目录

Statistics is itself a science—— the science of learning from data.

——From *Statistics: Challenges and Opportunities for the Twenty-First Century*

We are drowing in information and starving for knowledge.

——*Rutherford D. Roger*

第一章　基本概念

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.

数理统计不同于一般的资料统计，它更侧重于应用随机现象本身的规律性进行资料的收集、整理和分析. 由于大量随机现象必然呈现出它的规律性，因而从理论上讲，只要对随机现象进行足够多次观察，被研究的随机现象的规律性一定能清楚地呈现出来. 但客观情况只允许我们对随机现象进行次数不多的观察试验，也就是说, 我们获得的只是局部观察资料.

数理统计的任务就是研究怎样有效地收集、整理、分析所获得的有限的资料，对所研究的问题, 尽可能地作出精确而可靠的结论.

1.1　引言

例1.1. *Who Are Those Speedy Drivers?*

在Penn. State University 作了一个调查，被调查者要回答他们开车的最大速度？随机采访了87位男士和102位女士，得到数据如下：(单位：mph)

```
> male
 110 109  90 140 105 150 120 110 110  90 115  95 145 140 110 105 85  95 100
 115 124  95 100 125 140  85 120 115 105 125 102  85 120 110 120 115 94 125
  80  85 140 120  92 130 125 110  90 110 110  95  95 110 105  80 100 110 130
 105 120  90 100 105 100 120 100 100  80 100 120 105  60 125 120 100 115  95
 110 101  80 112 120 110 115 125  55  90 105
> female
  80  75  83  80 100 100  90  75  95  85  90  85  90  90 120  85  100 120 75
  85  80  70  85 110  85  75 105  95  75  70  90  70  82  85 100  90  75  90
 110  80  80 110 110  95  75 130  95 110 110  80  90 105  90 110  75  100 90
```

```
110  85  90  80  80  85  50  80  90 100  80  80  80  95 100  90  100  95 80
 80  50  88  90  90  85  70  90  30  85  85  87  85  90  85  75  90 102  80
100  80  95  90  80  95 110
```

从这些数据中我们能了解到什么呢？男士和女士开车最快速度有什么特点？

简单的数据总结得到

```
              male          Female
Min.    :     55.0           30.0
1st Qu.:      95.0           80.0
Median :     110.0           89.0
Mean    :    107.4           88.4
3rd Qu.:     120.0           95.0
Max.    :    150.0          130.0
```

显然, 有一半的男士开车的最快速度≥110, 有3/4 的人最快速度≥ 95, 而开车最快的速度为150, 最慢的速度为55. 对女士而言, 有一半的人开车的最快速度≥89, 有3/4的人的最快速度≥ 80, 而开车最快的速度为130,最慢的速度为30.

进一步，我们还以对这些数据的分布有如下了解

从这些分析我们可以认为男性开车速度数据是服从某个正态分布的。

例1.2. *在卢瑟福试验中,每隔一段时间观察一次由某种铀所放射的到达计数器的粒子数, 共观察100次, 得到结果如下:*

i	0	1	2	3	4	5	6	7	8	9	10	11	≥ 12
v_i	1	5	16	17	26	11	9	9	2	1	2	1	0

其中v_i表示观察到i个粒子的次数。由理论知识认为放射粒子数服从$Poisson$分布，试问是否真是这样？

例1和例2反映了统计的两个方面：**描述性统计**(Descriptive Statistics) 和**推断性统计**(Inferential Statistics)。

像例1那样对数据的特点(中心、方差、分位数、直方图等等)进行描述或者总结的方法，我们称为描述性统计。而像例2 那样，利用观察到的(部分)数据对总体作出某种

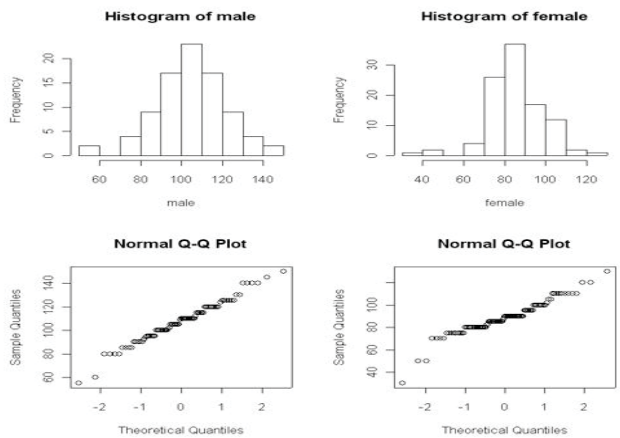


图 1.1: 直方图和正态Q-Q图

推断我们称为推断统计。概率论在推断统计中起着极其重要的作用。　因此，我们也可以这样定义数理统计学：

定义 1.1.1. 数理统计学是一门使用概率论和数学的方法，研究怎样有效地收集带有随机误差的数据，并在设定的模型下，对这种数据进行分析，以对所研究的问题作出推断的一门学科。

