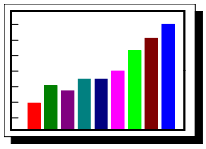
第1章 数理统计的基本概念

1.1 数理统计的基本问题

数理统计学是一门应用性很强的学科. 它是研究怎样以有效的方式收集、整理和分析带有随机性的数据，以便对所考察的问题作出推断和预测，直至为采取一定的决策和行动提供依据和建议.



数理统计------利用数据来分析对象满足的概率规律.

分为两大类：

一、如何科学地安排试验, 以获取有效的随机数据
-----描述统计学, 如：试验设计、抽样方法等。

二、研究如何分析所获得的随机数据, 对所研究的问题进行科学的、合理的估计和推断, 尽可能地为采取一定的决策提供依据, 作出精确而可靠的结论
--------推断统计学, 如:参数估计、假设检验等。

例：44某厂生产一型号的合金材料, 用随机的方法选取100个样品进行强度测试，于是面临下列几个问题：

1、估计这批合金材料的强度均值是多少?
(参数的点估计)

2、强度均值在什么范围内？(参数的区间估计)

3、若规定强度均值不小于某个定值为合格，那么这批材料是否合格？ (参数的假设检验)

4、这批合金的强度是否服从正态分布?(分布检验)

5、若这批材料是由两种不同工艺生产的，那么不同的工艺对合金强度有否影响？若有影响，那一种工艺生产的强度较好？(方差分析)

6、若这批合金由几种原料用不同的比例合成，那么如何表达这批合金的强度与原料比例之间的关系？
(回归分析问题)

1.2 总体和样本

一、总体、个体和样本

1. 总体　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)，总体中每个对象称为个体.

研究某批灯泡的质量
该批灯泡寿命的全体就是总体

考察国产 轿车的质量
所有国产轿车每公里耗油量的全体就是总体

不过在统计研究中，人们关心总体仅仅是关心其每个个体的一项(或几项)数量指标和该数量指标在总体中的分布情况. 这时，每个个体具有的数量指标的全体就是总体.

称总体中所含个体的数目为总体容量，总体容量有限的称为有限总体，总体容量无限的称为无限总体.

从另一方面看：

统计的任务，是根据从总体中抽取的样本，去推断总体的性质.

由于我们关心的是总体中的个体的某项指标(如人的身高、体重，灯泡的寿命，汽车的耗油量…)，　　所谓总体的性质，无非就是这些指标值集体的性质.

概率分布是刻划这种集体性质最适当的工具. 因此在理论上可以把总体与概率分布等同起来.

如研究某批灯泡的寿命时，关心的数量指标就是寿命，那么，此总体就可用描述其寿命的随机变量X或用其分布函数$F(x)$表示.

再如，若研究某地区中学生的营养状况时，关心的数量指标是身高和体重我们用X和Y分别表示身高和体重，那么此总体就可用二维随机变量(X,Y)或其联合分布函数$F(x,y)$来表示.

总体概念的要旨：　总体就是一个具有概率分布的随机变量。

2. 样本

为推断总体分布及各种特征，按一定规则从总体中抽取若干个体进行观察试验以获得有关总体的信息. 这一抽取过程称为抽样，所抽取的部分个体称为样本. 样本中所包含的个体数目称为样本容量.

从国产轿车中抽5辆进行耗油量试验　样本容量为5

抽到哪 5 辆是随机的！

样本是随机变量

容量为n的样本可以看作一n维随机变量$(X_1, X_2, \cdots, X_n)$.

但是，一旦取定一组样本，得到的是n个具体的数$x_1, x_2, \cdots, x_n$，称为样本$(X_1, X_2, \cdots, X_n)$的一组观测值，简称样本值.

抽样的目的是为了对总体进行统计推断，为了使抽取的样本能很好地反映总体的信息，必须考虑抽样方法.

最常用的一种抽样方法叫作简单随机抽样，　它要求抽取的样本$X_1, X_2, \cdots, X_n$满足下面两点：

1.独立性：　$X_1, X_2, \cdots, X_n$　是相互独立的随机变量；

2.代表性：　X_i $(i=1,2,\cdots,n)$ 与所考察的总体 X 同分布.

