§0 引言

数理统计是伴随着概率论的发展而发展起来的一个数学分支，它是研究如何有效地收集、整理和分析受随机因素影响的数据，并对所观察的问题作出推断或预测，为采取某种决策和行动提供依据或建议。

当前，数理统计的应用范围愈来愈广泛，已渗透到许多科学领域(如自然科学，工程技术，管理科学，人文社会科学等)，并成为科学研究不可缺少的工具。

特　点　　数理统计是以随机现象的观察试验取得的资料为出发点，以概率论为理论基础来研究随机现象，根据资料为随机现象选择数学模型，且利用数学资料来验证数学模型是否合理，在合理的基础上再研究它的特点、性质和规律性。

例如：讨论灯泡厂生产的灯泡寿命。

将灯泡厂所生产的某一批灯泡中抽出几个进行试验，试验前不知道该批灯泡的寿命有多长，概率分布是什么。试验后得到这几个灯泡的寿命，以此作为资料从而推测整批灯泡的使用寿命。为了研究它的分布，可以利用概率论中的指数分布作为灯泡寿命分布的数学模型，并用抽样试验数据来验证指数分布的合理性。如果合理，便可用指数分布来作为整批灯泡的寿命分布，进而对整批灯泡作出推断和预测等。

由此可以看出概率论与数理统计的区别与联系：

联 系　　二者都是研究和揭示随机现象的统计规律；概率论是理论基础，数理统计则是其实际应用。

区 别

- 概率论中，所研究**RV的分布是假设已知**的，以此为前提再去研究它的性质、特点和规律。

- 数理统计中，所研究**RV的分布是未知的或不完全知道**，通过对**RV**进行独立重复的观察，获得观察数据，对这些数据进行分析，从而对**RV**的分布作出种种合理的推断。

本课程主要介绍数理统计中统计推断内容：

估计问题　假设检验

第六章作为数理统计的基础，介绍一些基本概念，常用统计量及抽样分布。

第六章 样本与抽样分布

总体

随机样本

统计量

抽样分布

常用统计量及其抽样分布

§1 随机样本

一、总体与个体

定义 （一个统计问题总有它明确的研究对象）

研究对象的全体称为**总体**/**母体**（**Population**）；

总体中每个成员称为**个体**（**Unit**）；

总体中所包含的个体的个数称为**总体的容量**。

总体分类
- 有限总体：总体容量有限
- 无限总体：总体容量无限

从数理统计角度来说，我们关心的是总体中的个体的某项数量指标(如人的身高、灯泡的寿命,汽车的耗油量等)。由于每个个体的出现是随机的，所以相应的数量指标的出现也带有随机性。从而可以把这种数量指标看作一个RV X，因此RV X 的分布就是该数量指标在总体中的分布。

总体就可以用一个RV及其分布来描述。

因此，在理论上可以把总体与RV等同起来，对总体分布的研究就是对相应RV分布的研究。

例如：

某工厂生产的一批灯泡的寿命的全体——一个总体；

每个灯泡的寿命——一个个体。

若总体可以用**RV** ***X***表示，则对该批灯泡寿命的研究，就是对**RV** ***X***的研究。

又如：研究某地区中学生的营养状况时，若关心的数量指标是身高和体重，用**RV X**和**Y**分别来表示，那么总体就可以用二维**RV**（**X,Y**）来表示。

因此，统计中总体这个概念的要旨是：

总体就是一个随机变量(一维**or**多维**RV**)。

二、样本

1. 定义

对总体X进行n次重复独立的观察，其结果依次记为$X_1, X_2, \cdots, X_n$，若它们相互独立且与总体同分布，则称$X_1, X_2, \cdots, X_n$为取自总体X的一个容量为n的简单随机样本，简称样本 Sample（或子样），它们的观察值$x_1, x_2, \ldots, x_n$称为样本值，也称为总体X的n个独立的观测值。

说 明

① 总体分布的获取

理想方法 逐个对个体进行观察，实际往往不可行。

实际方法 抽样(即抽取一定容量的样本)，利用样本来推断总体分布。

例如：研究灯泡寿命的分布。

a. 无限总体时不可能用理想方法

b. 寿命试验为破坏性试验

抽样

② (简单随机)样本的获取

无限总体：采用不放回抽样得到样本。

有限总体：采用放回抽样得到样本。

但放回抽样使用不方便，实际中若满足：

$$\frac{\text{总体中个体总数}}{\text{抽取的样本容量}} \geq 10$$

即可将不放回抽样近似当作放回抽样处理。

③ 样本的特点

独立性　　样本 $X_1, X_2, \ldots, X_n$是相互独立的RV。

同分布性　　样本 $X_1, X_2, \ldots, X_n$与所考察总体X有相同的分布。

独立同分布 **Independent and identically distributed**

简记为 ***i.i.d.***

④ 样本的二重性

数的属性　在一次具体的观察或试验中，样本是一批观察值(测量值)，是已知的数。

随机性　在具体的观察和试验中，由于受到各种随机因素的影响，则在不同的观察中样本的取值可能不同，具有随机性。

综上

样本就是与总体独立同分布的*n*个相互独立的RV

2. 样本的JCDF和JPDF

设总体的CDF和PDF分别为$F(x)$和 $f(x)$，将样本可以看成是一个n维随机变量(向量)，即$(X_1, X_2, \ldots, X_n)$

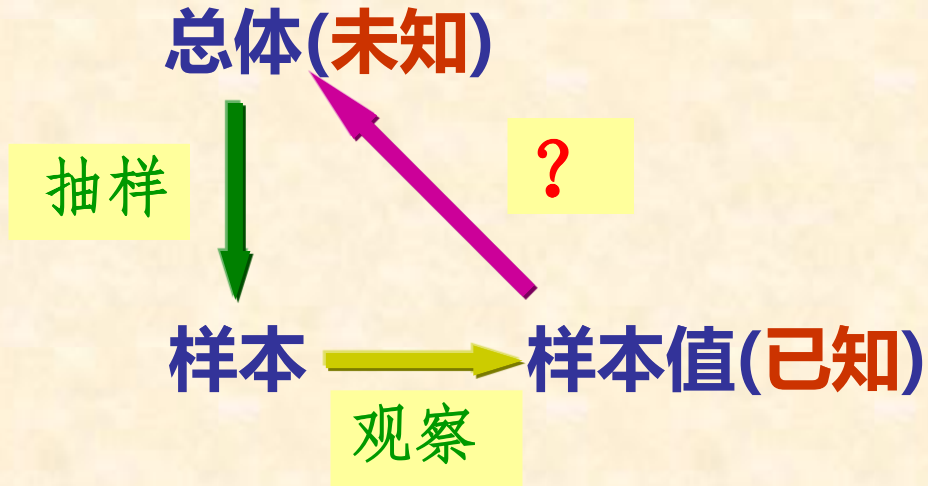
则：

样本的JCDF为 $F^*(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} F(x_i)$

样本的JPDF为 $f^*(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$

今后，当说到$X_1, X_2, \ldots, X_n$是取自某总体的样本时，若不特别说明，就指简单随机样本。

3. 总体、样本和样本值间的关系



① 为何可以用样本值去推断总体?

