第一讲　抽样分布

秦中峰
qinz@vip.163.com

1.0 概率论与数理统计

- 概率论：研究随机现象
  - 随机变量及其概率分布全面地描述了随机现象的统计规律性
  - 在概率论中，通常假定概率分布是已知的，一切计算及其推理均基于这个已知的分布进行
- 然而，当我们研究并解决实际问题时，情况往往并非如此

1.0 概率论与数理统计

- 例题：
  - 某公司要采购一批产品，每件产品不是合格品就是不合格品，但该批产品总有一个不合格品率p
  - 若从该批产品中随机抽取一件，用X表示这一件产品的不合格数，则X服从两点分布b(1,p)，但p未知
  - p的大小决定了该批产品的质量，直接影响采购行为的经济效益。
- 因此，人们会对p提出一些问题，比如：
  - **p的大小如何?**
  - **P大概落在什么范围内?**
  - **能否认为p满足设定要求（如不超过0.05）？**

1.0 概率论与数理统计

- 数理统计的研究内容：
  - 研究如何更合理、更有效地抽取样本，从而获得观测数据和资料的方法
  - 如何利用一定的数据资料，对所关心的问题，得出尽可能精确且可靠的统计结论
- 然而，当我们研究并解决实际问题时，会立即遇到一些问题：
  - 1. 这个随机现象可以用什么样的分布律来刻画？这种分布律的选用合理吗？
  - 2. 所选用的这一分布律的参数是多少？如何估计和确定这些参数？

1.1 统计推断的意义和问题

- 1. 总体与个体
  - 总体：研究对象的全体
  - 个体：总体中的每个成员
- 如何理解“总体就是一个分布”？
  - 个体可以用数量表示。为简单：把个体看成数量，把总体看成数量的集合
  - **例如，*X*是一个正态总体：** $X \sim N(\mu,\sigma^2)$
- 例如：调查某公司 500 位职工的工资收入
  - （1）总体：500名职工的收入集合
  - （2）个体：每一个职工的工资收入

1.1 统计推断的意义和问题

- 总体均值：常用μ表示
- 设总体含有N个个体，第*i*个个体用　表示
- 总体均值 $\mu = \dfrac{x_1 + x_2 + \cdots + x_N}{N}$
- 总体方差 $\sigma^2 = \dfrac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_N-\bar{x})^2}{N}$
- 总体标准差 $\sigma = \sqrt{\sigma^2}$
- 总体参数：描述总体的特征，要调查的指标

1.1 统计推断的意义和问题

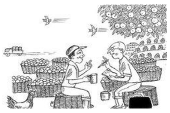
- **例题：**磁带的一个质量指标是一卷磁带（20m）上的伤痕数。每卷磁带都有一个伤痕数，全部磁带的伤痕数构成一个总体。该总体中相当一部分是0（无伤痕），但也有1、2、3等，但多于8个的伤痕数非常少见。
- 研究表明：一卷磁带上的伤痕数 X 服从泊松分布P（λ），但分布的参数 λ 却未知。显然，λ 的大小决定了一批产品的质量，直接影响生产方的经济效益。
- 在本例：总体分布的类型是明确的，但总体含有未知参数λ，因此总体不是一个特定的泊松分布。
- **统计的任务：确定λ，即确定最终的总体分布**

1.1 统计推断的意义和问题

- **例题：**考察常见的测量问题。
- 对物理量μ进行重复测量，此时一切可能的测量结果是实数集R，因此总体是一个取值于R的随机变量X
- 测量结果X可以看作物理量μ与测量误差ε的叠加，即 X = μ+ε。这里μ是一个确定但未知的量，称为参数。
  - （1）假定随机误差 ε~N(0，σ2），则X~N(μ，σ2）。如何**确定两个未知参数**是统计学要研究的问题。
  - （2）若没有理由认定误差服从正态分布，但可以认为误差的分布是关于0对称的，则总体分布就变为一个分布类型未知但带有某种限制的分布，**确定分布类型**是非参数统计

1.1 统计推断的意义和问题

问题：为什么要抽样?



普查(Census）的代价：

1. 费用昂贵
2. 时间过长
3. 观测值几乎是无穷个
4. 毁坏性实验
5. 精度：

由一个训练有素的调查人员得到的样本统计结果，可能比没有受过训练的人进行普查得到的结果更准确.

最关心的问题：

抽样调查结论的准确性？可靠性？

2. 随机样本与样本容量

可以从抽样框中抽取一部分个体进行观测统计，再根据这部分个体的观测信息推断总体的性质。

（1）**一个样本（Sample）：**

$$X \leftarrow (X_1, X_2, \cdots, X_n)$$

注意：由于X_i是从总体中随机抽取的，所以 $X_1, X_2, \ldots, X_n$ 是 n 个随机变量。

(2) 样本容量（Sample Size）：n

大样本：$n \geq 30$

小样本：$n < 30$

（3）**样本值：**一次实际抽取 $(x_1, x_2, \ldots, x_n)$

2. 随机样本与样本容量

Sample vs. Sampling

样本的二重性

（1）一方面，由于样本是从总体中随机抽取的，抽取前无法预知它们的数值。因此，样本是随机变量，用大写字母表示

（2）另一方面，样本在抽取后，经观测就有确定的观测值。因此，样本又是一组数值，此时用小写字母表示

3. 简单随机样本——“独立同分布”

independent and identically distributed **(i.i.d)**

(1) 随机性：

每一个随机变量 X_i 与总体 X **同分布**

(2) 独立性：

每一样品的取值不影响其他样品的取值

($X_1, X_2, \ldots, X_n$ 是相互独立且同分布的随机变量)

例：9个白球，1个黑球。抽出两个球：(X_1, X_2)

放回抽样 $P(X_1=\text{白})=\frac{9}{10}$，$P(X_2=\text{白}|X_1=\text{白})=\frac{9}{10}$

不放回抽样 $P(X_1=\text{白})=\frac{9}{10}$，$P(X_2=\text{白}|X_1=\text{白})=\frac{8}{9}$

一种近似情景： 总体的观测数量远远大于样本容量

由概率论知，

联合概率分布：

若总体 X 是离散型随机变量，

则 i.i.d 样本（$X_1, X_2, \dots, X_n$）具有**联合概率分布：**

$$P(X_1=x_1, X_2=x_2, \cdots, X_n=x_n) = P(X_1=x_1)P(X_2=x_2)\cdots P(X_n=x_n)$$

联合密度函数：

若总体 X 具有概率密度 $f(x)$，

则i.i.d 样本（$X_1, X_2, \dots, X_n$）具有**联合密度函数：**

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

---

4. 统计量（估计量）

用样本构造一个不含有任何未知参数的函数，用于推断总体参数。 $\hat{\theta} = g(X_1, \cdots, X_n)$

为什么要构造统计量？

样本来自总体，样本的观测值中含有总体各方面的信息，但这些信息较为分散，有时显得杂乱无章。将这些分散在样本中的有关总体的信息集中起来以反映总体的各种特征，需要对样本进行加工。

表和图是一类加工形式，人们从中获得对总体的初步认识；当需要从样本获得对总体各种参数的认识时，最常用的加工方法是构造样本函数，不同的函数反映总体的不同特征

---

4. 统计量（估计量）

尽管统计量不依赖于未知参数，但是它的分布一般是依赖于未知参数的。

例： $X \sim N(\mu, \sigma^2) \leftarrow (X_1, X_2, \cdots, X_n)$

μ ，σ未知。

则：$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ，$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$是统计量。

而 $\dfrac{n\sigma}{\sum_{i=1}^{n} X_i}$不是统计量

统计量的分布：抽样分布

常见的统计量：

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

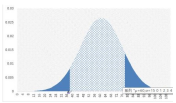
2. 样本方差 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

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

1.3 统计中常用的随机变量及其分布

1. 正态分布（Normal Distribution）

$X \sim N(\mu, \sigma^2)$



(1) $P\{a \le X \le b\}$ = 密度曲线下面的面积

(可以查表)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

(2) 标准正态分布： $Z \sim N(0,1)$

(3) $X \sim N(\mu, \sigma^2)$

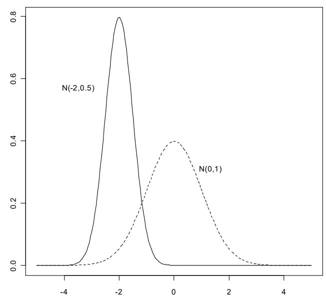
$E(X) = \mu$

$Var(X) = \sigma^2$

$Z = \dfrac{X-\mu}{\sigma} \sim N(0,1)$ （Z-Score）

正态分布的形态

例：左边是 $N(-2, 0.5)$ 分布，右边是 $N(0, 1)$ 分布



$E(X) = \mu$

$Var(X) = \sigma^2$

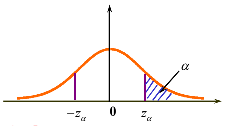
均值 μ 决定了正态分布的中心位置.

方差 σ^2决定了分布离散的程度.