由简单随机抽样得到的样本称为简单随机样本，它可以用与总体同分布的 n 个相互独立的随机变量$X_1, X_2, \cdots, X_n$ 表示.

简单随机样本是应用中最常见的情形，今后，说到“$X_1, \cdots, X_n$是取自某总体的样本”时，若不特别说明，就指简单随机样本.

随机向量（$X_1, X_2, \cdots, X_n$）所有可能取值的全体称为样本空间，一个样本值（$x_1, x_2, \cdots, x_n$）就是样本空间中的一个点。

若总体X的分布函数为$F(x)$，则其简单随机样本的联合分布函数为

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

若总体X的概率密度为$f(x)$，则其简单随机样本的联合概率密度为

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

例1　设总体 X 服从两点分布 $B(1, p)$，其中$0 < p < 1$，$(X_1, X_2, \cdots, X_n)$是来自总体的样本，求样本 $(X_1, X_2, \cdots, X_n)$的分布律.

解　总体 X 的分布律为

$$P\{X = x\} = p^x (1-p)^{1-x} \quad (x = 0, 1)$$

因为$X_1, X_2, \cdots, X_n$相互独立，

且与 X 有相同的分布，

所以$(X_1, X_2, \cdots, X_n)$的分布律为

$$P\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\}$$

$$= P\{X_1 = x_1\}P\{X_2 = x_2\}\cdots P\{X_n = x_n\}$$

$$= p^{\sum_{i=1}^{n} x_i} (1-p)^{n-\sum_{i=1}^{n} x_i}$$

其中 $x_1, x_2, \cdots, x_n$ 在集合 $\{0,1\}$ 中取值.

例2 设总体 X 服从参数为 $\lambda(\lambda>0)$ 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

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

3. 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．比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.

总体(理论分布)？ ← 样本值 → 样本

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.

分散、复杂

是总体的代表，含有总体的信息

1.3 统计量

1. 统计量

由样本值去推断总体情况，需要对样本值进行“加工一个有效的方法就是构造一些样本的函数，通过样本函数把样本中所含的(某一方面)的信息集中起来.

样本的函数

这种不含任何未知参数、完全由样本决定的量称为统计量

定义 设 $X_1,X_2,\cdots,X_n$ 是来自总体 X 的容量为 n 的样本，若样本函数 $g(x_1,\cdots,x_n)$ 中不含任何未知参数，则称 $g(x_1,\cdots,x_n)$ 是一个统计量.

例2 设 X_1,X_2,X_3 是取自正态总体 $X\sim(\mu,\sigma^2)$ 的一个样本，其中 μ 已知，σ 未知，问下列样本函数中哪些是统计量，那些不是？

X_1 ✓　　X_2+1 ✓　　$(X_1+X_2+X_3)/3$ ✓

$\sum_{i=1}^{3}(X_i-\mu)^2$ ✓,　　$\max\{X_1,X_2,X_3\}$ ✓,　　$\sum_{i=1}^{3}\left(\frac{X_i}{\sigma}\right)^2$ ✗

我们主要研究两种基本的统计量：样本矩 和 顺序统计矩

1^0 样本矩 —— 几个常见的统计量

它反映了总体标准差的信息

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

反映了总体均值的信息

样本标准差 $S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$

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

它反映了总体方差的信息

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

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

样本 k 阶原点矩 $A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$

样本 k 阶中心矩 $B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k$　$k=1,2,\cdots$

并称他们相应的观测值 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，　$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$,

$\sqrt{s}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$,　$a_k=\frac{1}{n}\sum_{i=1}^{n}x_i^k$,　$b_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k$,　$k=1,2,\cdots$

仍分别为：样本均值、样本方差、样本标准差、样本 k 阶原点矩、样本 k 阶中心矩.

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

再根据辛钦定理知

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

由关于依概率收敛的序列的性质知

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

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.

2^0 顺序统计量

定义 设 $x_1, \cdots, x_n$ 是样本 $(X_1, \cdots, X_n)$ 的一组观测值，将 $x_1, \cdots, x_n$ 按递增顺序排列如下：$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ 记 $X_{(k)}$ 是如下的随机变量：当 $(X_1, \cdots, X_n)$ 取值 $x_1, \cdots, x_n$ 时，$X_{(k)} = x_{(k)}$，$k = 1, 2, \cdots, n$，并称 $(X_{(1)}, \cdots, X_{(n)})$ 为样本 $X_1, \cdots, X_n$ 的顺序统计量，$X_{(k)}$ $(k = 1, 2, \cdots, n)$ 称为第 k 位顺序统计量.