这是由于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，则样本值反映和体现了总体的信息，因而可以由样本值去推断总体。

② 如何用样本值去推断总体?

虽然样本值中包含总体的信息，但较为分散，一般不宜直接用于统计推断。样本是联系总体和样本值的桥梁，需要对样本进行加工处理，通过构造一些样本的函数，把样本中所含的信息集中起来。因此，由样本值间接利用样本的函数来对总体进行推断。

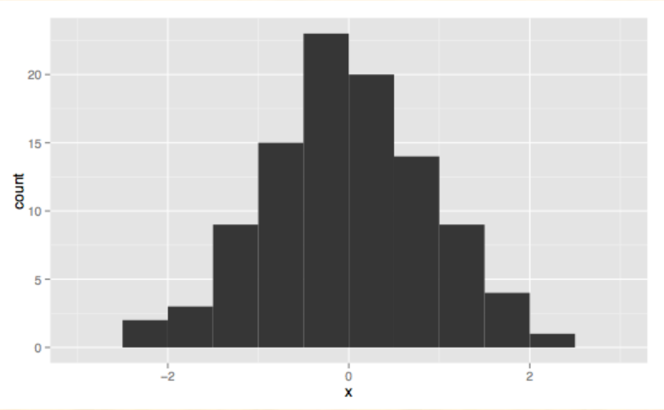
③ 如何对样本进行加工?

统计量及其分布（§3 内容）

§2 直方图和箱线图

两种概括RV分布的图示方法（自学）

直方图（Histogram）



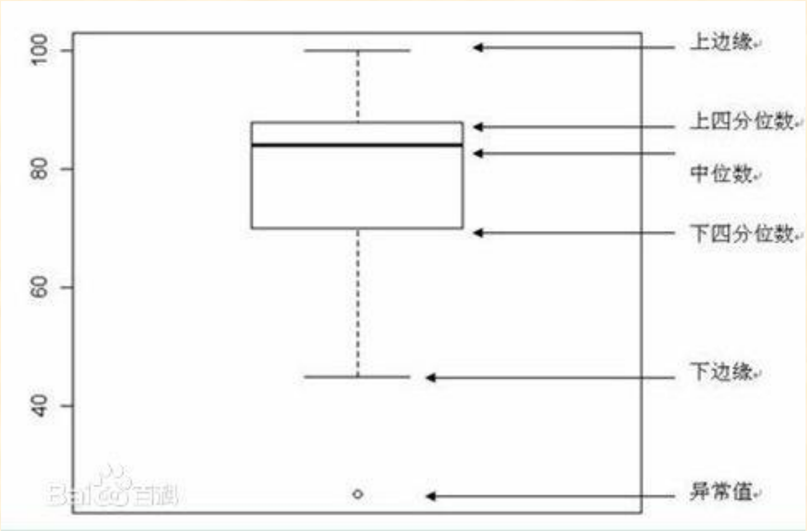
直方图，又称质量分布图，是一种统计报告图，由一系列高度不等的矩形表示数据分布的情况。直方图是数值数据分布的精确图形表示。

目 的

直方图是统计整理样本观察值的数据，用于初步了解总体的分布情况，以便对总体的分布型式做出假设。

箱线图（Box-plot）

箱线图，又称为盒须图、盒式图或箱形图，是一种用作显示一组数据分散情况资料的统计图。



主要包含六个数据节点：将一组数据从大到小排列，分别计算出他的上边缘，上四分位数Q_3，中位数，下四分位数Q_1，下边缘，还有一个异常值。

§3 抽样分布

一、统计量

定义　设$X_1, X_2, \ldots, X_n$是来自总体X的一个样本，$g(x_1, x_2, \ldots, x_n)$是不含未知参数的n元函数，则称$g(X_1, X_2, \ldots, X_n)$是一个**统计量**（**Statistic**）。

说　明

① 不含未知参数　是为了得到样本的观察值后能立即算出统计量的观察值，以便于利用样本值去推断总体，使其不受未知因素的影响。

说 明

② **统计量是一个RV。**

③ 若$x_1, x_2, \ldots, x_n$是样本$X_1, X_2, \ldots, X_n$的一组观察值，则$g(x_1, x_2, \ldots, x_n)$也是统计量 $g(X_1, X_2, \ldots, X_n)$的一个观察值。

已知来自总体X 的一个样本$X_1, X_2, \ldots, X_5$，判断下列函数是否为统计量？

$X_1 + X_2$ ✓ $\underset{1\le i\le 5}{Max}(X_i)$ ✓ $(X_5 - X_1)^2$ ✓

$\sum_{i=1}^{5}\left[X_i - E(X)\right]$ ✗ $\sum_{i=1}^{n} X_i^2 \Big/ D(X)$ ✗

二、常用统计量

设$X_1, X_2, \dots, X_n$是来自总体X的一个样本，$x_1, x_2, \dots, x_n$是其样本值。

1. 样本均值

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

它反映了总体均值的信息

$E(X)$

2. 样本方差

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

它反映了总体方差的信息

$D(X)$

样本标准差

$$S=\sqrt{S^2}$$

说 明

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

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

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

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

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

3. 样本k(原点)阶矩

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, (k = 1, 2, \ldots)$$

$E\left(X^k\right)$

它反映了总体k阶矩的信息

4. 样本k阶中心矩

$$B_k = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^k, (k = 1, 2, \ldots)$$

$E\left\{\left[X - E(X)\right]^k\right\}$

它反映了总体k阶中心矩的信息

$\bar{X}$和S^2是样本？矩

$\bar{X}$是样本一阶原点矩，S^2不是样本矩。

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

上述统计量的观察值的名称与统计量名称一致，只要将大写字母换成小写即可。

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

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

$$a_k=\frac{1}{n}\sum_{i=1}^{n}x_i^k,(k=1,2,\ldots)$$

$$b_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^k,(k=1,2,\ldots)$$

5. 次序统计量

将样本$X_1, X_2, \ldots, X_n$的观察值$x_1, x_2, \ldots, x_n$按由小到大排序，并重新编号为$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$，定义$X_{(k)}$的取值为$x_{(k)}$，$k=1,2,\ldots,n$，由此得到的$X_{(1)}, X_{(2)}, \ldots, X_{(n)}$称为样本$X_1, X_2, \ldots, X_n$的次序统计量，

$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$称为次序统计量的观察值，

$X_{(k)}$称为样本的第k个次序统计量，

$X_{(1)}=\mathrm{Min}(X_1, X_2, \ldots, X_n)$称为最小次序统计量，

$X_{(n)}=\mathrm{Max}(X_1, X_2, \ldots, X_n)$称为最大次序统计量。

注意 由于每个$X_{(k)}$都是样本$X_1, X_2, \ldots, X_n$的函数，所以次序统计量也是RV，并且他们一般不相互独立。

定理 设总体X的CDF为$F(x)$，PDF为$f(x)$，$X_{(1)}$, $X_{(2)}, \ldots, X_{(n)}$为总体X的样本$X_1, X_2, \ldots, X_n$的次序统计量，则：