此外,设$X\sim N(0,1)$,若Z_α满足条件$P\{X>Z_\alpha\}=\alpha, 0<\alpha<1$

则称点Z_α为标准正态分布的上α分位数。

$$Z_{1-\alpha}=-Z_\alpha$$

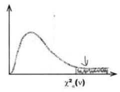


标准正态分布两个最常用的分位点:

$P\{|Z|\ge z_{\alpha/2}\}=\alpha,\quad \alpha=0.05,\quad z_{\alpha/2}=1.96$

$P\{|Z|\ge z_{\alpha/2}\}=\alpha,\quad \alpha=0.10,\quad z_{\alpha/2}=1.645$

2、χ^2分布



定义:

(1) 设 $X\sim N(0,1)$, 则 $X^2\sim\chi^2(1)$

可加性:

(2) 设 $X_j\sim N(0,1),\ j=1,2,\cdots,n\quad (i.i.d)$

$Z=X_1^2+X_2^2+\cdots+X_n^2$

则: $Z\sim\chi^2(n)$

(3) 设 $Z_1\sim\chi^2(n_1)$, $Z_2\sim\chi^2(n_2)$ 相互独立

则: $Z_1+Z_2\sim\chi^2(n_1+n_2)$

χ^2分布

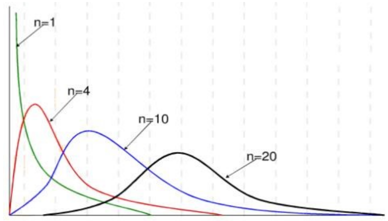
χ^2(n)的概率密度为

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

其中,

$$\Gamma(\alpha)=\int_0^{+\infty}x^{\alpha-1}e^{-x}dx$$

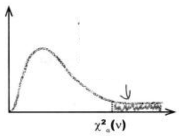
若$\chi^2\sim\chi^2$(n), 则有$\mathbf{E(\chi^2)=n,\ D(\chi^2)=2n.}$

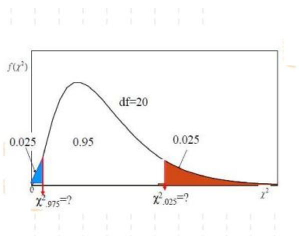


例题: $\chi^2\sim\chi^2(25)$

求: $P(\chi^2>x)=0.05$

$\chi^2_{0.05}(25)=37.652$





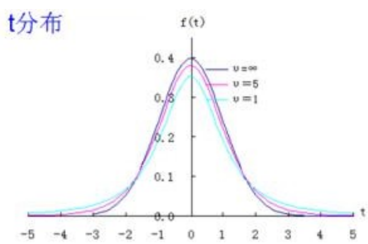
例题: $\chi^2\sim\chi^2(20)$

求: $P(\chi^2>x_{\alpha/2})=0.025$ $\quad P(\chi^2<x_{\alpha/2})=0.025$

$\chi^2_{0.025}(20)=34.170$ $\quad \chi^2_{0.975}(20)=9.591$

3、t-分布:

定义: 设$X\sim N(0,1)$, $Y\sim\chi^2(n)$, 并且X, Y相互独立,则称 $t=X/\sqrt{Y/n}$服从自由度为 n 的 t-分布,记作 t~t(n).



3、t-分布:

定义: 设$X \sim N(0,1)$, $Y \sim \chi^2(n)$, 并且X, Y相互独立,则称 $\boxed{t = X/\sqrt{Y/n}}$ 服从自由度为 n 的 t – 分布 ,记作 $t \sim t(n)$.

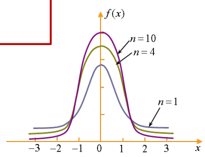
2. t(n)分布的概率密度函数

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

$$\Gamma(\alpha)=\int_0^{+\infty} x^{\alpha-1}e^{-x}dx$$

E(t) =0

D(t) = *n*/*n*-2



t分布的密度函数

例题:

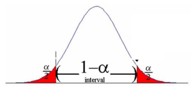
考虑一个自由度等于 9 的 t -分布（ n=9）

(1) 求 $P(T > t_{0.05}(9)) = 0.05$ **上α分位数**

解: $t_{0.05}(9) = 1.833$

(2) 求 $P(|T| > t_{0.025}(9)) = 0.05$

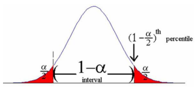
解: $t_{0.025}(9) = 2.262$



例题:

(3) 求 $P(|T| > t_{0.025}(20)) = 0.05$

解: $t_{0.025}(20) = 2.086$



(2) 求

解: $P(T > t_{0.05}(70)) = 0.05$

$t_{0.05}(70) = 1.96$

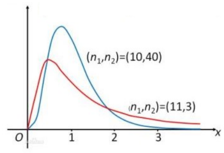
4、F-分布

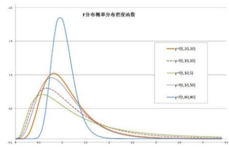
定义: 设$U \sim \chi^2(n_1)$, $V \sim \chi^2(n_2)$, 且U, V独立,

则称: $F = \dfrac{U/n_1}{V/n_2}$ 服从自由度为 (n_1, n_2)

的 F分布,

记作: $\boxed{F = \dfrac{U/n_1}{V/n_2} \sim F(n_1, n_2)}$





F-分布

$F(n_1, n_2)$分布的概率密度函数 :

$$\Gamma(\alpha)=\int_0^{+\infty} x^{\alpha-1}e^{-x}dx$$

$$\psi(u)=\begin{cases}\dfrac{\Gamma[(n_1+n_2)/2](n_1/n_2)^{n_1/2}}{\Gamma(n_1/2)\Gamma(n_2/2)}\left(1+\dfrac{n_1}{n_2}u\right)^{-(n_1+n_2)/2} u^{n_1/2-1}, & u \ge 0\\ 0 & \text{其它.}\end{cases}$$

$n_2 > 2 \quad E(u) = \dfrac{n_2}{n_2 - 1}$

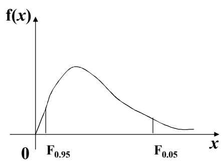
$n_2 > 4 \quad D(u) = \dfrac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)(n_2-4)}$

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

$n_1 = 8, \quad n_2 = 4$

$F_{0.05}(8,4) = 6.04$

$$F_{0.95}(8,4) = \frac{1}{F_{0.05}(4,8)} = \frac{1}{3.84} = 0.26$$



总结：四大分布之间的亲缘关系

X_1 与 X_2 相互独立

(1) $X_1 \sim N(\mu,\sigma^2),\ X_2 \sim N(\mu,\sigma^2) \to X_1+X_2 \to N(2\mu,2\sigma^2)$

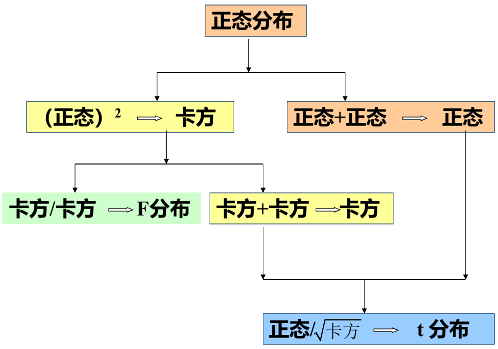
(2) $X \sim N(0,1) \ \to \ X^2 \sim \chi^2(1)$

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

(4) $X_1 \sim N(0,1),\ X_2 \sim \chi^2(n) \to \dfrac{X_1}{\sqrt{X_2/n}} \sim t(n)$

(5) $X_1 \sim \chi^2(n_1), X_2 \sim \chi^2(n_2) \ \to \ \dfrac{X_1/n_1}{X_2/n_2} \sim F(n_1,n_2)$

正态分布族谱：



本源问题：人们起意研究这些分布的原始动机？

5、二项分布 Binomial Distribution

贝努力试验:

Bernoulli Trials: $X = \begin{cases} 1 & p \\ 0 & (1-p) \end{cases}$

二项分布：在 n 次独立的贝努力试验中成功的次数 k

$$P(X=k) = C_n^k p^k (1-p)^{n-k} \quad ,k = 0,1,2,\cdots,n$$

例： 在一个男女各占一半的居民区中，随机抽取一个容量为 10的样本。问样本中正好有4 位女性居民的概率是多少？

$n = 10,\quad p = 0.5,\quad k = 4$

$$P(X=4) = C_{10}^4 0.5^4 (1-0.5)^{10-4} = \frac{10\times9\times8\times7}{4\times3\times2\times1} \times 0.5^{10} = 0.2051$$

例：一个多重测试选择题由10个问题组成。每一个问题有5个可供选择的答案。至少答对5道才能及格。如果完全靠猜测，及格的概率是多大？

n=10, $p = 1/5$, $k = 5,6,7,8,9,10$

二项分布的数学期望值与方差

问题：手上有一枚均匀硬币，连续抛掷100次，有多少次正面朝上？

（1） $E(X) = \sum_{k=0}^{n} k C_n^k p^k (1-p)^{n-k} = np$

（2） $D(X) = \sum_{k=0}^{n} (k-np)^2 C_n^k p^k (1-p)^{n-k} = np(1-p)$

1.4 常用统计量及其分布

1. $X \sim N(\mu,\sigma^2) \leftarrow (X_1,X_2,\cdots,X_n)$: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

则： $\overline{X} \sim N(\mu,\sigma^2/n)$ （放回抽样）

证明： 因为 $X_1, X_2, \ldots, X_n$ 服从正态分布，所以

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

$$\overline{\mu} = E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

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

问题：为什么有 $D(\overline{X}) = \sigma^2/n$

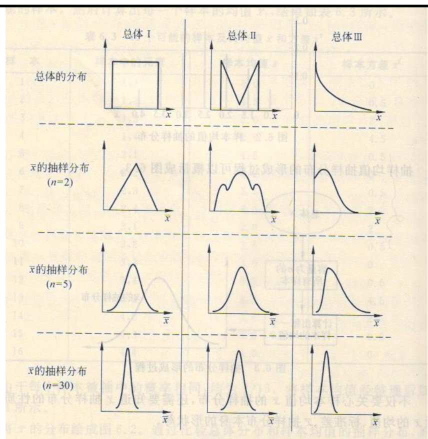


图 6.4 $\overline{x}$ 的抽样分布趋于正态分布的过程

在概率论中已经证明：

$$X \sim N(\mu,\sigma^2)$$

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

2、在统计问题中：

$$X \sim N(\mu,\sigma^2) \leftarrow (X_1, X_2, \cdots, X_n): \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

则： $\bar{X} \sim N(\mu, \sigma^2/n)$

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

$$X \sim N(\mu,\sigma^2) \leftarrow (X_1, X_2, \cdots, X_n): \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

则： $\bar{X} \sim N(\mu, \sigma^2/n)$

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

σ 未知的情形怎么办?