最大值？ 最小值？ 即按增序取值

例3 设 X_1, X_2, X_3 是总体 X 的一个样本，其三次观测值如下，则样本顺序统计量的观测值表为

观测值序号 \ 样本	X_1	X_2	X_3
1	1	0	1
2	3	1	0
3	1	2	0

观测值序号 \ 样本	X_1^*	X_2^*	X_3^*
1	0	1	1
2	0	1	3
3	0	1	2

样本极差的观测值 $r_1 = 1-0 = 1$;
$r_2 = 3$;
$r_3 = 2$.

称 $M = \begin{cases} X_{(\frac{n+1}{2})}, & n\text{为奇数}, \\ \dfrac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2}, & n\text{为偶数}, \end{cases}$ 为样本中位数，

称 $R = X_{(n)} - X_{(1)}$ 为样本极差.

设总体 X 的分布函数为 $F(x)$，$(X_1, \cdots, X_n)$ 为其样本，

$X_{(n)}$ 的分布函数记为 $F_{X_{(n)}}(x)$，则有

$$F_{X_{(n)}}(x) = P(X_{(n)} \le x) = [F(x)]^n$$

$X_{(1)}$ 的分布函数记为 $F_{X_{(1)}}(x)$，则有

$$F_{X_{(1)}}(v) = P(X_{(1)} \le x) = 1 - [1 - F(x)]^n$$

如果总体 X 有概率密度函数 $f(x)$，则 $X_{(n)}$ 及 $X_{(1)}$ 的概率密度函数分别为：

$$f_{X_{(n)}}(x) = nf(x)[F(x)]^{n-1}$$

$$f_{X_{(1)}}(x) = nf(x)[1 - F(x)]^{n-1}$$

$X_{(k)}$ 的概率密度函数和分布函数分别为：

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!}[F(x)]^{k-1}[1-F(x)]^{n-k} f(x),$$

$$F_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!}\int_0^{F(x)} u^{k-1}[1-u]^{n-k}dx.$$

例：设总体 X 服从区间 $(0, 1)$ 上的均匀分布，$X_1, X_2, \cdots, X_n$ 是来自 X 的一个样本，求第 k 个顺序统计量 $X_{(k)}$ 的概率密度函数 $f_{X_{(k)}}(x)$.

例：设总体 X 服从参数 λ 指数分布，$X_1, X_2, \cdots, X_n$ 是来自 X 的一个样本，求 $X_{(1)}, X_{(n)}$ 的概率密度函数.

1.4 由正态分布导出的抽样分布

一、χ^2 分布

1.定义 设 $X_1, \cdots, X_n$ 独立且都服从标准正态分布 $N(0,1)$，称随机变量

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

所服从的分布为自由度为 n 的 χ^2 分布，记为 $Y \sim \chi^2(n)$.

可以证明 χ^2 分布的密度为

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

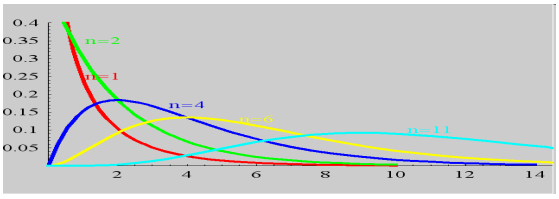
其中伽玛函数 $\Gamma(x)$ 通过积分 $\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$，$x > 0$ 定义的.

1^0 $\Gamma(x+1) = x\,\Gamma(x)$ $(x > 0)$；　2^0 $\Gamma(n+1) = n!$；　3^0 $\Gamma(1) = 1$.

4^0 $\Gamma(1/2) = \sqrt{\pi}$

χ^2 分布是由正态分布派生出来的一种分布

$\chi^2(n)$ 分布的概率密度曲线如图.