$X_{(1)}$的PDF为 $$f_{X_{(1)}}(x) = n\left[1 - F(x)\right]^{n-1} f(x)$$

$X_{(n)}$的PDF为 $$f_{X_{(n)}}(x) = n\left[F(x)\right]^{n-1} f(x)$$

补充　　**样本中位数**

$$x_{0.5}=\begin{cases} x_{(k+1)}, & n=2k+1 \\ \left[x_{(k)}+x_{(k+1)}\right]/2, & n=2k \end{cases}$$

说明

① 样本中$\geqslant x_{0.5}$和$\leqslant x_{0.5}$的观察值各为50%。

② 样本中位数描述了样本位置的特征，具有和样本均值类似的含义，但样本中位数不受样本异常值的影响，同时也容易计算，也可以作为总体均值的估计；但其分布不易计算，故在理论讨论时会带来一定困难。

三、经验分布函数

定义 设$X_1, X_2, \ldots, X_n$是总体X的一个样本，用$S(x)$表示$X_1, X_2, \ldots, X_n$中不大于x的RV个数，其中$x \in R$，**经验分布函数**$F_n(x)$为

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

经验分布函数$F_n(x)$的观察值

一般，设$X_1, X_2, \ldots, X_n$是总体X的一个样本，将其样本值$x_1, x_2, \ldots, x_n$按由小到大次序排序，并重新编号为$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$，则经验分布函数$F_n(x)$的观察值为

$$F_n(x) = \begin{cases} 0, & x < x_{(1)} \\ \dfrac{k}{n}, & x_{(k)} \le x < x_{(k+1)}, k = 1, 2, \cdots, n-1 \\ 1, & x \ge x_{(n)} \end{cases}$$

经验分布函数的性质

1. 当给定样本值 $x_1, x_2, \ldots, x_n$时，$F_n(x)$是一个分布函数，因为它具有CDF的特征，即

- $0 \leqslant F_n(x) \leqslant 1$；
- $F_n(-\infty)=0$，$F_n(+\infty)=1$；
- $F_n(x)$非减且右连续。

2. $F_n(x)$是样本的函数，则$F_n(x)$是RV，且 $nF_n(x) \sim b(n, F(x))$, 其中$F(x)$是总体X 的CDF，进而 $E[F_n(x)]=F(x),\ D[F_n(x)]=\dfrac{1}{n}F(x)[1-F(x)]$

经验分布函数的性质2说明

由经验分布函数定义知 $F_n(x)=\frac{1}{n}S(x) \Rightarrow S(x)=nF_n(x)$

$S(x)$表示$X_1,\ldots,X_n$中不大于x的RV个数, 这相当于我们进行了n重伯努利试验, 每次试验中要么事件$\{X_i \le x\}$发生, 要么事件$\{X_i > x\}$发生, 且事件$\{X_i \le x\}$发生的概率为$P\{X_i \le x\}=F(x)$（样本与总体独立同分布), 则在n重伯努利试验中事件$\{X_i \le x\}$发生的次数记为 $S(x)$, 故其分布为

$$S(x)\sim b(n,F(x)) \Rightarrow nF_n(x)\sim b(n,F(x))$$

经验分布函数的性质

3. 当$n\to\infty$时，$F_n(x)\xrightarrow{P}F(x)$ 。

说 明　当n充分大时，$F_n(x)$与$F(x)$差别微小。

　　数理统计中，就是利用已知样本来推断未知总体，由此性质可知：当n充分大时，就可以用$F_n(x)$近似总体的CDF $F(x)$，这也就是用样本来推断总体的理论依据。

四、样本矩与总体矩的关系

设总体X的k阶矩$E(X^k)=\mu_k$存在，则当$n\to\infty$时，样本的k阶矩

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

说明　由于$X_1,X_2,\cdots,X_n$独立且与X同分布，

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

且存在$E(X_i^k)=\mu_k\ (k=1,2,\cdots,n)$

由辛钦大数定律知：

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

利用依概率收敛的性质知：

$$g\left(A_1, A_2, \ldots, A_k\right) \xrightarrow{P} g\left(\mu_1, \mu_2, \ldots, \mu_k\right)$$

其中，g为连续函数。

这就是下一章中矩估计法的理论依据。

五、抽样分布

定义　统计量的分布称为**抽样分布**, 也称**诱导分布**。

抽样分布在研究统计量的性质和评价一个统计推断的优良性等方面十分重要。

总体分布已知时，抽样分布是确定，但要精确求出抽样分布却很困难。

本节主要介绍来自**正态总体**的几个常用统计量的分布：也即**统计学的三大抽样分布**。

1. χ^2分布（卡方分布，**Chi-square Distribution**）

① **定义**

设$X_1, X_2, \ldots, X_n$ 是来自总体N（0, 1）的样本，则称统计量

$$\chi^2=\sum_{i=1}^{n}X_i^2$$

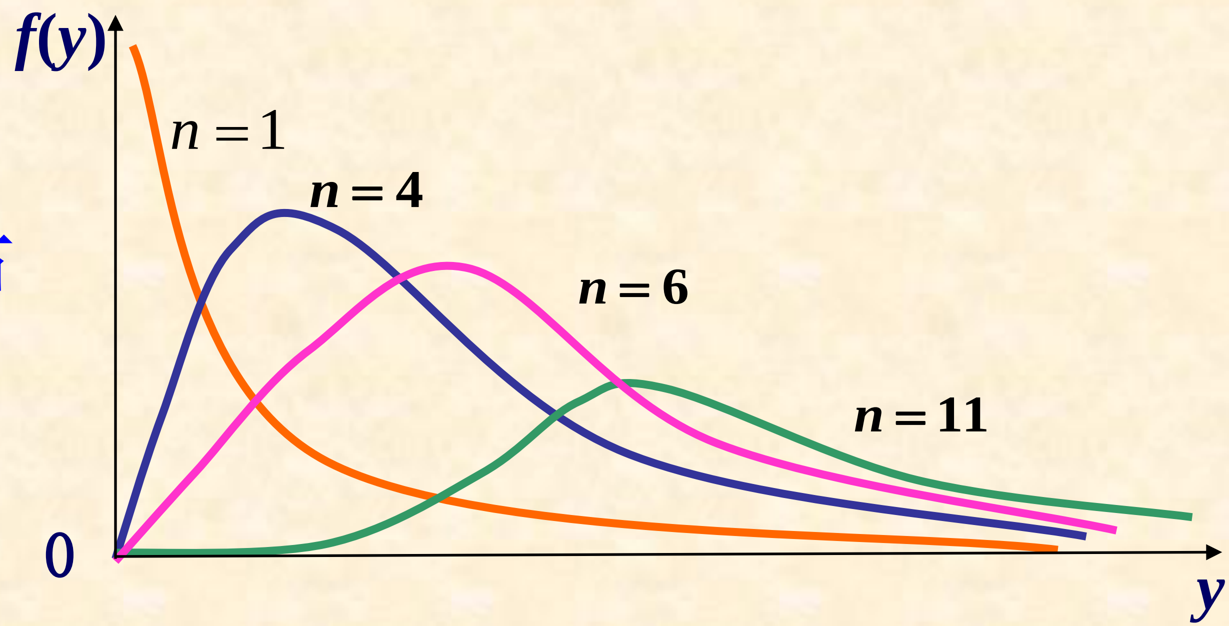
服从自由度为n的 χ^2分布，记为$\chi^2 \sim \chi^2(n)$。