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

问题：在什么情况下，有 $Z = \dfrac{\bar{X}-\mu}{s/\sqrt{n}} \sim N(0,1)$ ？？？

2、中心极限定理的应用

在服从**任意**分布的总体中，抽取容量为 n 的样本（i.i.d）。总体均值为 μ，标准差为 σ，如果 $n \to \infty$

$$\bar{X} \to N(\mu, \sigma^2/n)$$

在应用中， 当 $n \geq 30$ 可以认为是大样本

$s \approx \sigma$ $\qquad Z = \dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim N(0,1)$

3、t—分布的应用

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

问题：小样本 $Z = \dfrac{\bar{X}-\mu}{s/\sqrt{n}} \sim N(0,1)$ ？？？

$$X \sim N(\mu,\sigma^2) \leftarrow (X_1, X_2, \cdots, X_n): \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

则： $t = \dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

---

t—分布 （Student's Distribution，W.S. Gosset,1876_1937）

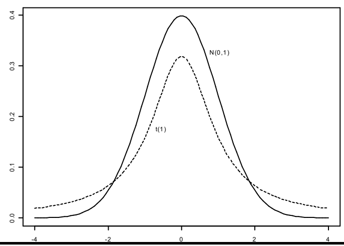
(1) 关于 **0** 对称，取值范围在 $-\infty$ to $+\infty$

(2) 钟形对称

（3） 当 $n \to \infty$, $T \sim N(0,1)$

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

标准正态分布和 t(1) 分布的密度图



都柏林Arthur Guinness and Son酿酒公司的酿酒师。发现从原材料到酵母发酵，每一个过程都影响啤酒产品的质量。

1908年发表论文“均值的可能误差”：**当样本容量很小时， s 是σ的一个不稳定的估计**。因此，**在小样本以及σ未知的情况下，应用 Z 统计量估计精度是失效的**。因此提出一种新的抽样分布，即学生 t 分布，并由此引入了小样本估计理论方法。

4、χ^2 分布的应用

$$X \sim N(\mu,\sigma^2) \leftarrow (X_1, X_2, \cdots, X_n):$$

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

则： $\chi^2 = \dfrac{S^2}{\sigma^2/(n-1)} \sim \chi^2(n-1)$

自由度：df = (*n*-1)

因为 $(X_1-\bar{X})+(X_2-\bar{X})+\cdots+(X_n-\bar{X})=0$

所以 $(X_1-\bar{X}),(X_2-\bar{X}),\cdots,(X_n-\bar{X})$ 自由取值的个数为（***n*-1**）.

5、F-分布的应用

$$X \sim N(\mu_X, \sigma_X^2) \leftarrow (X_1, X_2, \cdots, X_{n_1}):\quad s_X^2$$

$$Y \sim N(\mu_Y, \sigma_Y^2) \leftarrow (Y_1, Y_2, \cdots, Y_{n_2}):\quad s_Y^2$$

且：X与Y独立

则：$$F = \frac{s_X^2 / \sigma_X^2}{s_Y^2 / \sigma_Y^2} \sim F(n_1 - 1, n_2 - 1)$$

6. 样本比率 $\hat{p}$

例：在n个样品中，当废品的个数是 x 时，废品的比率是 $\hat{p} = x/n$

$$\mathrm{E}(\hat{p}) = \mathrm{E}\left(\frac{x}{n}\right) = \frac{1}{n}\mathrm{E}(x) = \frac{1}{n}np = p$$

$$\mathrm{D}(\hat{p}) = \mathrm{D}\left(\frac{x}{n}\right) = \frac{1}{n^2}\mathrm{D}(x) = \frac{1}{n^2}np(1-p) = \frac{1}{n}p(1-p)$$

根据中心极限定理，当样本容量较大时 $np \geq 5,\ np \geq 5$

$5 \leq np \leq n\text{-}5$

$$\hat{p} \sim N\left(p, \frac{1}{n}p(1-p)\right)$$

反例：$p = 0.5$，$n = 10$

$$Z = \frac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \sim N(0,1)$$

$q = (1\text{-}p)$

总结：几种常用的抽样分布

$X \sim N(\mu, \sigma^2) \leftarrow (X_1, X_2, \cdots, X_n)$

(1) $\bar{X} \sim N(\mu, \sigma^2 / n)$

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

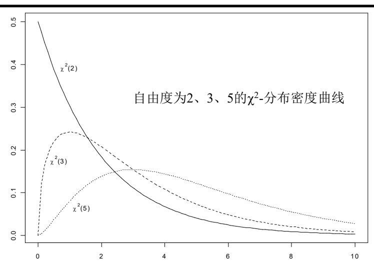
(3) $t = \dfrac{\bar{X} - \mu}{S / \sqrt{n}} \sim t(n-1)$

(4) $Z = \dfrac{\hat{p} - p}{\sqrt{pq/n}} \sim N(0,1)$ 大样本

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

(6) $F = \dfrac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$, if $\sigma_1^2 = \sigma_2^2$，X 与 Y 相互独立