2. χ^2 分布的基本性质

1⁰ 设 $Y_1 \sim \chi^2(m)$, $Y_2 \sim \chi^2(n)$, 且 Y_1, Y_2 相互独立, 则

$$Y_1 + Y_2 \sim \chi^2(m+n);$$

χ^2 分布的可加性

2⁰ 若 $Y \sim \chi^2(n)$, 则 $EY = n$, $DY = 2n$.

3⁰ 设 $X_1, \cdots, X_n$ 相互独立, 且都服从正态分布 $N(\mu, \sigma^2)$, 则

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

4⁰ 若 $Y \sim \chi^2$ 分布, 则当 n 充分大时, $\frac{Y-n}{\sqrt{2n}}$ 近似服从 $N(0,1)$.

应用中心极限定理可得

定理(柯赫伦定理): 设$X_1, X_2, \cdots, X_n$相互独立同服从$N(0,1)$,

$$\sum_{i=1}^{n} X_i^2 = Q_1 + Q_2 + \cdots + Q_k,$$

其中$Q_i, i = 1,2,\cdots,k$是$X_1, X_2, \cdots, X_n$的秩为n_i的二次型, 则

$Q_i, i = 1,2,\cdots,k$相互独立, 且$Q_i \sim \chi^2(n_i)$的充要条件是

$$n_1 + n_2 + \cdots + n_k = n.$$

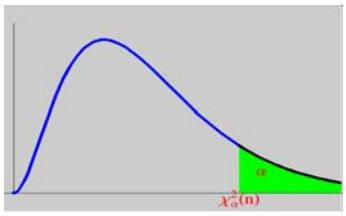
χ^2 分布的上侧分位数

对于给定的正数α, $0 < \alpha < 1$, 称满足条件

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

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

对于不同的α, n, 可以通过查表求得α 分位点的值.



例1 设 X 服从标准正态分布 $N(0,1)$, $N(0,1)$的上下侧α 分位数 z_α 满足 $P\{X \leq z_\alpha\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z_\alpha} \mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}x = \alpha$,

求 z_α 的值, 可通过查表完成.

$z_{0.95} = 1.645$,

$z_{0.975} = 1.96$,

根据正态分布的对称性知

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

例2设 $Z \sim \chi^2(n)$, $\chi^2(n)$的上侧α 分位数满足

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

求$\chi_\alpha^2(n)$的值, 可通过查表完成.

$\chi_{0.025}^2(8) = 17.535$,

$\chi_{0.975}^2(10) = 3.247$,

$\chi_{0.1}^2(25) = 34.382$.

附表只详列到 n=45 为止.

费舍尔(R.A.Fisher)证明:

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

其中 z_α 是标准正态分布的下侧α 分位数.

利用上面公式,

可以求得 $n > 45$时, 上侧α 分位数的近似值.

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 $X\sim\chi^2(15)$， 试确定 x 的值，使 $P(X\le x)=0.95$.

解 由题意知， $x=\chi^2_{0.05}(15)$

$n=15$， $\alpha=0.95$， 查附表知， $x=24.996$.

例 设 $X_1,\cdots,X_{10}$ 是来自正态总体 $X\sim N(0,0.3^2)$ 的一个样本，求 $P(\sum_{i=1}^{n}X_i^2>1.44)$.

解 $\because$ $X_i\sim N(0,0.3^2)$， $\therefore$ $\frac{X_i}{0.3}\sim N(0,1)$， $i=1,2,\cdots,10$.

又由于它们相互独立， $\therefore$ $\sum_{i=1}^{n}(\frac{X_i}{0.3})^2\sim\chi^2(10)$，

$$\Rightarrow P(\sum_{i=1}^{n}X_i^2>1.44)=P(\sum_{i=1}^{n}(\frac{X_i}{0.3})^2>\frac{1.44}{0.3^2})$$

$$=P(\chi^2(10)>16)=0.1.$$

二、 t 分布

定义2 设 $X\sim N(0,1)$, $Y\sim\chi^2(n)$， 且 X 与 Y 相互独立，称随机变量

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

所服从的分布为自由度为 n 的 t 分布，记为 $T\sim t(n)$.