自由度：指$\chi^2=\sum_{i=1}^{n}X_i^2$等式中包含的独立变量的个数。

② χ^2分布的PDF

$$f(y)=\begin{cases}\dfrac{1}{2^{n/2}\Gamma\left(\dfrac{n}{2}\right)}y^{\frac{n}{2}-1}e^{-\frac{y}{2}}, & y>0\\ 0, & else\end{cases}$$

PDF图如右



③ χ^2分布与Γ分布关系

若$X \sim \Gamma(\alpha,\beta)$, 其PDF为

$$f_X(x)=\begin{cases}\dfrac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}, & x>0\\ 0, & else\end{cases}$$

Γ分布特例 $\Gamma(1,\lambda) = Exp(\lambda)$

$$f_X(x)=\begin{cases}\lambda e^{-\lambda x}, & x>0\\ 0, & else\end{cases}$$

$$\Gamma\left(\frac{1}{2},\frac{1}{2}\right) = \chi^2(1)$$

$$f_X(x)=\begin{cases}\dfrac{1}{\sqrt{2\pi}}x^{-1/2}e^{-x/2}, & x>0\\ 0, & else\end{cases}$$

由于$X \sim N(0,1)$，$i=1,2,\ldots,n$

$$\therefore X_i^2 \sim \chi^2(1) = \Gamma\left(\frac{1}{2},\frac{1}{2}\right)$$

又$\because X_1, X_2, \ldots, X_n$相互独立

$\therefore X_1^2, X_2^2, \ldots, X_n^2$也相互独立

利用Γ分布的可加性，得$\chi^2 = \sum_{i=1}^{n} X_i^2 \sim \Gamma\left(\frac{n}{2},\frac{1}{2}\right)$

由此可知:

χ^2分布与Γ分布的关系为 $\Gamma\left(\frac{n}{2},\frac{1}{2}\right) = \chi^2(n)$

④ χ^2分布的性质

1）若$X_1 \sim \chi^2(n_1)$, $X_2 \sim \chi^2(n_2)$, 且X_1与X_2独立, 则

$$X_1 + X_2 \sim \chi^2(n_1 + n_2)$$

这个性质叫χ^2分布的可加性。*Additivity*

（多个独立χ^2分布时也适用）

2）若$X_1, X_2, \ldots, X_n$相互独立, 都服从正态分布

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

④ χ^2分布的性质

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

4) 若$X \sim \chi^2(n)$，则当n充分大时，$\dfrac{X-n}{\sqrt{2n}} \overset{\text{近似}}{\sim} N(0,1)$

也即 $X \overset{\text{近似}}{\sim} N(n, 2n)$。 应用中心极限定理

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

说明：$\chi^2=\sum\limits_{i=1}^{n}X_i^2$, 其中$X_i \sim N(0,1), i=1,2,\cdots,n,$

且相互独立,

$$E(X_i^2)=D(X_i)+E^2(X_i)=1$$

$$D(X_i^2)=E(X_i^4)-E^2(X_i^2)=3-1^2=2$$

$$\therefore E(\chi^2)=E\left(\sum_{i=1}^{n}X_i^2\right)=\sum_{i=1}^{n}E(X_i^2)=n$$

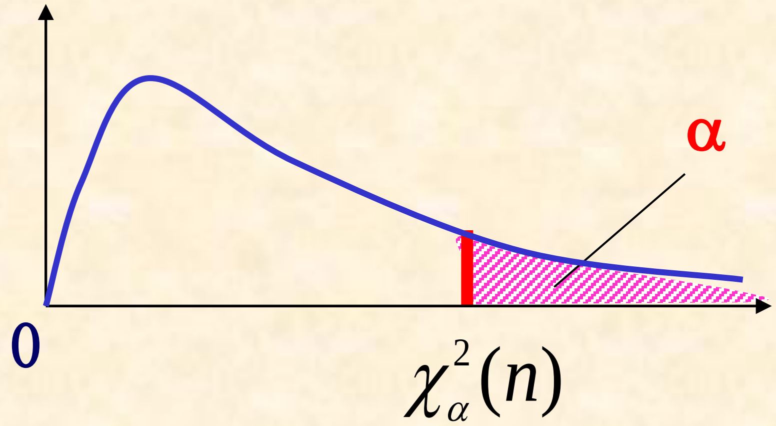
$$D(\chi^2)=D\left(\sum_{i=1}^{n}X_i^2\right)=\sum_{i=1}^{n}D(X_i^2)=2n$$

⑤ χ^2分布的分位点

定义　$\forall 0<\alpha<1$, 称满足

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

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



⑤ χ^2分布的分位点

对不同的α和n, χ^2分布的上α分位点$\chi_\alpha^2(n)$的值已制成表格，见P386附表5。

当$n \le 40$时，直接查表得$\chi_\alpha^2(n)$值；

当$n > 40$时，Fisher给出

$$\chi_\alpha^2(n) \approx \frac{1}{2}\left(z_\alpha + \sqrt{2n-1}\right)^2,$$

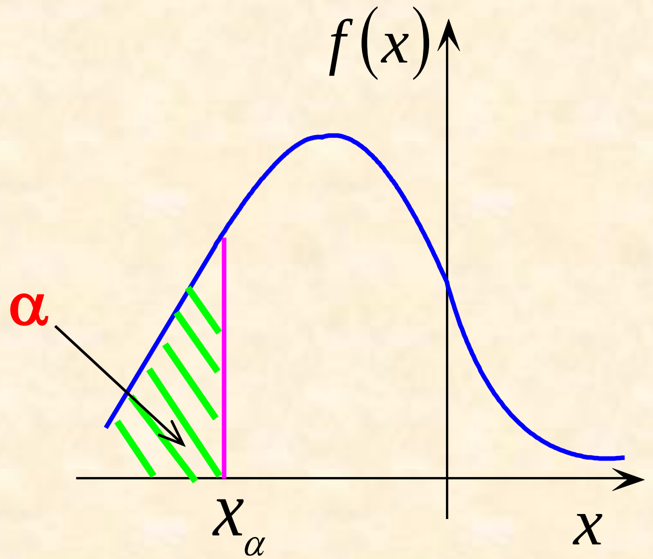
其中，z_α为$N(0,1)$的上α分位点；n越大，误差越小。

补充 下分位点(P118)