学习用Excel查表



例题： $\chi^2 \sim \chi^2(25)$

求： $P(\chi^2 > x) = 0.05$

$\chi_{0.5}^2(25) = 37.65$

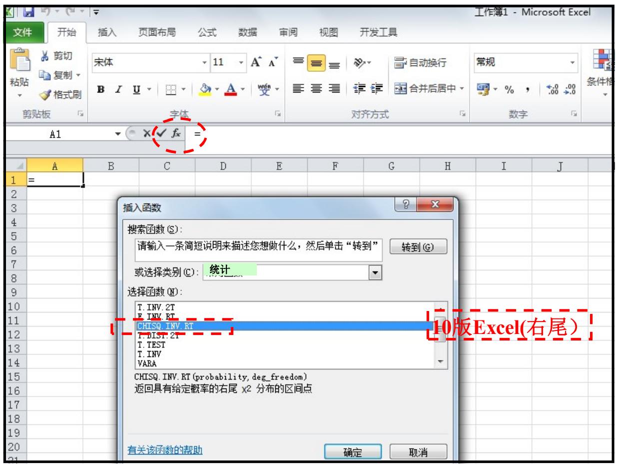
10 版 Excel: *f*(x) **统计**

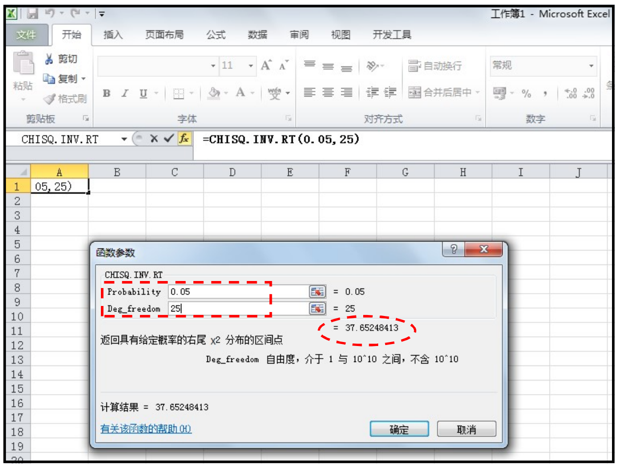
CHISQ. INV.RT: （右尾）

Probability: 0.05

Deg-freedom: 25

χ²分布的上α分位数表





例题：

10版 Excel: $f(x)$ 插入函数

——或选择类别(C): 统计

考虑一个自由度等于 9 的 t—分布（n=9）

(1) 求 $P(|T| > t_{0.025}(9)) = 0.05$

解: $t_{0.025}(9) = 2.262$

Excel: $f(x)$

T. INV.2T (双尾) :

Probability: 0.05

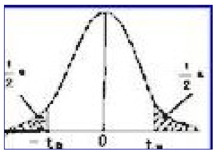
Deg-freedom: 9

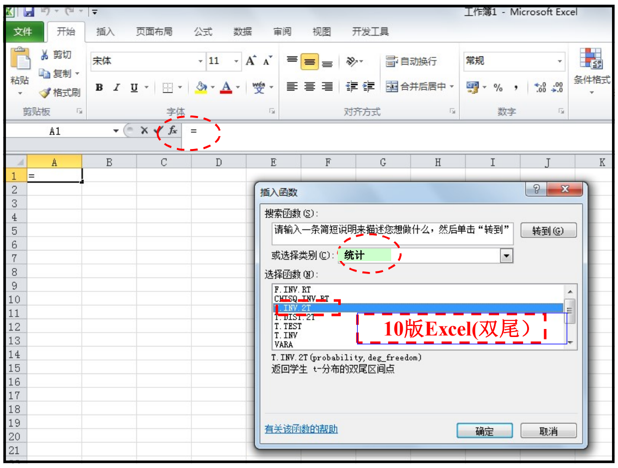
(2) 求 $P(T > t_{0.05}(9)) = 0.05$

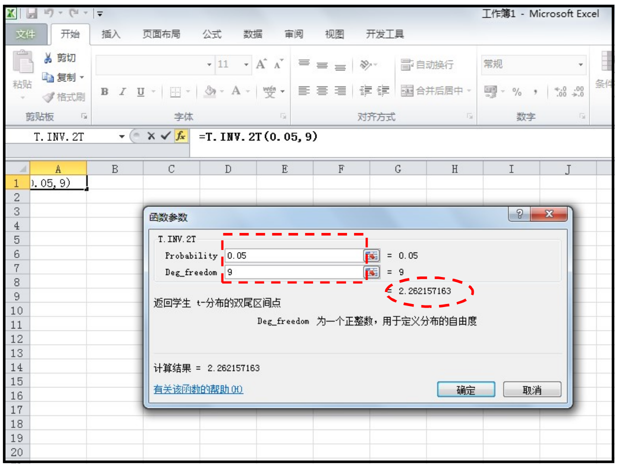
解: $t_{0.05}(9) = 1.833$

Probability: 0.10

Deg-freedom: 9







F-分布的分位点：

10版 Excel: $f(x)$

F. INV.RT: （右尾）

Probability: 0.05

Deg-freedom1: 8

Deg-freedom2: 4

$$n_1 = 8, \quad n_2 = 4$$

$$F_{0.05}(8,4) = 6.04$$

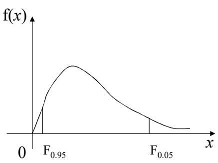
$$F_{0.95}(8,4) = 0.26$$

F. INV.RT:

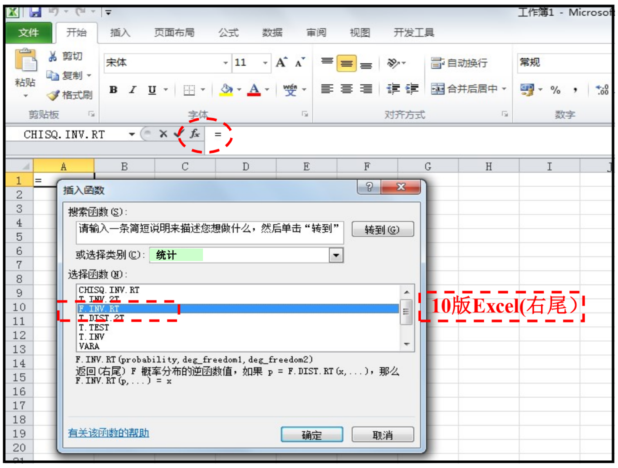
Probability: 0.95

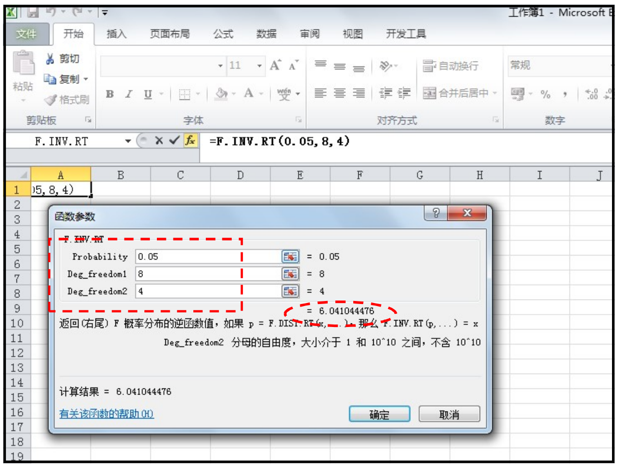
Deg-freedom1: 8

Deg-freedom2: 4



$$P(0.26 \le F \le 6.04) = 0.9$$





第二讲 参数估计

统计推断的任务：

根据样本信息，推断总体的统计规律

参数估计：问题提出

假设总体分布函数的形式已知，但它的一个或多个参数未知。如何根据样本数据，构造适当的样本函数来估计总体分布中的未知参数？

2.1 点估计的基本概念

参数估计的两种方法 $\begin{cases} \text{点估计} \\ \text{区间估计} \end{cases}$

点估计： 用样本构造一个统计量 $\hat{\theta} = f(X_1, \cdots, X_2)$，用于推断总体参数 θ。

称 $\hat{\theta}$ 为 θ 的**点估计**或**点估计量**，简称**估计**

称 $\hat{\theta}$ 的取值为 θ 的**点估计值**或**估计值**

问题：估计量 $\hat{\theta}$ 是常数还是随机变量？

基本概念

1. **总体参数** θ（Parameter） **客观存在**
2. 样本 （Sample） $X \leftarrow (X_1, X_2, \cdots, X_n)$
3. 样本容量 n （Sample Size）
4. 统计量（Sample Statistic） $\hat{\theta} = f(X_1, \cdots, X_2)$

   例如： $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$，　$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
5. 待估参数 (Estimated Parameter）
6. **估计量**（Estimator）：$\overline{X}$，S^2
7. 估计值（Estimate）：$\bar{x}$，s^2
8. **抽样误差**（Sampling Error）：$\boxed{|\bar{x} - \mu|}$

2.1 点估计的基本概念

例如：总体均值与总体方差的点估计

$X \leftarrow (X_1, X_2, \cdots, X_n)$

1. 总体均值的点估计——样本均值

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

2. 总体方差的点估计——样本方差

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

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

2.1 点估计的基本概念

如何构造统计量?

- 没有明确确定，只要满足一定的合理性即可
- 涉及两个问题：
  - 其一：如何给出估计，即估计的方法问题
  - 其二：如何对不同的估计进行分析，即估计的好坏判断标准

2.2 矩估计

1900年，英国统计学家K. Pearson提出替换原则，后人称之为矩法

- 替换原则：
  - 用样本矩替换总体矩，既可以是原点矩也可以是中心距
  - 用样本矩的函数替换相应的总体矩的函数
- 矩法的优点：
  - 简单！
    - 在总体分布形式未知场合也可对各种参数作出估计
  - 实质是用经验分布函数替换总体分布，理论基础是格里纹科定理

2.2 矩估计

例2.1：设总体X的均值μ和方差σ^2都存在，且$\sigma^2>0$，μ,σ^2均未知，$(X_1,X_2,\cdots,X_n)$是取自X的一个样本，试求μ,σ^2的矩估计。

解：先求总体矩： $\boxed{\mu_1 = E(X)} = \mu$

$$\because \quad Var(X) = E(X^2) - E^2(X)$$

$$\therefore \quad \boxed{\mu_2 = E(X^2)} = Var(X) + E^2(X) = \sigma^2 + \mu^2$$

再求样本矩：

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}, \qquad \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

$$\begin{cases} \hat{\mu} = \bar{X} \\ \hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \end{cases}$$

矩估计的的方差比 s^2 略小

例2.2：X服从泊松分布$\mathcal{P}(\lambda)$ $\quad P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$

$(X_1,X_2,\cdots,X_n)$是取自X的一个样本

试求：λ 的矩估计

解： 对于泊松分布$\mathcal{P}(\lambda)$

$$\lambda = \mathrm{E}(X) = \mu_1$$

$$\boxed{\hat{\lambda} = \hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i}$$

例题2.1（P147）： 某高校组织了12次科普报告会。统计了每次听报告的人数。如果每次报告会的听众数相互独立，都服从泊松分布P(λ)，试估计参数λ。

$$\hat{\lambda} = \hat{\mu}_1 = \frac{1}{12}\sum_{i=1}^{n} x_i = 162.083$$

例2.3：X服从指数分布$\mathcal{E}(\lambda)$ $\quad f(x) = \lambda e^{-\lambda x} \quad (当x>0)$

$(X_1,X_2,\cdots,X_n)$是取自X的一个样本

试求： λ 的矩估计

解： 对于指数分布$\mathcal{E}(\lambda)$:

$$\mathrm{E}(X) = 1/\lambda$$

$$\boxed{\hat{\lambda} = \frac{1}{\hat{\mu}_1} = \frac{1}{\bar{X}_n}}$$

矩估计可能不唯一，此时通常尽量采用低阶矩给出未知参数的估计。

例题2.2（P148）： 网民甲在搜狐网上发帖后就开始观察跟帖情况，并记录了跟帖的间隔时间（单位s）。假设跟帖间隔时间服从指数分布$\varepsilon(\lambda)$，试估计参数 λ

$$\bar{x}_n = 46.525 \;\Rightarrow\; \hat{\lambda} = 1/\hat{\mu}_1 = 0.0215$$

例2.4：X服从均匀分布$\mathcal{U}[a,b]$ $f(x)=1/(b-a)$ （当$a\le x\le b$）

$(X_1,X_2,\cdots,X_n)$是取自X的一个独立同分布样本

试求： a,b 的矩估计

对于均匀分布$\mathcal{U}[a,b]$：

$$E(X)=\frac{a+b}{2}=\mu_1$$

$$Var(X)=\frac{(b-a)^2}{12}=\mu_2-\mu_1^2$$

$$\Rightarrow \boxed{\begin{cases}\hat{a}=\hat{\mu}_1-\sqrt{3\left(\hat{\mu}_2-\hat{\mu}_1^2\right)}\\ \hat{b}=\hat{\mu}_1+\sqrt{3\left(\hat{\mu}_2-\hat{\mu}_1^2\right)}\end{cases}}$$

总结： 矩估计法

如果总体X的分布函数$F(x;\theta_1,\theta_2)$有两个未知参数θ_1,θ_2，则$\mu_1=EX$和$\mu_2=EX^2$经常和θ_1,θ_2有关。

如果 $g_1(s,t)$, $g_2(s,t)$ 是已知函数，且θ_1 和θ_2 能表示成：

$$\begin{cases}\mu_1=EX\\ \mu_2=EX^2\end{cases} \Rightarrow \begin{cases}\theta_1=g_1(\mu_1,\mu_2)\\ \theta_2=g_2(\mu_1,\mu_2)\end{cases}$$

则：

$$\boxed{\begin{cases}\hat{\theta}_1=g_1(\hat{\mu}_1,\hat{\mu}_2)\\ \hat{\theta}_2=g_2(\hat{\mu}_1,\hat{\mu}_2)\end{cases}}$$

矩估计的特点： 不运用总体 X 的分布信息

2.3 极大似然估计

- 极大似然估计，最早由高斯在1821年提出，但一般将之归功于费希尔（R.A. Fisher），因为费希尔在1922年再次提出这种想法，并证明了它的一些性质，从而使极大似然估计得到广泛的应用。
- **例2.5**：设有外形完全相同的两个箱子，甲箱中有99个白球和1个黑球，乙箱中有99个黑球和1个白球，今随机地抽取一箱，并从中随机地抽取一球，结果取得白球。问这球是从哪一个箱子中取出?