可以证明 T 的密度函数为：

$$f(x)=\frac{\Gamma[(n+1)/2]}{\Gamma(n/2)\sqrt{n\pi}}(1+\frac{x^2}{n})^{-\frac{n+1}{2}}.$$

偶函数

T 分布的密度关于 y 轴对称，且 $\lim_{n\to\infty}f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$，这表明 n 充分大时，其密度的图形类似于标准正态分布的图形， 即 n 充分大时，t 分布近似 $N(0,1)$.

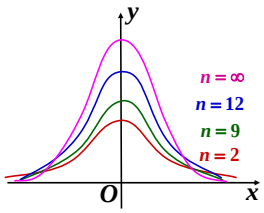
自由度为 n 的 t 分布的数学期望 和 方差为:

$E(T)=0$； $D(T)=n/(n-2)$， $(n>2)$.

t 分布的概率密度曲线如图.

显然图形是关于 $t=0$ 对称.

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.



t 分布的上侧分位数

对于给定的 α, $0<\alpha<1$, 称满足条件

$$P\{t\ge t_\alpha(n)\}=\int_{t_\alpha(n)}^{+\infty}f(t)\mathrm{d}t=\alpha$$

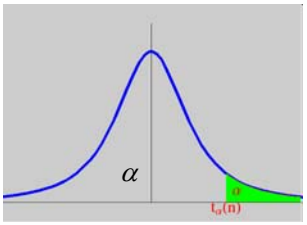
的点 $t_\alpha(n)$ 为 $t(n)$ 分布的上侧 α 分位数.

可以通过查表求得 α 分位数的值.

由分布的对称性知

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

当 $n>45$ 时， $t_\alpha(n)\approx z_{1-\alpha}$.



例 设 $T\sim t(n)$, $t(n)$ 的上侧 α 分位数满足

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

求 $t_\alpha(n)$ 的值，可通过查表完成.

$t_{0.05}(10)=1.8125$,

$t_{0.025}(15)=2.1315$.

例 设 $T\sim t(15)$， 试确定 x 的值，使 $P(|T|>x)=0.1$.

解 由 t 分布的偶性知 $P(T>x)=P(T<-x)$， $\therefore$ $P(T>x)=0.1/2$，

$n=15$， $\alpha=0.05$， 查附表知， $x=1.7531$.

例 设 X, Y_1, Y_2, Y_3, Y_4 相互独立，且 $X\sim N(2,1)$， $Y_i\sim N(0,4)$, $i=1,2,3,4$， 令 $Z=4(X-2)\Big/\sqrt{\sum_{i=1}^{4}Y_i^2}$， 求 Z 的分布.

解 $\because$ $X-2\sim N(0,1)$， $Y_i/2\sim N(0,1)$， $i=1,2,3,4$.

$$\because Z=\frac{4(X-2)}{\sqrt{\sum_{i=1}^{4}Y_i^2}}=\frac{X-2}{\sqrt{\sum_{i=1}^{4}(\frac{Y_i}{2})^2/4}}\sim t(4),$$

由 t 分布的定义

即 Z 服从自由度为 4 的 t 分布.

三、F 分布

定义3 设随机变量 X 与 Y 独立，且 $X\sim\chi^2(m)$, $Y\sim\chi^2(n)$, 则称统计量

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

所服从的分布为第一自由度为 m，第二自由度为 n 的 F 分布，记作 $F\sim F(m,n)$.

由 F 分布的定义可见，若 $X\sim F(m,n)$，则 X 的概率密度为

$$f(x)=\begin{cases}\dfrac{\Gamma\left(\frac{m+n}{2}\right)\left(\frac{m}{n}\right)^{\frac{m}{2}}x^{\frac{m}{2}-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)\left[1+\frac{mx}{n}\right]^{\frac{m+n}{2}}}, & x>0,\\ 0, & \text{其他.}\end{cases}$$

F 分布的数学期望 $E(X)=\frac{n}{n-2}$，$n>2$. 不依赖于第一自由度

F 分布的性质：

1° 若 $X\sim F(m,n)$，则 $1/X\sim F(n,m)$.