1.2　基本概念

现实世界中存在着形形色色的数据,分析这些数据需要多种多样的方法. 因此,数理统计中的方法和支持这些方法的相应理论是相当丰富的. 对推断性统计而言，主要有如下两大类:

- 参数估计——估计一些我们感兴趣的量. 表现为: 在概率模型(分布)假定下，根据数据用一些方法对分布的未知参数进行估计.

- 假设检验——对某种假设做出推断. 表现为: 根据数据,用一些方法对分布或分布

中的未知参数进行检验.

这两种推断渗透到了数理统计的每个分支.

在统计学里，有一些专门的术语来描述一个统计问题。我们来介绍一些常见的术语和一个问题的统计描述。

- 总体
- 样本
- 统计量

1.2.1 总体(Population), 样本(Sample)

在统计学中,**将我们研究的问题所涉及的对象的全体称为总体,而把总体中的每个成员称为个体**. 例如: 我们想要研究一家工厂的某种产品的废品率. 这种产品的全体就是我们的总体, 而每件产品则是个体.

因此直观上讲，总体就是所考察对象的全体。但是实际上, 我们真正关心的并不是总体或个体的**本身**, 而是其某项**数量指标**。比如例1, 我们要考察Penn. State University男士和女士开车的最快速度, 因此总体就是该校所有人。而我们真正关心的是该校每个人开车的最快速度这个数量指标。因此,我们应该把总体理解为那些研究对象上的某项数量指标的全体.

为了研究开车最快速度和性别之间的关系，通常的做法是从该校所有人中随机调查一些人，被调查人的全体就是一个样本。同上，我们实际是把样本理解为个体的数量指标. 因此从总体中抽出的一部分个体组成一个样本，总体包含个体的数目称为**总体容量**，样本包含个体的数目称为**样本容量**或者**样本大小(Sample size)**。

例1.3. *研究某地区N个农户的年收人. 在这里,总体既指这N个农户,又指我们关心的数量指标——他们的年收入这N个数字. 如果我们从这N个农户中随机地抽出n个农户作为调查对象, 那么, 这n个农户以及我们关心的数量指标——他们的年收入这n个数字就是样本.*

注意：在上面的例子中, 总体是很直观的, 是看得见摸得着的. 但是客观情况并不总是这样.

例1.4. 用一把尺子去量一个物体的长度. 假定n次测量值为$X_1, \cdots, X_n$。显然, 在这个问题中, 我们把测量值$X_1, \cdots, X_n$看成了样本, 但是, 总体是什么呢? 事实上, 这里没有一个现实存在的个体的集合可以作为我们的总体. 可是, 我们可以这样考虑, 既然n个测量值$X_1, \cdots, X_n$是样本, 那么总体就应该理解为一切所有可能的测量值的全体.

对一个总体,如果我们用X表示它的数量指标, 那么X的值对不同的个体取不同的值. 因此, 如果我们随机地抽取个体, 则X的值也就随着抽取的个体的不同而不同.

所以X是一个随机变量!

既然总体是随机变量X,自然就有其概率分布. 我们把X的分布称为总体的分布. 总体的特性是由总体分布来刻画的. 因此, 我们常把总体和总体分布视为同义语.