2.3 极大似然估计

极大似然估计的原理

考察以下例子：

假设在一个罐中放着许多黑球和红球，并假定已经知道两种球的数目之比是1:3，但不知道哪种颜色的球多。如果用放回抽样方法从罐中任取 n 个球，则其中红球的个数为 k 的概率为：

记： p 是红球的概率

$P(k;p)=C_n^k p^k q^{n-k}$，其中$q=1-p$,由假设知，$p=\frac{1}{4}$或$\frac{3}{4}$

若取$n=3$，如何通过 k 来估计p 值。先计算抽样的可能结果k在这两种p 值之下的概率：

k	0	1	2	3
$P(k,\frac{3}{4})$	1/64	9/64	27/64	27/64
$P(k,\frac{1}{4})$	27/64	27/64	9/64	1/64

记： p 是红球的概率

k	0	1	2	3
$P(k,\frac{3}{4})$	1/64	9/64	27/64	27/64
$P(k,\frac{1}{4})$	27/64	27/64	9/64	1/64

$n=3$, $p=3/4$ 或 $p=1/4$

$k=$ 0, 1, 2, 3

从上表看到：

$k=0$, $P\left(0,\frac{1}{4}\right)=\frac{27}{64}>P\left(0,\frac{3}{4}\right)=\frac{1}{64}$,取 $\hat{p}=\frac{1}{4}$更合理；

$k=1$ 类似；

$k=2$, $P\left(2,\frac{1}{4}\right)=\frac{9}{64}<P\left(2,\frac{3}{4}\right)=\frac{27}{64}$, 取$\hat{p}=\frac{3}{4}$更合理；

$k=3$类似；

于是有：$\hat{p}(X=k)=\begin{cases}\frac{1}{4} & k=0,1\\ \frac{3}{4} & k=2,3\end{cases}$

例2.6. 甲、乙两人下棋，用p表示甲在每局中获胜的概率。如果在5局中，甲胜了3局，请问：$p=?$

$$L(p)=C_5^3p^3(1-p)^2$$

现在已知 X=3, 所以 p 应该使 X=3 发生的概率最大。

对$L(p)$ 求导数，令：

$$L'(p)=C_5^3\left[3p^2(1-p)^2-2p^3(1-p)\right]$$
$$=C_5^3p^2(1-p)\left[3(1-p)-2p\right]=0$$

解： $\left.\begin{array}{l}p=0\\ p=1\end{array}\right\}$ 不符合题意

$$\boxed{p=\frac{3}{5}}$$

极大似然估计方法

(一) 离散分布的情况

定义2.1 设**离散**随机变量 $X_1,X_2,\cdots,X_n$ 有**联合分布**

$$p(x_1,x_2,\cdots,x_n;\theta)=P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n)$$

其中，θ是未知参数。给定观测数据 $x_1,x_2,\cdots,x_n$ 后，

称 θ 的函数为**似然函数**：

$$\boxed{L(\theta)=p(x_1,x_2,\cdots,x_n;\theta)}$$

称$\hat{\theta}$ 为 θ 的**极大似然估计MLE**(maximum likelihood estimator)：

$$\hat{\theta}=\underset{\theta\in\Theta}{Max}\ L(\theta)$$

可以解出MLE： $\boxed{L'(\theta)=0}$

例2.7 设 $X_1,X_2,\cdots,X_n$ **独立同分布**，都服从泊松分布 $\mathcal{P}(\lambda)$

给定 $X_1,X_2,\cdots,X_{12}$ 的观察值：

169 167 157 196 163 151 154 157 163 154 162 165

计算 λ 的MLE 。

解：λ 的似然函数为

$$\begin{aligned}L(\lambda)&=P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n)\\&=P(X_1=x_1)P(X_2=x_2)\cdots P(X_n=x_n)\\&=\frac{\lambda^{x_1}}{x_1!}e^{-\lambda}\frac{\lambda^{x_2}}{x_2!}e^{-\lambda}\cdots\frac{\lambda^{x_n}}{x_n!}e^{-\lambda}\\&=\frac{\lambda^{x_1+x_2+\cdots+x_n}}{x_1!x_2!\cdots x_n!}e^{-n\lambda}\end{aligned}$$

对数似然函数：

$$l(\lambda)=\ln L(\lambda)=(x_1+x_2+\cdots+x_n)\ln\lambda-n\lambda-c_0$$

求导数：

$$\frac{x_1+x_2\cdots+x_n}{\lambda}-n=0 \Rightarrow \lambda=\frac{x_1+x_2\cdots+x_{12}}{12}=163.167$$

总结： 离散分布的MLE

似然函数： $L(\theta)=p(x_1,x_2,\cdots,x_n;\theta)$

对数似然函数： $\boxed{l(\theta)=\ln L(\theta)}$

似然方程： $\boxed{l'(\theta)=0}$

最大似然估计（MLE）： 似然方程的解 θ

注意： 因为 $\ln x$ 是严格单调增函数，

所以：$l(\theta)$与$\ln L(\theta)$有相同的最大值点

例2.8 设 $X_1,X_2,\cdots,X_n$ 独立同分布，都服从两点分布 $\mathcal{B}(1,p)$。

给定观察值 $x_1,x_2,\cdots,x_n$ ，计算 p 的MLE 。

解：

$$\because\ P(X_i=x_i)=p^{x_i}(1-p)^{1-x_i}$$

$$\begin{aligned}\therefore\ L(p)&=P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n)\\&=P(X_1=x_1)P(X_2=x_2)\cdots P(X_n=x_n)\\&=p^{n\bar{x}}(1-p)^{n-n\bar{x}}\end{aligned}$$

对数似然函数： $l(p)=\ln L(p)=n\bar{x}\ln p+(n-n\bar{x})\ln(1-p)$

求导数： $l'(p)=n\bar{x}/p-(n-n\bar{x})/(1-p)=0$

$$\Rightarrow \hat{p}=\bar{x} \qquad \hat{p}=\bar{x}=\frac{k}{n}$$

(二) 连续分布的情况

定义2.2 设随机向量 $\mathbf{X}=(X_1,X_2,\cdots,X_n)$ 有**联合密度** $f(\mathbf{X},\boldsymbol{\theta})$

其中 $\boldsymbol{\theta}=(\theta_1,\theta_2,\cdots,\theta_m)$ 是未知参数。

当得到 $\mathbf{X}$ 的观测值 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$ 后，

称**似然函数**： $\mathrm{L}(\boldsymbol{\theta})=f(\mathbf{X},\boldsymbol{\theta})$

设总体 X 有概率密度 $f(x,\theta)$， 则 i.i.d 样本 $X_1,X_2,\cdots,X_n$

有联合密度 $f(x_1,x_2,\cdots,x_n;\boldsymbol{\theta})=\boxed{\prod_{i=1}^{n}f(x_i;\boldsymbol{\theta})}$

似然函数： $L(\boldsymbol{\theta})=\prod_{i=1}^{n}f(x_i;\boldsymbol{\theta})$ **对数似然函数**： $l(\boldsymbol{\theta})=\ln L(\boldsymbol{\theta})$

似然方程： $\boxed{\frac{\partial l(\boldsymbol{\theta})}{\partial\theta_j}=0 \quad j=1,2,\cdots,m}$

例2.9 设 $X_1,X_2,\cdots,X_n$ 独立同分布，都服从指数分布 $\mathcal{E}(\lambda)$

给定观察值 $x_1,x_2,\cdots,x_n$ ，计算 λ 的MLE

解：指数分布 $\mathcal{E}(\lambda)$的概率密度是： $f(x;\lambda)=\lambda e^{-\lambda x},\quad x\ge 0$

λ 的似然函数： $L(\lambda)=\lambda^n\exp\left(-\lambda\sum_{i=1}^{n}x_i\right)$

对数似然函数： $l(\lambda)=\ln L(\lambda)=n\ln\lambda-\lambda\sum_{i=1}^{n}x_i$

求导数： $l'(\lambda)=\frac{n}{\lambda}-\sum_{i=1}^{n}x_i=0 \Rightarrow \frac{1}{\lambda}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}$

MLE $\Rightarrow$ $\lambda=\frac{1}{\bar{x}}$

例2.10 设$X_1,X_2,\cdots,X_n$ 独立同分布，都服从正态分布$N\left(\mu,\sigma^2\right)$

给定观察数据 $x_1,x_2,\cdots,x_n$，试估计 μ,σ^2

解：正态分布 $N\left(\mu,\sigma^2\right)$的概率密度

$$f\left(x;\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

似然函数： 基于观测 $x_1,x_2,\cdots,x_n$

$$L\left(\mu,\sigma^2\right)=\prod_{i=1}^{n}f\left(x_i;\mu,\sigma^2\right)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$=\frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n}\exp\left[-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

对数似然函数： 记：$a=\sigma^2$

$$l\left(\mu,a\right)=\ln L\left(\mu,a\right)=-\frac{n}{2}\ln a-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2a}+c_0$$

对 μ,a 求偏导：

$$\begin{cases}\dfrac{\partial l}{\partial\mu}=\dfrac{1}{a}\sum_{i=1}^{n}(x_i-\mu)=0\\ \dfrac{\partial l}{\partial a}=-\dfrac{n}{2a}+\dfrac{1}{2a^2}\sum_{i=1}^{n}(x_i-\mu)^2=0\end{cases}\Rightarrow\begin{cases}\sum_{i=1}^{n}x_i=n\mu\\ \dfrac{n}{2a}=\dfrac{1}{2a^2}\sum_{i=1}^{n}(x_i-\mu)^2\end{cases}$$

MLE ➡

$$\begin{cases}\hat{\mu}=\dfrac{1}{n}\sum_{i=1}^{n}x_i\\ \hat{\sigma}^2=\dfrac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2\end{cases}$$

例2.11 设总体X服从$[a,b]$上的均匀分布，a，b未知，试由样本 $x_1,x_2,\cdots,x_n$求出a，b的极大似然估计

解： X的概率密度为：$f\left(x;a,b\right)=\frac{1}{b-a}\mathrm{I}\left(a\le x\le b\right)$

定义： $x_{(1)}=\min\left\{x_1,x_2,\cdots,x_n\right\}$ $x_{(n)}=max\left\{x_1,x_2,\cdots,x_n\right\}$

似然函数：$L\left(a,b\right)=\frac{1}{(b-a)^n}\prod_{i=1}^{n}\mathrm{I}\left(a\le x_i\le b\right)$

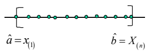
$=\frac{1}{(b-a)^n}\mathrm{I}\left(a\le x_{(1)}\le x_{(n)}\le b\right)$

所有样本点x_i都要在[a, b]区间里。
否则，似然函数就会等于0!