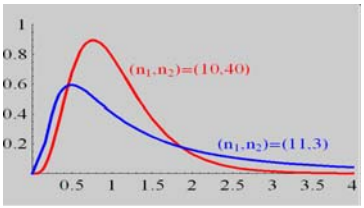
2° 若 $X\sim t(n)$，则 $X^2\sim F(1,n)$；$\frac{1}{F}=\frac{Y/n}{X/m}$

F 分布的概率密度曲线如图

根据定义可知，

若 $F\sim F(m,n)$,

则 $\frac{1}{F}\sim F(n,m)$.



F 分布的上侧分位数

对于给定的 α, $0<\alpha<1$, 称满足条件

$$P\{F\geq F_\alpha(m,n)\}=\int_{F_\alpha(m,n)}^{+\infty}f(y)\mathrm{d}y=\alpha$$

的点 $F_\alpha(n_1,n_2)$ 为 $F(m,n)$ 分布的长侧 α 分位数.

例 设 $F(m,n)$分布的 α分位数满足

$$P\{F\geq F_\alpha(m,n)\}=\int_{F_\alpha(m,n)}^{+\infty}f(y)\mathrm{d}y=\alpha,$$

求 $F_\alpha(m,n)$ 的值, 可通过查表完成.

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

$F_{0.05}(14,30)=2.04.$

F 分布的 α 分位数具有如下性质:

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

证明 因为 $F\sim F(m,n)$,

所以 $\alpha=P\{F>F_\alpha(m,n)\}$

$$=P\left\{\frac{1}{F}<\frac{1}{F_\alpha(m,n)}\right\}$$

因为 $\frac{1}{F}\sim F(n,m)$, 所以 $P\left\{\frac{1}{F}\leq F_{1-\alpha}(n,m)\right\}=\alpha,$

比较后得 $\frac{1}{F_\alpha(m,n)}=F_{1-\alpha}(n,m),$

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

用来求分布表中未列出的一些 α 分位数.

例 $F_{0.95}(12,9)=\frac{1}{F_{0.05}(9,12)}=\frac{1}{2.8}=0.357.$

例 设 $F\sim F(24,15)$，求 F_1,F_2,F_3，使其分别满足 $P(F>F_1)=0.025$，$P(F<F_2)=0.025$，$P(F>F_3)=0.95$.

解 (1) 由 $m=24, n=15, \alpha=0.025$，查附表知 $F_1=F_{0.025}(24,15)=2.70$;

(2) 无法直接查表获得，但 $P(F<F_2)=P(\frac{1}{F}>\frac{1}{F_2})=0.025$，由 F 分布性质知 $1/F\sim F(15,24)$，查附表知

$\therefore F_2=1/2.44=0.41$;

(3) 查表可知

$\therefore F_3=\frac{1}{2.11}=0.474.$

统计三大分布的定义和基本性质在后面的学习中常用到，要牢记！

四. 正态总体的抽样分布

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

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

定理四 设 $X_1, X_2, \cdots, X_m$ 与 $Y_1, Y_2, \cdots, Y_n$ 分别是两正态总体 $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ 的样本, 且这两个样本互相独立, 设 $\overline{X} = \dfrac{1}{m}\sum_{i=1}^{m} X_i$, $\overline{Y} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$ 分别是这两个样本的均值,

$S_1^2 = \dfrac{1}{m-1}\sum_{i=1}^{m}(X_i-\overline{X})^2$, $S_2^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2$ 分别是这两个样本的方差, 则有

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

(2) $$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{m}+\dfrac{\sigma_2^2}{n}}} \sim N(0,1).$$

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

$$\frac{S_1^2}{S_2^2} \sim F(m-1, n-1);$$

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}} \sim t(m+n-2),$$

其中 $S_w^2 = \dfrac{(m-1)S_1^2+(n-1)S_2^2}{m+n-2}$, $S_W = \sqrt{S_W^2}$.