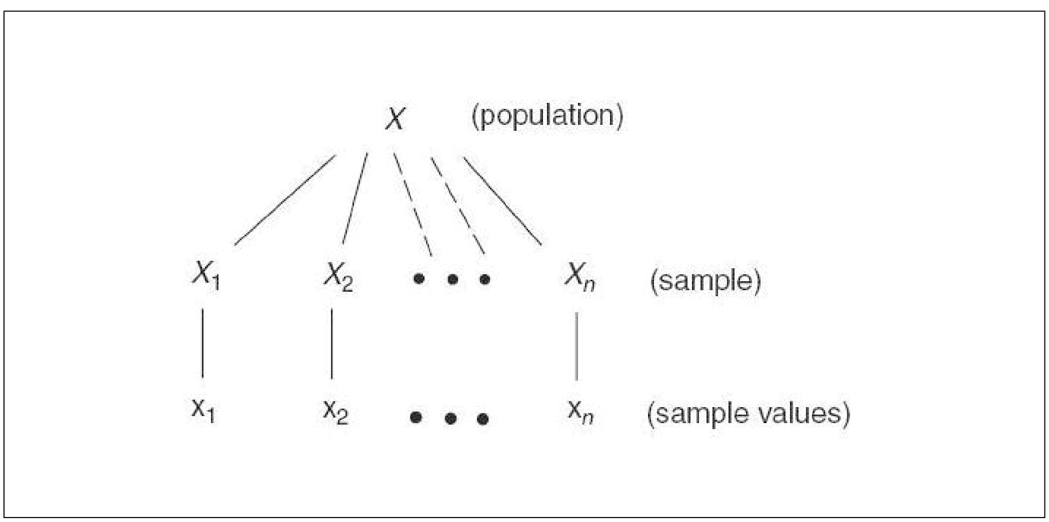
例1.5. 例*1.3*中，若农户年收入以万元计, 假定N户中收入X为以下几种取值:

$$0.5,\ 0.8,\ 1,\ 1.2 \text{和} 1.5.$$

取这些值的农户个数分别为：n_1, n_2, n_3, n_4, n_5,(这里$n_1 + n_2 + n_3 + n_4 + n_5 = N$). 则总体$X$的分布为离散型分布,其分布律为:

X	0.5	0.8	1	1.2	1.5
P	n_1/N	n_2/N	n_3/N	n_4/N	n_5/N

因此抽象地说，**总体是一个分布**。从总体中抽取一个个体就是做一次随机试验，而抽取样本容量为n的一个样本，就是做n次随机试验，记为$X_1, \cdots, X_n$。而试验得到的值$x_1, \cdots, x_n$则称为该样本的观察值。如下表所示:



- 如果总体所包含的个体数量是有限的, 则称该总体为有限总体. 有限总体的分布显然是离散型的,如例1.5.

- 如果总体所包含的个体数量是无限的,则称该总体为无限总体. 无限总体的分布可以是连续型的, 也可以是离散型的. 通常在总体所含个体数量比较大时, 我们就把它近似地视为无限总体, 并且用连续型分布去逼近总体的分布,这样便于做进一步的统计分析. 这种逼近所带来的误差,从应用观点来看,可以忽略不计.

当总体为某个确定的分布F时, 则也称该总体为F总体。比如总体分布为正态分布时, 则称为正态总体；而总体分布为指数分布时, 则称为指数总体等等。

- **样本的二重性**

1. 假设$X_1, X_2, \cdots, X_n$是从总体X中抽取的样本, 在一次具体的观测或试验中,它们是一批测量值, 是一些已得到的数. 这就是说, 样本具有数的属性.

2. 另一方面,由于在具体的试验或观测中,受到各种随机因素的影响,在不同的观测中样本取值可能不同. 因此,当脱离开特定的具体试验或观测时,我们并不知道样本$X_1, X_2, \cdots, X_n$的具体取值到底是多少, 因此,可以把它们看成随机变量.

样本$X_1, X_2, \cdots, X_n$既可被看成数又可被看成随机变量, 这就是所谓**样本的二重性**.

当试验是独立重复的进行时，则称样本$X_1, \cdots, X_n$为简单样本。即$X_1, \cdots, X_n$独立同分布。以后我们若无特殊说明，所说的样本都是指简单样本。

综上，我们给出如下定义

定义 1.2.2. *若用$r.v.X$表示所研究对象的某一指标，则总体即为$r.v.X$(的分布)。从此总体中抽取的n个随机变量$X_1, \cdots, X_n$称为样本，而样本$X_1, \cdots, X_n$的值$x_1, \cdots, x_n$称为样本的观察值。*

设总体X有概率函数(离散型即为分布律，连续场合下即为概率密度)$f(x)$，则在简单样本情形下，样本$X_1, \cdots, X_n$的联合分布为

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

1.2.2 统计量(Statistic)

只依赖于样本的量称为统计量。比如设$X_1, \cdots, X_n$为从总体$F_\theta(x)$中抽取的一个样本，其中θ为未知的参数，则$\sum_{i=1}^{n} X_i$为一个统计量，而$\sum_{i=1}^{n} X_i - \theta$ 就不是统计量。

统计量既然是依赖于样本的，而后者又是随机变量，故统计量也是随机变量，因而就有一定的分布，这个分布叫做统计量的“抽样分布”.