定义 对连续型RV, $\forall 0<\alpha<1$, 称满足

$$P\{X \le x_\alpha\} = \int_{-\infty}^{x_\alpha} f(y)\,\mathrm{d}y = \alpha$$

的点x_α就是此分布的下α分位点。



特别地 当$\alpha=0.5$时, $x_{0.5}$称为此分布的中位数。

2. t 分布

① 定义

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

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

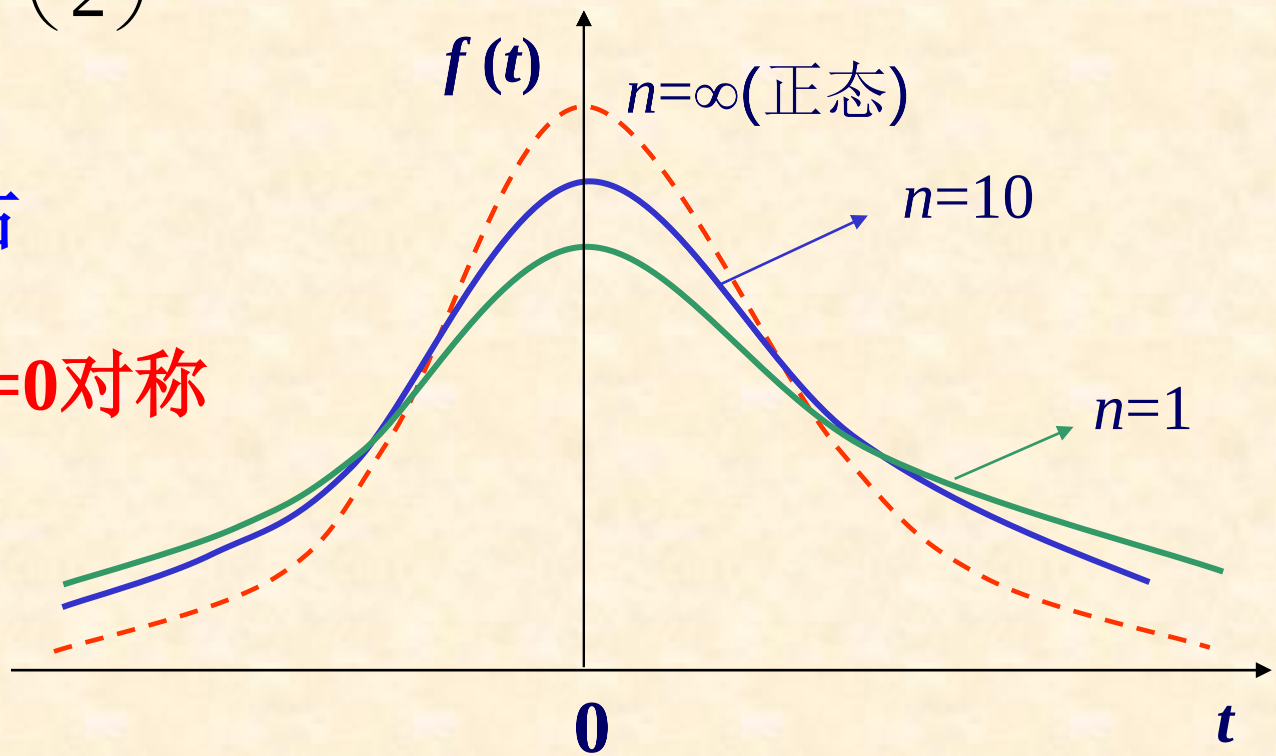
服从自由度为n的 t 分布，又称为学生氏分布（**Student Distribution**），记为 $t \sim t(n)$。

② *t* 分布的PDF

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

PDF图如右

***h*(*t*)关于*t* =0对称**



③ t 分布的性质

1) 若$t \sim t(n)$, 则$E(t) = 0$, $D(t) = \dfrac{n}{n-2}, (n > 2)$。

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

即：当$n \to \infty$时，$t \overset{近似}{\sim} N(0,1)$。

③ t分布的性质

3) 若$t \sim t(n)$, 则t只存在$k\,(k<n)$阶矩。

特别地, 当$n=1$时的t分布称为标准柯西分布;

柯西分布的任意阶矩都不存在。

t分布的$k\,(k<n)$阶矩:

$$E\left(t^k\right)=\begin{cases} n^{k/2}\dfrac{\Gamma\left(\dfrac{k+1}{2}\right)\Gamma\left(\dfrac{n-k}{2}\right)}{\Gamma\left(\dfrac{1}{2}\right)\Gamma\left(\dfrac{n}{2}\right)}, & k\text{为偶数} \\ 0, & k\text{为奇数} \end{cases}$$

④ t 分布的分位点

定义 $\forall 0<\alpha<1$, 称满足

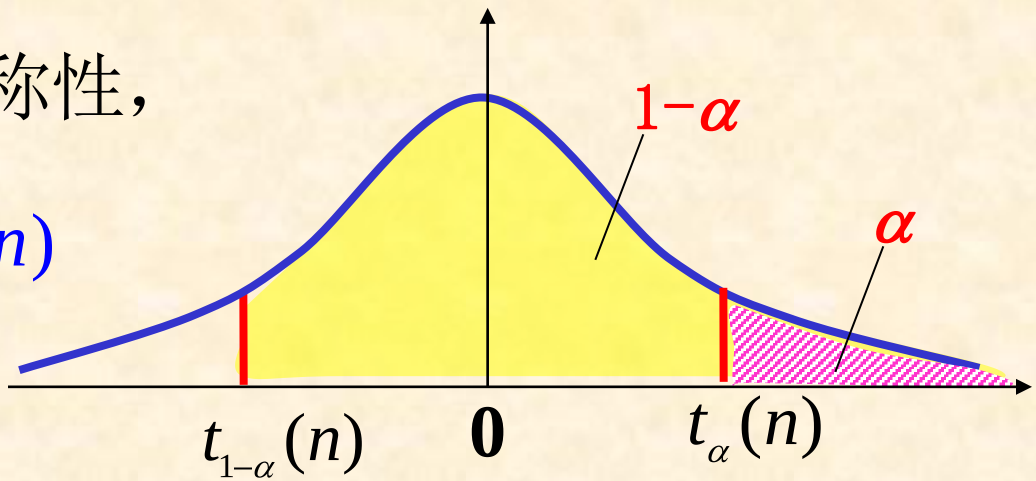
$$P\left\{t>t_\alpha(n)\right\}=\int_{t_\alpha(n)}^{\infty} h(t)\,dt=\alpha$$

的点$t_\alpha(n)$就是$t(n)$分布的上α分位点. (查P385附表4)