证明 (1) 由定理二

$$\frac{(m-1)S_1^2}{\sigma_1^2}\sim\chi^2(m-1),\qquad \frac{(n-1)S_2^2}{\sigma_2^2}\sim\chi^2(n-1),$$

由假设 S_1^2, S_2^2 独立， 则由 F 分布的定义知

$$\frac{(m-1)S_1^2}{(m-1)\sigma_1^2}\Big/\frac{(n-1)S_2^2}{(n-1)\sigma_2^2}\sim F(m-1,n-1),$$

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

(2), (3) 因为 $\bar{X}-\bar{Y}\sim N\left(\mu_1-\mu_2,\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}\right)$

所以 $U=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}$ **~ N(0,1),**

当 $\sigma_1^2=\sigma_2^2=\sigma^2$

$$U=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}}\sim N(0,1),$$

由 $\frac{(m-1)S_1^2}{\sigma^2}\sim\chi^2(m-1),\quad \frac{(n-1)S_2^2}{\sigma^2}\sim\chi^2(n-1),$

且它们相互独立，故由 χ^2 分布的可加性知

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

由于 U 与 V 相互独立，按 t 分布的定义.

$$\frac{U}{\sqrt{V/(m+n-2)}}$$

$$=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\frac{1}{m}+\frac{1}{n}}}\sim t(m+n-2).$$

非正态总体的抽样分布定理

定理 ——样本均值、方差的分布 设 $X_1,\cdots,X_n$ 是取自均值为 μ，方差为 σ^2 的总体的一个样本，则当 n 充分大时，近似地有

(1) $\bar{X}\sim N(\mu,\frac{\sigma^2}{n})$; **(2)** $\frac{\bar{X}-\mu}{S/\sqrt{n}}\sim N(0,1)$.

独立同分布中心极限定理

独立同分布中心极限定理

1.5 经验分布函数和直方图

设 $x_1,x_2,\ldots,x_n$ 为总体X的一组观察值，将它们按有小到大的顺序排列，得到

$$x_{(1)}\le x_{(2)}\le\cdots\le x_{(n)}$$

称它为顺序统计量。则

$$F_n^*(x)=\begin{cases}0, & \text{当}x\le x_{(1)}\\ 1/n, & \text{当}x_{(1)}\le x<x_{(2)}\\ \cdots\\ k/n, & \text{当}x_{(k)}\le x<x_{(k+1)}\\ \cdots\\ 1, & \text{当}x\ge x_{(n)}\end{cases}$$

称它为总体X的经验分布。

1.5 经验分布函数和直方图

1 经验分布函数

设 $x_1,x_2,\ldots,x_n$ 是取自总体分布函数为 $F(x)$ 的样本，若将样本观测值由小到大进行排列，为 $x_{(1)},x_{(2)},\ldots,x_{(n)}$，则称 $x_{(1)},x_{(2)},\ldots,x_{(n)}$ 为有序样本，用有序样本定义如下函数

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

则$F_n(x)$是一非减右连续函数，且满足

$$F_n(-\infty)=0 \text{ 和 } F_n(+\infty)=1$$

由此可见，$F_n(x)$是一个分布函数，
并称$F_n(x)$为经验分布函数。

例 某食品厂生产听装饮料，现从生产线上随机抽取5听饮料，称得其净重（单位：克）

351 347 355 344 351

这是一个容量为5的样本，经排序可得有序样本：

$x_{(1)}=344,\ x_{(2)}=347,\ x_{(3)}=351,\ x_{(4)}=354,\ x_{(5)}=355$

其经验分布函数为

$$F_n(x)=\begin{cases}0, & x<344\\ 0.2, & 344\le x<347\\ 0.4, & 347\le x<351\\ 0.6, & 351\le x<354\\ 0.8, & 354\le x<355\\ 1, & x\ge 355\end{cases}$$

由伯努里大数定律：
只要 n 相当大，$F_n(x)$依概率收敛于$F(x)$。

更深刻的结果也是存在的，这就是格里纹科定理。

定理（格里纹科定理） 设$x_1,x_2,\ldots,x_n$是取自总体分布函数为$F(x)$的样本，$F_n(x)$是其经验分布函数，当$n\to\infty$时，有

$$P\{\sup|F_n(x)-F(x)|\to 0\}=1$$

格里纹科定理表明：当n相当大时，经验分布函数是总体分布函数$F(x)$的一个良好的近似。
经典的统计学中一切统计推断都以样本为依据，其理由就在于此。

2 直方图

直方图能反映总体密度函数的大致形状。

设X的分布函数和密度函数分布用$F(x), f(x)$表示：

$$f(x)=F'(x)=\lim_{\Delta x\to 0}\frac{F(x+\Delta x)-F(x)}{\Delta x}$$

$$=\lim_{\Delta x\to 0}\frac{P(x<X\le x+\Delta x)}{\Delta x}$$

当Δx很小时有 $f(x)\approx\dfrac{P(x<X\le x+\Delta x)}{\Delta x}$

设来自X的样本观测值为$x_1,x_2,\cdots,x_n$，则

$$f(x)\approx\frac{x_1,x_2,\cdots,x_n\text{这组值中属于区间}(x,x+\Delta x]\text{的频率}}{\Delta x}$$