抽样分布就是通常的随机变量函数的分布. 只是强调这一分布是由一个统计量所产生的. 研究统计量的性质和评价一个统计推断的优良性，完全取决于其抽样分布的性质. 抽样分布有精确抽样分布（小样本问题中使用）和渐近分布（大样本问题中使用)。

统计量的作用在于集中有用的信息，降低数据的维数。

- 常见的统计量

以下我们设$X_1, \cdots, X_n$为样本。

1. 样本均值 $\bar{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$, 反映了总体均值的信息
2. 样本方差 $S^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i - \bar{X})^2$, 反映了总体方差的信息
3. 次序统计量$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$, 反映了总体分布的分位数信息

   3-1. 样本中位数

$$m = \begin{cases} X_{(\frac{n+1}{2})}, & n \textit{ is odd} \\ \frac{1}{2}[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}], & n \textit{ is even} \end{cases}$$

   3-2. 样本p $(0 < p < 1)$分位数$X_{[(n+1)p]}$, 此处$[a]$表示不超过a的最大整数.

   3-3. 样本极大值和样本极小值: $X_{(n)}$和$X_{(1)}$

   3-4. 极差: $X_{(n)} - X_{(1)}$

4. 样本k阶矩, 反映了总体k 阶矩的信息

   4-1. 样本k阶原点矩 $a_k = \frac{1}{n}\sum\limits_{i=1}^{n} X_i^k$

   4-2. 样本k阶中心矩 $m_k = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \bar{X})^k$

5. 经验分布函数

$$F_n(x) = \{X_1, \cdots, X_n\text{中} \le x\text{的个数}\}/n$$

例1.6. *公司用机器向瓶子里灌装液体洗净剂, 规定每瓶装m毫升. 但实际灌装量总有一定的波动. 假定灌装量的方差$\sigma^2 = 1$, 如果每箱装25瓶这样的洗净剂. 求: 这25瓶洗净剂的平均灌装量与标定值m相差不超过0.3毫升的概率是多少？又: 如果每箱装50瓶时呢?*

解: 记一箱中25瓶洗净剂灌装量为$X_1, X_2, \cdots, X_{25}$, 它们是来自均值为m, 方差为1的总体中的样本, 则由中心极限定理有

$$Z = \frac{\sqrt{n}(\bar{X} - m)}{\sigma} \text{ 近似服从 } N(0,1)$$

其中$n = 25$。因此这25瓶洗净剂的平均灌装量与标定值m相差不超过0.3毫升的概率近似是

$$P(|\bar{X} - m| \le 0.3) = P(|Z| \le 0.3\sqrt{n}/\sigma) \approx 2\Phi(1.5) - 1 = 0.86638$$

又当$n = 50$时，上述概率近似为0.966.

1.3 收集和加工有用的数据*

1.3.1 数据的有效性

使用数据进行推断的基本准则: 利用现有的数据能对一个较大的总体进行推断的前提是，可以认为这些数据在感兴趣的问题下能够代表这个总体。

即要求**好的试验设计** (以保证得到数据能够代表总体).

另一方面，当从这些数据从发进行统计推断时，我们还需要对数据的信息进行提炼。即需要构造合适的统计量，一个理想的统计量是完全包含了样本的信息，没有损失任何样本包含的有关参数的信息。换句话说，只要算出了这个统计量的值，就算把原来的样本都丢掉了，也没有任何损失。这种统计量我们称为**充分统计量**:

1.3.2 充分统计量

定义 1.3.3. 设$T(X)$为一统计量，$X = (X_1, \cdots, X_n)$为从总体$F_\theta(x)$里抽取的样本，θ为参数。如果

$$P(X_1 \le x_1, \cdots, X_n \le x_n | T(X) = t) \text{ 与参数 } \theta \text{ 无关}$$

则称$T(X)$为参数θ的一个充分统计量.