要求 L(*a*,*b*) 取到最大：

（1）必须有示性函数等于1，即 $a\le x_{(1)}$，$x_{(n)}\le b$

（2） 需要$\frac{1}{(b-a)^n}$达到最大：$\boxed{\hat{a}=X_{(1)},\ \hat{b}=X_{(n)}}$



(b-a)应达到最小：$b=x_{(n)}$，$a=x_{(1)}$； 所有xi都要在[a, b]里

2.3 极大似然估计

- **极大似然估计的不变性**
- 如果 $\hat{\theta}$ 是θ的极大似然估计，则对任一函数g(θ)，其极大似然估计为g($\hat{\theta}$)
- 例2.12 设 $X_1,X_2,\cdots,X_n$ 是来自正态总体 $N\left(\mu,\sigma^2\right)$ 的样本。试求标准差 σ 和概率$P(X<3)$ 的极大似然估计
  - 由例2.10知，$\hat{\mu}=\bar{X}=\frac{1}{n}\sum_{i=1}^{n}x_i$，$\hat{\sigma}^2=s_n^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$
  - 因此，标准差的MLE是$\hat{\sigma}=s_n=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2}$
  - $P(X<3)$ 的MLE是$\Phi\left(\frac{3-\bar{X}}{s_n}\right)$

2.4 估计量的评价标准

对总体的未知参数可用不同方法求得不同的估计量，如何评价好坏?

常用的三条标准：无偏性，有效性，相合性

1、无偏性： $\hat{\theta}$ 是 θ 的无偏估计量：

$$\mathrm{E}(\hat{\theta})=\theta$$

若$E\left(\hat{\theta}\right)\ne\theta$,那么$\left|E\left(\hat{\theta}\right)-\theta\right|$称为估计量$\hat{\theta}$的偏差

例2.13： 总体中有5个数据：1，2，3，4，5

总体均值： $\mu=\frac{1}{5}(1+2+3+4+5)=\frac{15}{5}=3$

抽取**容量为3**的样本：$\left(x_1,x_2,x_3\right)$

样本	第1次抽样	第2次抽样	第3次抽样	第4次抽样	第5次抽样	第6次抽样
x_1	1	1	1	2	2	3
x_2	2	3	3	3	3	4
x_3	3	4	5	4	5	5
Σx_i	6	8	9	9	10	12
$\hat{\mu}$	2	2.666667	3	3	3.333333	4

$$E\left(\hat{\mu}\right)=\frac{1}{6}(2+2.666667+3+3+3.333333+4)=\frac{18}{6}=3$$

例2.14：设总体X的一阶和二阶矩存在，分布是任意的，记$E(X)=\mu, D(X)=\sigma^2$，

证明：样本均值$\bar{X}$ 和样本方差S^2 分别是μ和σ^2的无偏估计。

证：因$X_1, X_2, \cdots, X_n$与X同分布，故有：

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

$$(\bar{X}-\mu)\left[n\bar{X}-n\mu\right]$$

故$\bar{X}$是μ的无偏估计

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

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

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

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

$$=\frac{1}{n-1}\left(n\sigma^2-n\times\frac{\sigma^2}{n}\right)=\sigma^2$$

故S^2是σ^2的无偏估计

可以证明：

$s_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$ **不是无偏估计量，因为** $\mathrm{E}(s_n^2)=\frac{n-1}{n}\sigma^2$

注：当样本量趋于无穷时，有 $\mathrm{E}(s_n^2)\to\sigma^2$，
称 s_n^2 为σ^2 的渐变无偏估计

总结：（1）对任一总体而言，样本均值是总体均值的无偏估计；当总体k阶矩存在时，样本k阶原点矩是总体k阶原点矩的无偏估计。k阶中心距则不一样。
（2）无偏性不具有不变性：即若$\hat{\theta}$是θ的无偏估计，一般而言，g($\hat{\theta}$)不是g(θ)的无偏估计，除非g(θ)是θ的线性函数。

2、有效性：

定义2.3．$\hat{\theta}_1$和$\hat{\theta}_2$都是θ的无偏估计量，若：

$$\mathrm{D}(\hat{\theta}_1)<\mathrm{D}(\hat{\theta}_2)$$

则称 $\hat{\theta}_1$比 $\hat{\theta}_2$更有效。

定义2.4. 若 $\hat{\theta}$ 是总体参数 θ的无偏估计中方差最小的，称 $\hat{\theta}$ 是 θ的最小方差无偏估计量 (**Minimum Variance Unbiased Estimator**) 。

$X\sim N(\mu,\sigma^2)\leftarrow(X_1,X_2,\cdots,X_n)$: $\bar{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i$
(i.i.d)

则：$\bar{X}_n\sim N(\mu,\sigma^2/n)$

例2.15. $X\leftarrow(X_1,X_2,\cdots,X_n)$ (i.i.d)

记：$\bar{X}_{n-1}=\frac{1}{n-1}\sum_{i=1}^{n-1}X_i$

它也是无偏估计量：

$$E(\bar{X}_{n-1})=\frac{1}{n-1}\sum_{i=1}^{n-1}E(X_i)=\mu$$

但是因为少用了一个数据，因此**它的方差比较大：**

$$\boxed{Var(\bar{X}_{n-1})=\frac{\sigma^2}{n-1}}>Var(\bar{X}_n)=\frac{\sigma^2}{n}$$

- **例2.16** 由例2.11可知，均匀总体U(0,θ)中θ的极大似然估计是$X_{(n)}$。
- 但是$X_{(n)}$不是θ的无偏估计，是θ的渐近无偏估计，因为$E(X_{(n)})=\frac{n}{n+1}\theta$
- 修偏后可得θ的一个无偏估计：$\hat{\theta}_1=\frac{n+1}{n}X_{(n)}$
- 由矩估计，可得θ是另个无偏估计：$\hat{\theta}_2$=2$X_{(n)}$
- 哪个无偏估计更有效?

问题：有偏估计一定是不好的估计吗?

一般而言，在样本量一定时，评价点估计的最一般标准是点估计值$\hat{\theta}$ 与真实值θ的距离的函数。最常用的函数是距离的平方。

均方误差： $\mathrm{MSE}(\hat{\theta})=E(\hat{\theta}-\theta)^2$

注意： $\mathrm{MSE}(\hat{\theta})=D(\hat{\theta})+(E\hat{\theta}-\theta)^2$

均方误差由点估计的方差与偏差的平方两部分组成。
对于无偏估计， $\mathrm{MSE}(\hat{\theta})=D(\hat{\theta})$

- **例2.17** 在例2.16中，对于均匀总体U(0,θ)，由θ的极大似然估计得到的无偏估计是$\hat{\theta}=\frac{n+1}{n}X_{(n)}$
- 它的均方误差是 $MSE(\hat{\theta})=D(\hat{\theta})=\frac{\theta^2}{n(n+2)}$
- 考虑形如 $\hat{\theta}_\alpha=\alpha\cdot X_{(n)}$ 的估计，其均方误差为

$$MSE(\hat{\theta}_\alpha)=\alpha^2\frac{n}{(n+1)^2(n+2)}\theta^2+\left(\frac{n\alpha}{n+1}-1\right)\theta^2$$

- 易证，当 $\alpha=(n+2)/(n+1)$时，上述均方误差达到最小， $MSE\left(\frac{n+2}{n+1}X_{(n)}\right)=\frac{\theta^2}{(n+1)^2}$
- 在均方误差标准下，有偏估计vs无偏估计

3、相合性

复习：随机变量依概率收敛的定义（P110）

定义2.5： 设 $U_1,U_2,\cdots U_n,\cdots$ 是随机变量，如果 ∀ε > 0，有：

$$\lim_{n\to\infty}P\{|U_n-U|\geq\varepsilon\}=0$$

则称 U_n 依概率收敛到 U，记做：

$U_n\xrightarrow{P}U$ **依概率收敛**

复习：大数定律

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

弱大数定律： 设随机变量 $X_1,X_2,\cdots X_n,\cdots$ 独立同分布，μ=E(X_1)，则 ∀ε > 0，有：

随着样本中人数增多，样本组同学的平均高度会更接近全班平均高度。但也不排除一组大个子的同学被抽中。

$$\lim_{n\to\infty}P\{|\bar{X}_n-\mu|\geq\varepsilon\}=0$$

依概率收敛 $\bar{X}_n\xrightarrow{P}\mu$

强大数定律： 设随机变量 $X_1,X_2,\cdots X_n,\cdots$ 独立同分布，μ=E(X_1)，则：

样本均值趋于总体均值的事件必然会发生

$$P\left(\lim_{n\to+\infty}\bar{X}_n=\mu\right)=1$$

几乎处处收敛 $\lim_{n\to+\infty}\bar{X}_n=\mu \quad a.s$

估计量的相合性

定义2.6：设 $\hat{\theta}_n$ 为参数θ的估计量，$\forall\varepsilon>0$，如果有：

$\lim_{n\to+\infty}P\{|\hat{\theta}_n-\theta|\geq\varepsilon\}=0$成立，

依概率收敛 $\theta_n\xrightarrow{P}\theta$

则称$\hat{\theta}_n$为θ的相合估计量，或一致估计量

强相合估计量：

$\lim_{n\to+\infty}\theta_n=\theta \quad a.s$ **几乎处处收敛**

注意：相合性在样本容量较大时才适用

例，对于简单随机样本，样本均值和样本方差分别是总体均值与总体方差的相合估计量。

◆ A. 样本均值

强大数定律： 如果$X_1,X_2,\cdots,X_n\cdots$ 是独立同分布的随机变量，μ=E(X_1)，则：

$$\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^n X_i=\mu \quad a.s$$

几乎处处收敛

所以，样本均值是总体均值的强相合估计，因此也是相合估计

◆ B. 样本方差

对于：$X_1^2,X_2^2,\cdots,X_n^2$ i.i.d, $E(X_1^2)=E(X^2)$

根据强大数定律： $\lim_{n\to+\infty}\frac{1}{n-1}\sum_{i=1}^n X_i^2\to E(X^2) \quad a.s$

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

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

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

$$\to E(X^2)-\mu^2=\sigma^2 \quad a.s$$ **几乎处处收敛**

样本方差是总体方差的强相合估计；也是相合估计

◆ C. 样本标准差

根据强大数定律： $\lim\limits_{n\to+\infty}\frac{1}{n-1}\sum_{i=1}^{n}X_i^2 \to E(X^2)\quad a.s$

$$s=\sqrt{s^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\hat{\mu})^2}$$

因为： $s^2\to\sigma^2\ \ a.s$

所以： $s\to\sigma\ \ a.s$ **几乎处处收敛**

样本标准方差是总体标准差的强相合估计；

也是相合估计

但是，样本标准差不是无偏估计量！

施瓦茨不等式：$EX^2<\infty$, $EY^2<\infty$，则有

$$|E(XY)|\le\sqrt{EX^2EY^2}$$

并且，**等号成立的充分必要条件**是有不全为零的常数 a, b,

使得： $P(aX+bY=0)=1$

当 $\sigma>0$，不存在不全为零的常数 a, b, 使得：

$$P(a\times s+b\times 1=0)=1$$

$$E(s)<\sqrt{Es^2E1^2}=\sqrt{\sigma^2}=\sigma$$

于是 $E(s)<\sigma$，这时 s 低估了 σ

如何判断估计的相合性?