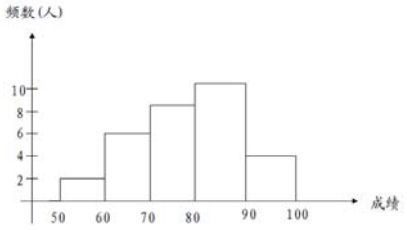
即：$f(x)$与单位长度的频率近似相等，称单位长度的频率为频率密度。

分布**直方图**是频数分布的图形表示，它的横坐标表示所关心变量的取值区间，纵坐标有三种表示方法：频数，频率，最准确的是频率/组距（频率密度），它可使得诸长条矩形面积和为1。凡此三种直方图的差别仅在于纵轴刻度的选择，直方图本身并无变化。

例：某班概率论考试频数分布表

按成绩分组（分）	频数（人）	频率（%）	频率密度（%）
50~60	2	6.7	0.67
60~70	6	20	2.0
70~80	8	26.7	2.67
80~90	10	33.3	3.33
90~100	4	13.3	1.33
合计	30	100	10

频数分布直方图



类似有：频率分布直方图，频率密度分布直方图

样本数据的整理是统计研究的基础，整理数据的最常用方法之一是给出其频数分布表或频率分布表。

例 参见书P17例1.4.2。某厂生产一种25瓦的白纸灯泡，其光通量(单位：流明)用X表示，从这批灯泡中抽取容量为60的样本，进行观察得到光通量数据。

对这60数据(样本)进行整理,具体步骤如下:

(1) 对样本进行分组：作为一般性的原则，组数通常在5~20个；（计算极差：224−193=31）

(2) 确定每组组距：近似公式为
组距d=(最大观测值 − 最小观测值)/组数;
(d=31/7=4.43)

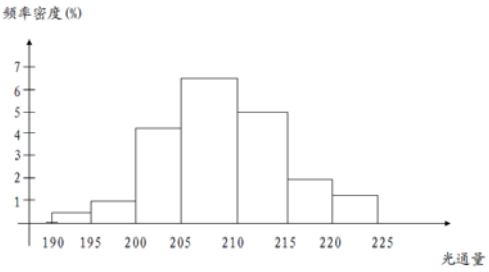
(3) 确定每组组限： 各组区间端点为
a_0, $a_1=a_0+d$, $a_2=a_0+2d$, …, $a_k=a_0+kd$,
形成如下的分组区间
$(a_0, a_1]$, $(a_1, a_2]$, …, $(a_{k-1}, a_k]$

其中a_0略小于最小观测值，a_k略大于最大观测值.
(190, 195], (195, 200], (200, 205], (205, 210], (210, 215], (215, 220], (220, 225],

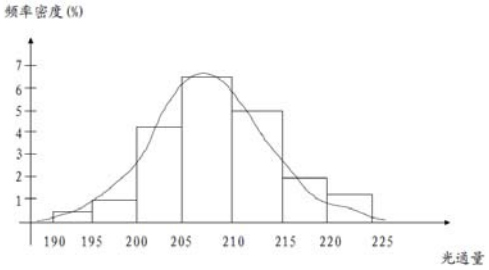
(4) 统计样本数据落入每个区间的个数——频数，并列出其频数，频率，频率密度分布表。

按光通量分组	频数(个)	频率密度(%)
(190,195]	2	0.67
(195,200]	3	1
(200,205]	12	4
(205,210]	19	6.3
(210,215]	14	4.7
(215,220]	6	2
(220,225]	4	1.3
合计	60	

频率密度分布直方图



三、频率密度分布曲线图



- 作业：P19
  4, 8, 14, 16, 19
  22, 26, 27, 30, 31, 33