利用PDF曲线$h(t)$的对称性,

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

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



3. F 分布

① 定义

设$U \sim \chi^2(n_1)$, $V \sim \chi^2(n_2)$, 且U与V独立, 称RV

$$F = \frac{U/n_1}{V/n_2}$$

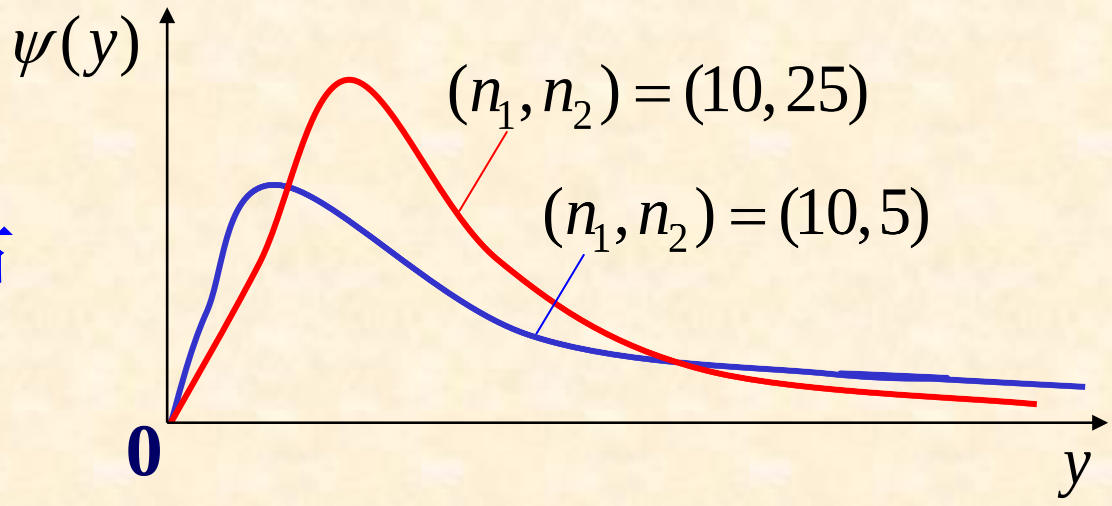
服从自由度为(n_1, n_2)的 F 分布,

记为 $F \sim F(n_1, n_2)$。

② *F* 分布的PDF

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

PDF图如右



③ *F* 分布的性质

1) 若$F \sim F(n_1, n_2)$, 则$E(F) = \dfrac{n_2}{n_2 - 2}, (n_2 > 2)$

$$D(F) = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)}, (n_2 > 4)。$$

2) $F \sim F(n_1, n_2)$, 则$\dfrac{1}{F} \sim F(n_2, n_1)$。

$$F = \frac{U/n_1}{V/n_2}$$

3) 若$t \sim t(n)$, 则$t^2 \sim F(1, n)$。

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

④ F 分布的分位点

定义　$\forall 0 < \alpha < 1$, 称满足

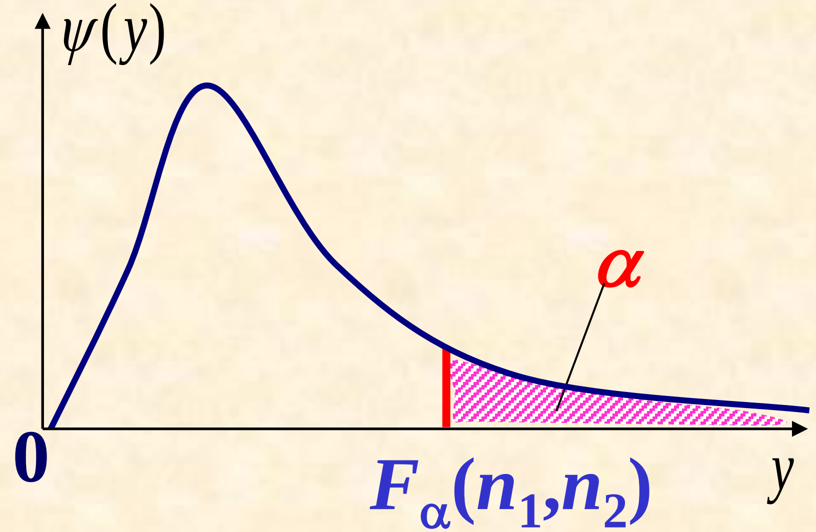
$$P\{F > F_\alpha(n_1,n_2)\} = \int_{F_\alpha(n_1,n_2)}^{\infty} \psi(y)\,dy = \alpha$$

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

F分布的上α分位点$F_\alpha(n_1,n_2)$可查P387附表6求得。

F分布上α分位点的性质：

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



证：$1-\alpha \overset{\text{定义}}{=} \mathrm{P}\left\{F > F_{1-\alpha}(n_1,n_2)\right\}$

$$=\mathrm{P}\left\{\frac{1}{F} < \frac{1}{F_{1-\alpha}(n_1,n_2)}\right\} = 1-\mathrm{P}\left\{\frac{1}{F} \geq \frac{1}{F_{1-\alpha}(n_1,n_2)}\right\}$$

由此可得：$\mathrm{P}\left\{\frac{1}{F} \geq \frac{1}{F_{1-\alpha}(n_1,n_2)}\right\} = \alpha \quad (1)$

又$\because$当$F \sim F(n_1,n_2)$，有$\frac{1}{F} \sim F(n_2,n_1)$

$$\therefore \mathrm{P}\left\{\frac{1}{F} \geq F_{\alpha}(n_2,n_1)\right\} = \alpha \qquad (2)$$

对比式(1)和(2)，得证 $F_{1-\alpha}(n_1,n_2) = \dfrac{1}{F_{\alpha}(n_2,n_1)}$

六、重要的抽样分布定理

1. 样本均值与样本方差的数字特征

设总体X的均值$E(X)=\mu$，方差$D(X)=\sigma^2$，

$X_1, X_2, \ldots, X_n$是来自总体X的一个样本，则

样本均值$\bar{X}$和样本方差S^2有

$$E(\bar{X})=\mu \qquad D(\bar{X})=\frac{\sigma^2}{n}$$

$$E(S^2)=\sigma^2$$

注：总体X为任意分布，上述3个等式都成立。

已知：$E(X)=\mu$，$D(X)=\sigma^2$

证：$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\cdot\mu = \mu$

$D(\bar{X}) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{1}{n^2} n\cdot\sigma^2 = \frac{\sigma^2}{n}$

$E(S^2) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = E\left[\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)\right]$

$= \frac{1}{n-1}\left[\sum_{i=1}^{n} E(X_i^2) - nE(\bar{X}^2)\right]$

$= \frac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right] = \sigma^2$ 得证

2. 样本均值的分布

1) 若总体X非正态分布，当n很大时，由独立同分布的中心极限定理知：

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \overset{\text{近似}}{\sim} N\left(\mu,\frac{\sigma^2}{n}\right)$$

其中，$E(X)=\mu$，$D(X)=\sigma^2$。

2) 若总体$X\sim N(\mu,\sigma^2)$，则样本均值 **P142 定理一**

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i\sim N\left(\mu,\frac{\sigma^2}{n}\right)$$

也即 $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$

3. 样本方差的分布

定理二 设$X_1, X_2, \ldots, X_n$是总体$X \sim N\left(\mu, \sigma^2\right)$的一个样本，$\bar{X}$和$S^2$分别是样本均值和样本方差，则

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

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

(2) $\bar{X}$与S^2独立。 **证略**

定理三 设$X_1, X_2, \ldots, X_n$是总体$X \sim N\left(\mu, \sigma^2\right)$的一个样本，$\bar{X}$和$S^2$分别是样本均值和样本方差，则

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

定理一 $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

定理四 设总体$X \sim N\left(\mu_1, \sigma_1^2\right), Y \sim N\left(\mu_2, \sigma_2^2\right)$, 且$X$与$Y$独立, $X_1, X_2, \ldots, X_{n_1}$是来自$X$的一个容量为$n_1$的样本, $Y_1, Y_2, \ldots, Y_{n_2}$是来自$Y$的一个容量为$n_2$的样本, $\bar{X}$和$\bar{Y}$分别是这两个样本的样本均值, S_1^2和S_2^2分别是两个样本的样本方差，则

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

定理二 $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

$$\left.\begin{aligned} \frac{(n_1-1)S_1^2}{\sigma_1^2} &\sim \chi^2(n_1-1) \\ \frac{(n_2-1)S_2^2}{\sigma_2^2} &\sim \chi^2(n_2-1) \end{aligned}\right\} \text{独立} \Rightarrow \frac{\frac{(n_1-1)S_1^2}{\sigma^2} \Big/ (n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma^2} \Big/ (n_2-1)} \sim F\left(n_1-1, n_2-1\right)$$