- **定理2.1** 设 $\hat{\theta}_n=\hat{\theta}_n(X_1,X_2,\cdots,X_n)$ 是 θ 的一个估计量，若
$$\lim_{n\to+\infty}E(\hat{\theta}_n)=\theta,\quad \lim_{n\to+\infty}D(\hat{\theta}_n)=0$$
则 $\hat{\theta}_n$ 是 θ 的相合估计。
- **定理2.2** 若 $\hat{\theta}_{n1},\hat{\theta}_{n2},\cdots,\hat{\theta}_{nk}$ 是 $\theta_1,\theta_2,\cdots,\theta_k$ 的相合估计，$\eta=g(\theta_1,\theta_2,\cdots,\theta_k)$ 是 $\theta_1,\theta_2,\cdots,\theta_k$ 的连续函数，则 $\hat{\eta}_n=g(\hat{\theta}_{n1},\hat{\theta}_{n2},\cdots,\hat{\theta}_{nk})$ 是 η 的相合估计。

定理2.3 设 $X_1,X_2,\cdots,X_n$ 是总体 X 的样本，

$\mu=EX$, $\sigma^2=Var(X)>0$

则有：

(1) 样本均值： 是总体均值的强相合、无偏估计

(2) 样本方差： 是总体方差的强相合、无偏估计

(3) 样本标准差： 是总体标准差的强相合估计，
但是 $E(S)<\sigma$

例2.18： $X\leftarrow X_1,X_2,\cdots,X_n$ 独立同分布样本

$\mu_k=EX^k$ 存在

称： **X 的 k 阶原点矩：** $\boxed{\mu_k=EX^k}$ 总体意义上

X 的 k 阶样本原点矩： $\boxed{\hat{\mu}_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k}$ 样本意义上

因为 $X_1^k,X_2^k,\cdots,X_n^k$ 独立同分布；并且与 X^k 同分布 $\mu_k=EX^k$

根据定理2.1（1），知道：

$\hat{\mu}_k$ 是 μ_k 的强相合无偏估计

称： $\hat{\mu}_k$ 是 $\mu_k=EX^k$ 的“矩估计”

第三讲 参数的区间估计

参数估计问题:

问题：想象你经营一个食品商店,问能否根据下面的市场调查结果进行决策

(1) 点估计：软饮料的每日平均需求量是 300 瓶;

(2)软饮料的每日平均需求量是每日 300 ±50瓶;

你认为可以利用上面的信息进行决策吗?

3.1 区间估计的概念

Interval Estimation

在总体 X 抽取一个容量为 n 的随机样本

X: $\leftarrow X_1, X_2, \ldots, X_n$

利用样本构造两个统计量 $(\hat{\theta}_1, \hat{\theta}_2)$

使得: $P\left\{\hat{\theta}_1 \le \theta \le \hat{\theta}_2\right\} = 1-\alpha$

置信区间 Confidence Interval (C.I.): $(\hat{\theta}_1, \hat{\theta}_2)$

置信度 Level of Confidence: $1-\alpha$

已知总体方差，总体均值μ的置信区间

给定 正态总体: $X \sim N(\mu, \sigma^2)$

总体方差 σ^2 已知

在总体中抽取一个容量为 n 的样本:

$X \leftarrow (X_1, X_2, \ldots, X_n)$

根据样本均值的性质，有:

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

枢轴量：其分布与未知参数无关

已知总体方差，总体均值μ的置信区间

于是得到: $P\left\{\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| \le z_{\alpha/2}\right\} = P\{|z| \le z_{\alpha/2}\} = 1-\alpha$

即以1-α 的概率保证 $\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| \le z_{\alpha/2}$

也就是说，以1-α 的概率保证

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

此时，$\theta = \mu$，$\hat{\theta}_1 = \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \hat{\theta}_2 = \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

注意正确的叙述方法:

$$P\left\{\hat{\theta}_1 \le \theta \le \hat{\theta}_2\right\} = 1-\alpha$$

置信区间 **覆盖**参数 θ 的概率是 $(1-\alpha)$.

— θ 不是随机变量

— C.I. 是随机区间，会随样本的不同而发生变化.

置信度为95% 意味着：大约有 95%的 置信区间包含参数 θ, 而5%的置信区间不包含 θ.

假设：总体比例为 p　　点估计值为75%

已经求得其95%的置信区间为：(72%, 78%)

可否说：这个置信区间包含总体比例的概率为95%?

如何理解：“置信区间覆盖参数 θ 的概率是 $(1-\alpha)$”

吴喜之教授强调：“这里的区间（72%, 78%）是固定的，而总体比例 p 也是固定的值。因此只有两种可能：或者该区间包含总体比例，或者不包含；这当中没有任何概率可言。至于区间（72%, 78%）是否覆盖真正比例，除非一个不漏地调查所有选民，否则永远也无法知道。”

发令枪生产企业	甲工厂	乙工厂
合格率	95%	20%

置信度：　用于评价“估计方法”的可信程度

假设：总体比例为 p　　点估计值为75%

已经求得其95%的置信区间为：(72%, 78%)

可否说：这个置信区间包含总体比例的概率为95%?

如何理解：“置信区间覆盖参数 θ 的概率是 $(1-\alpha)$”

比较准确的表达是：　点估计值为75%；

此次估计的误差范围是±3%；

用该方法估计的可靠程度是95%

置信度：　用于评价“估计方法”的可信程度

回顾：已知总体方差，总体均值μ的置信区间

【例3.1】：一家食品生产企业生产袋装食品。已知正常生产条件下，产品重量服从正态分布，总体标准差为10 g。现从某天生产的食品中随机抽取了25袋，计算出样本均值为105.36 g 。试估计该批产品平均重量的置信区间，置信水平为95%。

已知：$\sigma=10$，$n=25$，$\bar{x}=105.36$

选取 $(1-\alpha)=0.95$;

查表得到：$z_{\alpha/2}=1.96$

置信区间：$(105.36-1.96\frac{10}{\sqrt{25}},\ 105.36+1.96\frac{10}{\sqrt{25}})$　3.92

$\Rightarrow (101.44, 109.28)$

这个计算结果说明什么问题?

回顾：已知总体方差，总体均值μ的置信区间

$$\left[\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

基本结论：

（1）置信区间的中心是样本均值

（2）置信水平 $1-\alpha$ 越高，则置信区间越长

（3）样本量n越大，则置信区间越短

为了避免置信区间过长带来的不足，同时考虑置信水平不能太低，人们一般使用置信水平为 $1-\alpha=0.95$ 的置信区间，此时 $z_{\alpha/2}=z_{0.025}=1.96$

关于区间估计概念的几点讨论

问题1：在实际应用中，为什么要做区间估计?

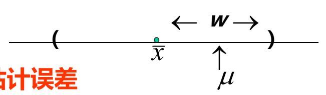
想象你经营一个食品商店.问能否根据下面的市场调查结果进行决策?

(1) 点估计：软饮料的每日平均需求量是 300 瓶

(2)软饮料的每日平均需求量是每日 300 ±10瓶

(3)置信度95%，每日平均需求量是300 ±150瓶

$w=|\bar{x}-\mu|$



置信区间——估计误差

问题2：置信区间越大，置信度越大还是越小?

讨论 1:　C.I↑，　$(1-\alpha)$↑;

C.I $\rightarrow\infty$，　$(1-\alpha)\rightarrow 1$;

讨论 2:　在同样的置信度下，置信区间越窄，估计的精度越高。

分析人员的期望：

1. 置信区间越窄越好 (抽样误差越小)
2. 置信度 $(1-\alpha)$ 越高越好（关于误差估计的结论更可靠）

定义: 总体均值的置信区间的宽度:

$$w = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$(1-\alpha)=0.90$: $z_{\alpha/2}=1.65$
$(1-\alpha)=0.95$: $z_{\alpha/2}=1.96$
$(1-\alpha)=0.99$: $z_{\alpha/2}=2.58$

均值的置信区间的宽度取决于三个因素:

1. 置信水平 (1-α): $z_{\alpha/2}$
2. 总体方差: σ
3. 样本容量: n **(是可控制的)**

问题3：为什么样本容量越大，置信区间越窄?

置信区间的宽窄和 $\bar{x}$ 的抽样分布有关: $\bar{X}\sim N(\mu,\sigma^2/n)$

事实上, 如果 $n=N$: 则 $\bar{x}=\mu$

例3.1： 一家食品生产企业生产袋装食品。按规定每袋重量为100g。已知正常生产条件下，产品重量服从正态分布，总体标准差为10 g。现从某天生产的食品中随机抽取了25袋，计算出样本均值为105.36 g 。试估计该批产品平均重量的置信区间，置信水平为95%。

置信区间 (101.44,109.28)

置信区间只是此次点估计的误差范围!!!

比较准确的表达: 平均重量的点估计值为105.36

此次估计的绝对误差是3.92

用该方法估计的可靠程度是95%

注意：上述总体均值置信区间 C.I. 的构造受以下两个限制条件的约束。

1. 必须是正态总体: $X\sim N(\mu, \sigma^2)$;
2. 总体方差已知;

但是，当样本容量充分大时，这两个条件就都不重要了.