例1.7. *设一批产品有N件，其中次品有M件(M未知)，现从中随机抽取n件产品，其中恰好有m件次品。用X_i取1或0表示第i件产品是合格品还是次品($i = 1, \cdots, n$)。试证明无论是有放回抽样还是不放回抽样，$T = \sum_{i=1}^{n} X_i$ 是充分统计量。*

证明: (1) 有放回抽样情形: 记$p=(N-M)/N, q=M/N$，$x_i(i=1,\cdots,n)$只取0和1且$\sum\limits_{i=1}^{n} x_i=m$. 则因为

$$\begin{aligned}
P(X_1=x_1,\cdots,X_n=x_n|T=m) &= \frac{P(\{X_1=x_1,\cdots,X_n=x_n\}\{T=m\})}{P(T=m)} \\
&= \frac{P(X_1=x_1,\cdots,X_n=x_n)}{P(T=m)} \\
&= \frac{p^m q^{n-m}}{\binom{n}{m} p^m q^{n-m}} = 1\Big/\binom{n}{m} \text{ 与}M\text{无关}
\end{aligned}$$

(2) 不放回抽样情形

$$\begin{aligned}
&P(X_1=x_1,\cdots,X_n=x_n|T=m) = \frac{P(\{X_1=x_1,\cdots,X_n=x_n\}\{T=m\})}{P(T=m)} \\
&= \frac{P(X_1=x_1,\cdots,X_n=x_n)}{P(T=m)} \\
&= \frac{M(M-1)\cdots(M-m+1)(N-M)\cdots(N-M-n+m+1)}{N(N-1)\cdots(N-m+1)(N-m)\cdots(N-n+1)} \Big/ \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} \\
&= 1\Big/\binom{n}{m} \text{ 与}M\text{无关}
\end{aligned}$$

从而按充分统计量的定义知$T=\sum\limits_{i=1}^{n} X_i$为充分统计量。 □

充分性原则 在存在充分统计量的情形下，所有的统计推断都可以基于充分统计量进行.

定理 1.3.1. 设样本$X=(X_1,\cdots,X_n)$的概率函数$f_\theta(x_1,\cdots,x_n)$依赖于参数θ，$T=T(X)$为一统计量，则T为参数θ的充分统计量的充要条件为$f_\theta(x_1,\cdots,x_n)$可以分解为

$$f_\theta(x_1,\cdots,x_n)=g_\theta(T(x_1,\cdots,x_n))h(x_1,\cdots,x_n)$$

其中h仅与$x=(x_1,\cdots,x_n)$有关。

例1.8. 常见总体下的充分统计量(设样本为$X=(X_1,\cdots,X_n)$)

[1] 二项分布$B(n,p)$ 参数p的充分统计量为$\sum\limits_{i=1}^{n} X_i$

[2] 均匀分布$U(0,\theta)$ 参数θ的充分统计量为$X_{(n)}$

[3] 指数分布$Exp(\lambda)$ 参数λ的充分统计量为$\sum\limits_{i=1}^{n} X_i$

[4] 正态分布$N(\mu,\sigma^2)$ 参数(μ,σ^2)的充分统计量为$\bar{X},S^2$

1.3.3 对数据作预处理

在实际试验中，得到的观察数据若有某种相关性或者和假定的分布相差较远时，需要对数据进行预处理。比如X表示某种产品的性能指标，则X非负。若假定X服从正态，因为正态分布定义域为R而可能不太理想。因此我们若对X作log变换，则变换后的数据可能更切合正态分布假定。

1.4 统计三大分布

1.4.1 χ^2, t, F分布

在数理统计学里，有三个非常重要的分布:

1. χ^2 分布

定义 1.4.4. 设$X_1, \cdots, X_n$为相互独立且具有共同的分布*(i.i.d)* $N(0,1)$的随机变量，则称$X = \sum\limits_{i=1}^{n} X_i^2$的分布为自由度是$n$的$\chi^2$分布，记为$X \sim \chi_n^2$。

下面我们求解X的pdf:

•直接计算

作如下变换

$$
\begin{aligned}
&x_1 = rcos\theta_1 \\
&x_2 = rsin\theta_1 cos\theta_2 \\
&x_3 = rsin\theta_1 sin\theta_2 cos\theta_3 \\
&\ \vdots \\
&x_{n-1} = rsin\theta_1 \cdots sin\theta_{n-2} cos\theta_{n-1} \\
&x_n = rsin\theta_1 \cdots sin\theta_{n-2} sin\theta_{n-1} \\
&0 \le r < \infty,\ \ 0 < \theta_i \le \pi, i = 1, \cdots, n-2;\ \ 0 < \theta_{n-1} \le 2\pi
\end{aligned}
$$

