(7,8) = 4.53,$$



F 分布的上 α 分位点具有如下性质:

$$F_{1-\alpha}(n_1, n_2) = \frac{1}{F_{\alpha}(n_2, n_1)}.$$

用来求分布表中未列出的一些上 α 分位点.

例 $F_{0.95}(12,9) = \dfrac{1}{F_{0.05}(9,12)} = \dfrac{1}{2.80} = 0.357$.

4. 正态总体的样本均值与样本方差的分布

定理一

设 $X_1, X_2, \cdots, X_n$ 是来自正态总体 $N(\mu,\sigma^2)$ 的样本, $\overline{X}$ 是样本均值, 则有 $\overline{X} \sim N(\mu, \sigma^2/n)$.

正态总体 $N(\mu,\sigma^2)$ 的样本均值和样本方差有以下两个重要定理.

定理二

设 $X_1, X_2, \cdots, X_n$ 是总体 $N(\mu, \sigma^2)$ 的样本, $\overline{X}, S^2$ 分别是样本均值和样本方差, 则有

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

(2) $\overline{X}$ 与 S^2 独立.

定理三 设 $X_1, X_2, \cdots, X_n$ 是总体 $N(\mu, \sigma^2)$ 的样本, $\overline{X}, S^2$ 分别是样本均值和样本方差, 则有

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

证明 因为 $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$,

且两者独立, 由 t 分布的定义知

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

定理四 设 $X_1, X_2, \cdots, X_{n_1}$ 与 $Y_1, Y_2, \cdots, Y_{n_2}$ 分别是具有相同方差的两正态总体 $N(\mu_1, \sigma^2), N(\mu_2, \sigma^2)$ 的样本, 且这两个样本互相独立, 设 $\overline{X} = \dfrac{1}{n_1}\sum_{i=1}^{n_1} X_i$,

$\overline{Y} = \dfrac{1}{n_2}\sum_{i=1}^{n_2} Y_i$ 分别是这两个样本的均值,

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

分别是这两个样本的方差, 则有

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

(2) 当$\sigma_1^2=\sigma_2^2=\sigma^2$时,

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \sim t(n_1+n_2-2),$$

其中 $S_w^2=\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}, \quad S_w=\sqrt{S_w^2}.$

证明 (1) 由定理二

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1), \quad \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1),$$

由假设 S_1^2, S_2^2 独立， 则由 F 分布的定义知

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

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

(2) 因为 $\overline{X}-\overline{Y} \sim N\left(\mu_1-\mu_2, \dfrac{\sigma^2}{n_1}+\dfrac{\sigma^2}{n_2}\right)$

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

由 $\dfrac{(n_1-1)S_1^2}{\sigma^2} \sim \chi^2(n_1-1)$, $\dfrac{(n_2-1)S_2^2}{\sigma^2} \sim \chi^2(n_2-1)$,

且它们相互独立, 故由 χ^2 分布的可加性知

$$V=\frac{(n_1-1)S_1^2}{\sigma^2}+\frac{(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_1+n_2-2),$$

由于 U 与 V 相互独立, 按 t 分布的定义.

$$\frac{U}{\sqrt{V/(n_1+n_2-2)}}$$

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

例：设总体$X \sim N(\mu,\sigma^2)$, $X_1, X_2, \ldots X_{16}$

是X的一个样本，求：

$$(1)\ P\{\frac{\sigma^2}{2} \le \frac{1}{16}\sum_{i=1}^{16}(X_i-\mu)^2 \le 2\sigma^2\}$$

$$(2)\ P\{\frac{\sigma^2}{2} \le \frac{1}{16}\sum_{i=1}^{16}(X_i-\bar{X})^2 \le 2\sigma^2\}$$

解：因为 $X_i \sim N(\mu,\sigma^2), i=1,2\ldots16$

所以$\dfrac{X_i-\mu}{\sigma} \sim N(0,1), i=1,2\ldots16$

所以 $\sum_{i=1}^{16}(\frac{X_i-\mu}{\sigma})^2 \sim \chi^2(16)$

则 $P\{\frac{\sigma^2}{2} \le \frac{1}{16}\sum_{i=1}^{n}(X_i-\mu)^2 \le 2\sigma^2\}$

$$= P\{\frac{1}{2} \le \frac{1}{16}\sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2 \le 2\}$$

$$= P\{8 \le \sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2 \le 32\}$$

$$= P\{8 \le \chi^2(16) \le 32\}$$

$$= P\{\chi^2(16) > 8\} - P\{\chi^2(16) > 32\}$$

$$=0.95-0.01=0.94$$

$$(2)\quad \because \sum_{i=1}^{16}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 \sim \chi^2(15)$$

$$\therefore P\left\{\frac{\sigma^2}{2} \le \frac{1}{16}\sum_{i=1}^{16}(X_i-\bar{X})^2 \le 2\sigma^2\right\}$$

$$= P\left\{\frac{1}{2} \le \frac{1}{16}\sum_{i=1}^{16}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 \le 2\right\}$$

$$= P\left\{8 \le \sum_{i=1}^{16}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 \le 32\right\}$$

$$= P\{8 \le \chi^2(15) \le 32\}$$

$$= P\{\chi^2(15) > 8\} - P\{\chi^2(15) > 32\}$$

$$=0.9-0.005=0.895$$

三、小结

两个最重要的统计量:

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

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

三个来自正态分布的抽样分布:

χ^2 分布, t 分布, F 分布.