定理四 设总体$X \sim N\left(\mu_1, \sigma_1^2\right), Y \sim N\left(\mu_2, \sigma_2^2\right)$, 且$X$与$Y$独立, $X_1, X_2, \ldots, X_{n_1}$是来自$X$的一个容量为$n_1$的样本, $Y_1, Y_2, \ldots, Y_{n_2}$是来自$Y$的一个容量为$n_2$的样本, $\bar{X}$和$\bar{Y}$分别是这两个样本的样本均值, S_1^2和S_2^2分别是两个样本的样本方差，则

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

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

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)$$

定理一$\bar{X}\sim N\left(\mu_1,\frac{\sigma^2}{n_1}\right)$, $\bar{Y}\sim N\left(\mu_2,\frac{\sigma^2}{n_2}\right)$, $\bar{X}-\bar{Y}\sim N\left(\mu_1-\mu_2,\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}\right)$

$$\boxed{U=\frac{\bar{X}-\bar{Y}-\mu_1-\mu_2}{\sqrt{\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}}}}\sim N(0,1),\qquad \overset{\text{独立}}{\Rightarrow}\ \frac{U}{\sqrt{\frac{V}{n_1+n_2-2}}}\sim t(n_1+n_2-2)$$

得证

定理二

$$\left.\begin{array}{l}\dfrac{(n_1-1)S_1^2}{\sigma^2}\sim\chi^2(n_1-1)\\[2ex]\dfrac{(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_2-1)\end{array}\right\}\text{独立}\quad \boxed{V=\frac{(n_1-1)S_1^2}{\sigma^2}+\frac{(n_2-1)S_2^2}{\sigma^2}}\sim\chi^2(n_1+n_2-2)$$

设$\boxed{\text{总体}X \sim N\left(\mu,\sigma^2\right)}$, $X_1, X_2, \ldots, X_n$是来自X的一个样本, 样本方差S^2, 试求$D\left(S^2\right)$?

由定理二知: $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

且还可知: $\chi^2(n-1)$分布的方差为$2(n-1)$

即 $D\left[\dfrac{(n-1)S^2}{\sigma^2}\right] = 2(n-1)$

$$\Rightarrow \left[\frac{(n-1)}{\sigma^2}\right]^2 D\left(S^2\right) = 2(n-1) \qquad \boxed{\Rightarrow D\left(S^2\right) = \frac{2\sigma^4}{(n-1)}}$$

例1　设总体$X \sim N\left(\mu,\sigma^2\right), X_1, X_2,\ldots, X_n$是来自$X$的一个样本，样本均值$\bar{X}$，样本方差$S^2$，样本的二阶中心距 $B_2 = S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$，试求下列统计量的分布？

$$(1)\frac{nS_n^2}{\sigma^2},\quad (2)\frac{\bar{X}-\mu}{S_n\big/\sqrt{n-1}},\quad (3)\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}。$$

解：$(1)\ S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2 \Rightarrow nS_n^2 = \left(n-1\right)S^2$

由定理二知 $\frac{nS_n^2}{\sigma^2} = \frac{\left(n-1\right)S^2}{\sigma^2} \sim \chi^2(n-1)$

总体$X \sim N\left(\mu,\sigma^2\right)$,(2)$\dfrac{\bar{X}-\mu}{S_n \big/ \sqrt{n-1}}$

由定理三知 $\dfrac{\bar{X}-\mu}{S \big/ \sqrt{n}} \sim t(n-1)$

由于$nS_n^2 = \left(n-1\right)S^2 \quad \Rightarrow \quad \dfrac{S}{\sqrt{n}} = \dfrac{S_n}{\sqrt{n-1}}$

故$\dfrac{\bar{X}-\mu}{S_n \big/ \sqrt{n-1}} \sim t(n-1)$

(3) 由χ^2分布的性质2知 $\dfrac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n)$

例2 设样本$X_1, X_2, \ldots, X_{10}$来自总体$X \sim N\left(0, 0.3^2\right)$，试求：

(1) $P\left\{-1.2 < \bar{X} < 1.5\right\}$, (2) $P\left\{\sum_{i=1}^{10} X_i^2 > 1.44\right\}$,

(3) $P\left\{\frac{1}{10}\sum_{i=1}^{10}(X_i - \bar{X})^2 \le 2 \times 0.3^2\right\}$。

思路 **计算与统计量有关事件的概率，需知道这些统计量的分布**

解：(1) 可知 $\bar{X} \sim N\left(0, 0.3^2/10\right)$

$$P\left\{-1.2 < \bar{X} < 1.5\right\} = \Phi\left(\frac{1.5-0}{0.3/\sqrt{10}}\right) - \Phi\left(\frac{-1.2-0}{0.3/\sqrt{10}}\right)$$

$$= \Phi\left(5\sqrt{10}\right) - \Phi\left(-4\sqrt{10}\right) \approx 1$$

总体$X \sim N\left(0, 0.3^2\right)$,试求:(2) $P\left\{\sum_{i=1}^{10} X_i^2 > 1.44\right\}$?

易知$\frac{1}{0.3^2}\sum_{i=1}^{10} X_i \sim \chi^2(10)$

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

$$P\left\{\sum_{i=1}^{10} X_i^2 > 1.44\right\} = P\left\{\frac{1}{0.3^2}\sum_{i=1}^{10} X_i^2 > \frac{1.44}{0.3^2}\right\}$$

$$= P\left\{\chi^2(10) > 16\right\} \overset{\text{查附表5}}{\approx} 0.1$$

总体$X \sim N\left(0, 0.3^2\right)$，试求：(3) $P\left\{\frac{1}{10}\sum_{i=1}^{10}(X_i - \bar{X})^2 \le 2 \times 0.3^2\right\}$?

由定理二知 $\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2(n-1)$

已知$\sigma^2 = 0.3^2$，则

$$P\left\{\frac{1}{10}\sum_{i=1}^{10}(X_i - \bar{X})^2 \le 2 \times 0.3^2\right\} = P\left\{\frac{1}{0.3^2}\sum_{i=1}^{10}(X_i - \bar{X})^2 \le 2 \times 10\right\}$$

$$= P\left\{\chi^2(9) \le 20\right\} = 1 - P\left\{\chi^2(9) > 20\right\} \overset{\text{查附表5}}{\approx} 1 - 0.025$$

例3 设样本$X_1, X_2, \ldots, X_9$来自总体$X \sim N\left(\mu, \sigma^2\right), Y_1 = \frac{1}{6}\sum_{i=1}^{6} X_i,$

$Y_2 = \frac{1}{3}\left(X_7 + X_8 + X_9\right), S^2 = \frac{1}{2}\sum_{i=7}^{9}(X_i - Y_2)^2,$试求$\frac{\sqrt{2}\left(Y_1 - Y_2\right)}{S}$的分布。

解:可知$Y_1 \sim N\left(\mu, \frac{1}{6}\sigma^2\right)$, $Y_2 \sim N\left(\mu, \frac{1}{3}\sigma^2\right)$, 且$Y_1$与$Y_2$独立