则有

$$
\begin{aligned}
P(X \le x) &= P(\sum_{i=1}^{n} X_i \le x) = \int\!\!\cdots\!\!\int_{\sum x_i \le x} (2\pi)^{-n/2} exp[-\frac{1}{2}\sum_{i=1}^{n} x_i^2] dx_1 \cdots dx_n \\
&= c_n \int_0^{\sqrt{x}} r^{n-1} e^{-r^2/2} dr
\end{aligned}
$$

为求c_n，令$x \to \infty$，由Gamma函数易知$c_n = \frac{1}{2^{n/2-1}\Gamma(n/2)}$. 求导得到$X$的pdf

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2} I(x>0) \tag{1.4.1}$$

•归纳计算

参见课本P98例4.9.

> $X \sim \chi^2_n$的性质：
>
> 1. $EX = n, D(X) = 2n$
> 2. 关于参数n具有再生性，即若$Y \sim \chi^2_m$且与X相互独立，则$X+Y \sim \chi^2_{n+m}$.

2. t 分布 (Student t 分布)

定义 1.4.5. 设随机变量$X \sim N(0,1)$, $Y \sim \chi^2_n$且X和Y相互独立，令

$$T = \frac{X}{\sqrt{\frac{1}{n}Y}}$$

则称T的分布为自由度是n的t分布，记为$T \sim t_n$。

可以计算出T的概率密度为

$$g(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} I_{\{-\infty<t<\infty\}} \tag{1.4.2}$$

> t_n的性质
>
> 1. t分布关于$t=0$对称；
> 2. $\lim_{n\to\infty} g(t) = \phi(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$.

3. F 分布

定义 1.4.6. 设随机变量X, Y相互独立而且分别服从χ^2_n和χ^2_m，令

$$Z = \frac{1}{n}X \Big/ \frac{1}{m}Y$$

则称Z的分布为自由度是n和m(第一自由度是n，第二自由度是m)的F分布，记为$F \sim F(n,m)$.

同样，可以得到$F(n,m)$具有概率密度：

$$p(x) = \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} m^{m/2} n^{n/2} x^{m/2-1} (n+mx)^{-\frac{n+m}{2}} I_{0<t<\infty} \tag{1.4.3}$$

分位数

定义 1.4.7. 设随机变量X的分布函数为F，$0<\alpha<1$，称数x_α为随机变量X的(上)α分位数，如果

$$1 - F(x_\alpha) = P(X \geq x_\alpha) = \alpha$$

> 记$F(n,m)$的上α分位数为$F_\alpha(n,m)$，则有$F_\alpha(n,m) = F^{-1}_{1-\alpha}(m,n)$。

对标准正态分布，χ^2和t分布，其上α分位数分别记为$u_\alpha, \chi^2_n(\alpha), t_n(\alpha)$。

1.4.2 正态总体下$\bar{X}$与S^2的分布

定理 1.4.2. 设$X_1,\cdots,X_n$为从正态总体$N(\mu,\sigma^2)$中抽取的简单样本，$\bar{X}$与S^2分别为样本均值和样本方差，则

$$\begin{aligned}
&(1) \quad \bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \\
&(2) \quad (n-1)S^2 \Big/ \sigma^2 \sim \chi^2_{n-1} \\
&(3) \quad \bar{X} \quad \text{与} \quad S^2 \quad \text{相互独立} \\
&(4) \quad \sqrt{n}(\bar{X}-\mu) \Big/ S \sim t_{n-1}
\end{aligned}$$

证明需要多元正态分布的性质[注1]：

$$Let \quad X \sim N(\theta, \Sigma), Q > 0, \quad then \quad QX \sim N(Q\theta, Q\Sigma Q')$$

证明: 由如上多元正态的性质，我们记$X = (X_1,\cdots,X_n)', \theta = (\mu,\cdots,\mu)'$，则$X \sim$

[注1]感兴趣的可以参看南开大学杨振明的《概率论》中的多元正态分布一节

$N(\theta, \sigma^2 I_n)$。又令Q为如下正交阵

$$Q^{[注2]} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ q_{21} & q_{22} & \cdots & q_{2n} \\ & & \vdots & \\ q_{n1} & q_{n2} & \cdots & q_{nn} \end{pmatrix}$$

从而有$QX \sim N(Q\theta, \sigma^2 I_n)$, 令$Y = (Y_1, \cdots, Y_n)' = QX$，显然$Y_1, \cdots, Y_n$相互独立且

$$Y_1 \sim N(\sqrt{n}\mu, \sigma^2)$$
$$Y_i \sim N(0, \sigma^2), i = 2, \cdots, n$$

又由

$$Y_1 = \sqrt{n}\bar{X}$$
$$(n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X} = \sum_{i=2}^{n} Y_i^2$$

因此得到(3)，再由正态分布、χ^2和t分布的性质定义易知其他结论成立。 □

进而可以得到

定理 1.4.3. 设$X_1, \cdots, X_n$为从正态总体$N(\mu_1, \sigma_1^2)$中抽取的样本，$Y_1, \cdots, Y_m$为从正态总体$N(\mu_2, \sigma_2^2)$中抽取的样本, 而且两组样本独立，用$\bar{X}, S_X^2, \bar{Y}, S_Y^2$分别表示两组样本的样本均值和样本方差，则

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

(2) $\frac{S_X^2}{S_Y^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F(n-1, m-1)$

(3) 当$\sigma_1^2 = \sigma_2^2$, 有

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{nm}{(n+m)(n+m-2)}[(n-1)S_X^2 + (m-1)S_Y^2]}} \sim t_{n+m-2}$$

证明: 证明由定理1.4.2，以及正态分布、t分布及F分布的定义立得。 □

[注2]这样的正交矩阵是存在的,比如Q可以取为

$$\begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{1\cdot 2}} & -\frac{1}{\sqrt{1\cdot 2}} & 0 & \cdots & 0 \\ & & \vdots & & \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & \cdots & -\frac{n-1}{\sqrt{n(n-1)}} \end{pmatrix}$$

例1.9. 假设某物体的实际重量为m,但它是未知的.现在用一架天平去称它,共称了n次,得到$X_1, X_2, \cdots, X_n$. 假设每次称量过程彼此独立且没有系统误差, 则可以认为这些测量值都服从正态分布$N(m, \sigma^2)$, 方差σ^2反映了天平及测量过程的总精度. 考虑样本均值与真值m的偏差。

解: 由前面的定理知

$$\frac{\sqrt{n}(\bar{X}-m)}{\sigma} \sim N(0,1)$$

因此，在给定精度下，样本均值和真值m的偏差的界随着称量次数的增加而减小。即若设α为给定的称量精度，则

$$P(|\bar{X}-m| \le x) = P\left(\left|\frac{\sqrt{n}(\bar{X}-m)}{\sigma}\right| \le \frac{\sqrt{n}x}{\sigma}\right) = 2\Phi(\sqrt{n}x/\sigma) - 1 = \alpha$$

由此在已知σ, α时可以解出x的值。比如$\sigma = 0.1, \alpha = 0.97$时，若$n = 10$，则可以得到$x = 0.0686$. 若$n = 100$, 则$x = 0.0217$.

1.5 总结

数据在使用前要注意其收集的合法性(主要的是设计好的试验，感兴趣可以参看参考文献[3])。在合法的数据下，才能展开统计推断工作。

在给定统计模型(根据理论知识或其他知识来源确定)假设的前提下，一个统计推断问题可以按照如下的步骤进行：

1. 寻求用于统计推断的统计量(在存在充分统计量的情形下使用充分统计量)；
2. 统计量的分布；
3. 基于该统计量和统计推断方法作出推断；
4. 根据统计推断结果对问题作出解释。

统计三大分布及正态总体下样本均值和样本方差的分布，在我们后面的学习中占着重要的地位和应用。

学习统计无须把过多时间化在计算上，可以更有效地把时间用在基本概念、方法原理的正确理解上. 国内外著名的统计软件包：SAS，SPSS，MATLAB, STAT等，都可以让你快速、简便地进行数据处理和分析.

参考文献

[1] 陈希孺, 概率论与数理统计. 合肥: 中国科学技术大学出版社, 1995.

[2] 杨振明, 概率论, 南开大学数学教学丛书. 北京: 科学出版社, 2001.

[3] T.T. Soong, Fundamentals Of Probability And Statistics For Engineers, New York: John Wile & Sons, 2004.

[4] 王万中, 茆诗松, 试验的设计与分析. 上海: 华东师范大学出版社, 1997.