故$Y_1 - Y_2 \sim N\left(0, \frac{1}{6}\sigma^2 + \frac{1}{3}\sigma^2\right) = N\left(0, \frac{1}{2}\sigma^2\right) \Rightarrow \boxed{\frac{Y_1 - Y_2}{\sigma/\sqrt{2}}} \sim N(0,1)$

对样本X_7, X_8, X_9由定理二知 $\frac{(n-1)S^2}{\sigma^2} = \boxed{\frac{2S^2}{\sigma^2}} \sim \chi^2(n-1) = \chi^2(2)$

还知Y_2与S^2独立，又Y_1与S^2独立 $\Rightarrow \frac{Y_1 - Y_2}{\sigma/\sqrt{2}}$与$\frac{2S^2}{\sigma^2}$独立

由t分布定义 $\frac{(Y_1 - Y_2)/\left(\sigma/\sqrt{2}\right)}{\sqrt{\frac{2S^2}{\sigma^2}\Big/2}} \sim t(2)$ 即$\frac{\sqrt{2}\left(Y_1 - Y_2\right)}{S} \sim t(2)$

例4 设总体$X\sim N\left(\mu,\sigma^2\right)$, $X_1,X_2,\ldots,X_{2n}$是来自X的样本,

$Y=\frac{1}{2n}\sum_{i=1}^{n}\left(X_{2i}-X_{2i-1}\right)^2$, 求$E\left(Y\right)$和$D\left(Y\right)$。

解:法(1) $Y=\frac{1}{2n}\sum_{i=1}^{n}\left(X_{2i}-X_{2i-1}\right)^2$

$$\overset{\text{展开}}{=}\frac{1}{2n}\sum_{i=1}^{n}\left(X_{2i}{}^2-2X_{2i}X_{2i-1}+X_{2i-1}{}^2\right)\overset{\text{整理}}{=}\frac{1}{2n}\left(\sum_{i=1}^{2n}X_i{}^2-\sum_{i=1}^{n}2X_{2i}X_{2i-1}\right)$$

$$E\left(Y\right)=E\left[\frac{1}{2n}\left(\sum_{i=1}^{2n}X_i{}^2-\sum_{i=1}^{n}2X_{2i}X_{2i-1}\right)\right]=\frac{1}{2n}\left[\sum_{i=1}^{2n}E\left(X_i{}^2\right)-2\sum_{i=1}^{n}E\left(X_{2i}\right)E\left(X_{2i-1}\right)\right]$$

$$\overset{\text{代入}}{=}\frac{1}{2n}\left[2n\left(\sigma^2+\mu^2\right)-2n\mu^2\right]=\sigma^2$$

$D\left(Y\right)$?

$$E\left(X_i{}^2\right)=D\left(X_i\right)+E^2\left(X_i\right)=\sigma^2+\mu^2$$

总体$X \sim N\left(\mu,\sigma^2\right), Y=\frac{1}{2n}\sum_{i=1}^{n}\left(X_{2i}-X_{2i-1}\right)^2$，求$E\left(Y\right)$和$D\left(Y\right)$。

法(2) 易知$X_{2i} \sim N\left(\mu,\sigma^2\right)$和$X_{2i-1} \sim N\left(\mu,\sigma^2\right)$, 且相互独立

则$X_{2i}-X_{2i-1} \sim N\left(0,2\sigma^2\right)$, 设$Y_i=\frac{X_{2i}-X_{2i-1}}{\sqrt{2}\sigma} \sim N\left(0,1\right)$， $i=1,\ldots,n$

由于$Y_1,\ldots,Y_n$相互独立，利用χ^2分布

$$\sum_{i=1}^{n}Y_i^2=\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(X_{2i}-X_{2i-1}\right)^2=\frac{nY}{\sigma^2} \sim\chi^2(n)$$

故 $E\left(\frac{nY}{\sigma^2}\right)=\frac{n}{\sigma^2}E\left(Y\right)=n \Rightarrow E\left(Y\right)=\sigma^2$

$$D\left(\frac{nY}{\sigma^2}\right)=\left(\frac{n}{\sigma^2}\right)^2 D\left(Y\right)=2n \Rightarrow D\left(Y\right)=\frac{2\sigma^4}{n}$$

小结

- 本节介绍了统计量的概念；
- 介绍了几个常见的统计量；
- 重点掌握几个重要统计量及其分布 。

第一次作业

Pages 147, 148：

第1, 2, 3, 10题

第二次作业

Pages 147, 148：

第4, 6, 7, 9题

练习1：设$X_1,\ldots,X_8$为来自总体$N(0,1)$的一个样本, 试求统计量

$$Y=\frac{(X_1+X_2)^2+(X_3+X_4)^2}{(X_5+X_6)^2+(X_7+X_8)^2}$$的分布。

解: 可知$X_1+X_2\sim N(0,2)$, 则$\frac{X_1+X_2}{\sqrt{2}}\sim N(0,1)\Rightarrow Y_1=\frac{(X_1+X_2)^2}{2}\sim\chi^2(1)$

类似可知$Y_2=\frac{(X_3+X_4)^2}{2}, Y_3=\frac{(X_5+X_6)^2}{2}, Y_4=\frac{(X_7+X_8)^2}{2}$也服从$\chi^2(1)$

Y_1,Y_2,Y_3,Y_4相互独立 $\Rightarrow Y_1+Y_2\sim\chi^2(2)$， $Y_3+Y_4\sim\chi^2(2)$

$\frac{(Y_1+Y_2)/2}{(Y_3+Y_4)/2}\sim F(2,2)$，也即$Y=\frac{(X_1+X_2)^2+(X_3+X_4)^2}{(X_5+X_6)^2+(X_7+X_8)^2}\sim F(2,2)$

直方图

目　的

直方图是统计整理样本观察值的数据，用于初步了解总体的分布情况，以便对总体的分布型式做出假设。

绘制直方图步骤：

1. 找出数据的最大值M和最小值m, 选择一个合适的区间$[a,b]$, a比m略小, b比M略大, 则区间$[a,b]$包含了所有样本观察值。

2. 将区间$[a,b]$等分为k个小区间, 小区间的长度$\Delta=(b-a)/k$称为组距, 端点称为组限。数出落在每个小区间内的数据的个数f_i， f_i为第i个子区间的频数 f_i。

3. 计算出第i个子区间的频率 f_i/n。

4. 在区间$[a, b]$上从左至右, 依次在各个子区间上, 以$(f_i/n)/\Delta$(频率/组距,即单位区间上的频率)为高作小矩形。即画出了直方图。

显然，每个小区间上矩形的面积就等于数据落在该小区间的频率f_i/n。n很大时，频率接近于概率。因此，一般来说每个小区间上的矩形面积接近于概率密度函数曲线之下该小区间之上的曲边梯形的面积。

于是，一般来说：

直方图的外轮廓曲线接近于总体的概率密度曲线。

说明 (1) 划分子区间数目

$$n<50,\ k=5\sim6;\quad n\geq50,\ k=10\sim20.$$

(2) 子区间端点通常取比数据精度高一位，以免数据落在分点上。