3.2 未知总体方差σ^2，求总体均值的置信区间

一、大样本: $n \geq 30$

(1) 根据 **中心极限定理**

$$\bar{X}\sim N(\mu,\sigma^2/n)$$

(2) 样本标准差是总体标准差 σ 的良好估计:

$$s\approx\sigma$$

总体均值的置信区间是

$$(\bar{X}-\boxed{z_{\alpha/2}}\frac{s}{\sqrt{n}},\bar{X}+z_{\alpha/2}\frac{s}{\sqrt{n}})$$

例3.2：某大学从该校本科生中随机抽取100人。他们平均每日体育锻炼时间为26分钟，样本方差为34。试以95%置信水平的置信区间估计学生每日锻炼时间。

n=100, $\bar{x}$=26, s^2=34

构造总体均值的 95% 置信区间 :

$$(\bar{X}-z_{\alpha/2}\frac{s}{\sqrt{n}},\bar{X}+z_{\alpha/2}\frac{s}{\sqrt{n}})$$

$$=\left(26-1.96\sqrt{\frac{34}{100}},\ 26+\underbrace{1.96\sqrt{\frac{34}{100}}}_{1.14}\right)=26\pm1.14$$

$$\left|\frac{\bar{x}-\mu}{\mu}\right|\cong\left|\frac{w}{\bar{x}}\right|$$

w = 1.14

1.14 / 26 = 4.38%

3.2 未知总体方差σ^2，求总体均值的置信区间

二、一种新的情况:

— 正态总体

— **总体方差 σ^2 未知**

— **小样本**

问题: 使用什么统计量?

$$z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$???

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

$$P\{|t|<t_{\alpha/2}\}=1-\alpha$$

$$P\left\{-t_{\alpha/2}<\frac{\bar{X}-\mu}{s/\sqrt{n}}<t_{\alpha/2}\right\}=1-\alpha$$

$$P\left\{\bar{X}-t_{\alpha/2}\frac{s}{\sqrt{n}}<\mu<\bar{X}+t_{\alpha/2}\frac{s}{\sqrt{n}}\right\}=1-\alpha$$

置信区间:

$$\left(\bar{X}-\boxed{t_{\alpha/2}}\frac{s}{\sqrt{n}},\bar{X}+t_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

例3.3：已知某种灯泡的寿命服从正态分布，现从一批灯泡中随机抽取16只。

n=16, $\bar{x}$ =1490(小时), s =24.77 (小时)

建立该批灯泡平均寿命的置信区间，置信度为95%。

解: df = 15, (1-α) = 0.95 ⇒ α= 0.05 $t_{\alpha/2}(15)=2.131$

则置信度为95% 的置信区间为：

3.196

$$(1490-2.131\frac{24.77}{\sqrt{16}},1490+2.131\frac{24.77}{\sqrt{16}})=1490\pm3.196$$

$3.196/1490=0.886\%$

在小样本的情况下，为什么抽样误差还这么小？

3.3 总体比率的置信区间 (Large Samples)

- **总体比率 Population Proportion :** p
- **样本比率 Sample Proportion:** $\hat{p}$

如果是大样本，经验判断：$n\hat{p}\geq5$

则 $\hat{p}\sim N(p,pq/n)$

其中 $q=(1-p)$

因此

$$Z=\frac{\hat{p}-p}{\sqrt{pq/n}}\sim N(0,1)$$

总体比率的置信区间

置信度为(1-α)的置信区间为：

$$(\hat{p}-z_{\alpha/2}\sqrt{pq/n},\hat{p}+z_{\alpha/2}\sqrt{pq/n})$$

由于 p 和 q 都是未知的，因此置信区间近似为：

$$(\hat{p}-z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n},\hat{p}+z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n})$$

例3. 4：某城市想要估计下岗职工中女性所占的比例，随机抽取了100名下岗职工，其中65人为女性。试估计该城市下岗职工中女性比例，并指出估计误差。置信水平要求为95%。

已知 n=100, α= 0.05, $z_{\alpha/2}=1.96$

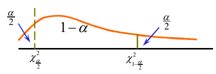
$\hat{p}=65\%$

$$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}=1.96\sqrt{\frac{0.65\times0.35}{100}}=0.0935$$

置信区间为：65% ± 9.35%

下岗职工中女性比例为65%，估计误差为9.35%

3.4 方差σ^2的置信区间



设μ未知

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

有$P\left\{\chi^2_{1-\alpha/2}(n-1)<\frac{(n-1)S^2}{\sigma^2}<\chi^2_{\alpha/2}(n-1)\right\}=1-\alpha$

即$P\left\{\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}<\sigma^2<\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right\}=1-\alpha$

置信区间为：$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)},\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right)$

例3.5 一个园艺科学家正在培养一个新品种的苹果,这种苹果除了口感好和颜色鲜艳以外,另一个重要特征是单个重量差异不大。为了评估新苹果,她随机挑选了25个测试重量(单位:克),其样本方差为$S^2=4.25$.试求σ^2的置信度为95%和的99%的置信区间。

解:置信度为95%时

$$P\left\{\frac{(n-1)S^2}{\chi^2_{1-0.025}}\le\sigma^2\le\frac{(n-1)S^2}{\chi^2_{0.025}}\right\}=1-0.05$$

查表得:$\chi^2_{0.975}(24)=39.4,\chi^2_{0.025}(24)=12.4;$

又:$\dfrac{(25-1)\times4.25}{39.4}=2.59,\dfrac{(25-1)\times4.25}{12.4}=8.23$

σ^2的置信区间为$(2.59,8.23)$

总结:

1. **正态总体,未知 σ^2, 小样本,求总体均值的置信区间**

$$(\bar{X}-t_{\alpha/2}\frac{s}{\sqrt{n}},\bar{X}+t_{\alpha/2}\frac{s}{\sqrt{n}})$$

2. **大样本的总体均值置信区间**

$$(\bar{X}-z_{\alpha/2}\frac{s}{\sqrt{n}},\bar{X}+z_{\alpha/2}\frac{s}{\sqrt{n}})$$

3. **总体均值置信区间:正态总体,σ^2已知**

$$(\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

4. 总体比率的置信区间 (大样本)

$$(\hat{p}-z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n},\hat{p}+z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n})$$

5. 总体方差的置信区间

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)},\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right)$$

置信区间的意义:估计抽样误差

置信度:

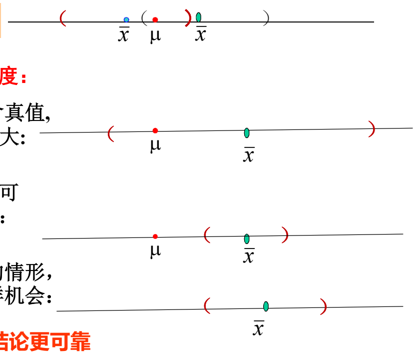
$(1-\alpha)=95\%$

置信区间的宽度:

过宽,虽然包含真值,但抽样误差过大:

置信区间也有可能不覆盖真值:

实际工作时的情形,只有一次抽样机会:



置信度高,则结论更可靠

关于置信区间的注意点

例题1: Gallup公司(1991)就消费者对美国产品质量的看法,对美国、德国、日本的消费者分别进行调查,结果表明:有55%的美国人相信美国产品的质量非常好,而持有同样看法的德国人和日本人的比例分别是26%和17%。美联社在报道这项调查结果时曾经提到"抽样误差在正、负三个百分点"。

(1)报道中"抽样误差在正、负三个百分点"这句话的含义?

(2)这个报道全面吗?还应该补充什么信息,为什么?

在正式报道中,除了估计值和抽样误差以外,还应该包含有关置信度的信息才是全面的。

置信区间的内涵:区间 ⊕ 置信度

例题2:一项有10000个人回答调查,同意某种观点的人的比例为70%(有7000人同意),可以算出总体中同意该观点的比例的95%置信区间为(0.691,0.709);

另一个调查者调查了50个人。他声称有70%的比例反对该种观点,并说总体中反对该观点的置信区间也是(0.691,0.709);

来自现实世界的数据量越大，我们对现实世界的了解就越清楚

例题3：如果在置信度不变的情况下，你要使目前所得到的置信区间的长度减少一半，样本量应增加到目前样本容量的多少倍？如果保持置信区间的长度不变，样本容量的增加会使什么发生变化？

由于：$w = z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} \Rightarrow \boxed{n = \dfrac{z_{\alpha/2}^2\sigma^2}{w^2}}$

因此
$$n_1 = \frac{z_{\alpha/2}^2\sigma^2}{\left(\frac{1}{2}w\right)^2} = 4\times\frac{z_{\alpha/2}^2\sigma^2}{w^2} = 4n$$

样本量应增加到目前样本量的4倍。

如果保持置信区间的长度不变，样本量的增加会使置信度增加。

扩展练习

一、 两个正态总体$N(\mu_1,\sigma_1^2), N(\mu_2,\sigma_2^2)$的情形

$X_1,X_2,\cdots,X_{n_1}$来自$N(\mu_1,\sigma_1^2)$，$Y_1,Y_2,\cdots,Y_{n_2}$来自$N(\mu_2,\sigma_2^2)$，

$\bar{X}=\frac{1}{n_1}\sum_{i=1}^{n_1}X_i, \bar{Y}=\frac{1}{n_2}\sum_{j=1}^{n_2}Y_j$，$S_1^2$和$S_2^2$分别为第一,二个总体的样本方差,置信度为 $1-\alpha$.

1. $\mu_1-\mu_2$的置信区间

(1) σ_1^2,σ_2^2已知时

由 $\bar{X}-\bar{Y}\sim N\left(\mu_1-\mu_2,\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}\right)$

有 $\dfrac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}\sim N(0,1)$

置信区间为：$\left[(\bar{X}-\bar{Y})\pm Z_{\frac{\alpha}{2}}\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}\right]$

(2) $\sigma_1^2=\sigma_2^2=\sigma^2, \sigma^2$未知

此时由第六章定理6.8, $\dfrac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim t(n_1+n_2-2)$

置信区间为：$\left[(\bar{X}-\bar{Y})\pm t_{\frac{\alpha}{2}}(n_1+n_2-2)S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}\right]$

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

二、 $\dfrac{\sigma_1^2}{\sigma_2^2}$的置信区间　设$\mu_1,\mu_2$未知

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

有 $P\left\{F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)<\dfrac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2}<F_{\frac{\alpha}{2}}(n_1-1,n_2-1)\right\}=1-\alpha$

即 $P\left\{\dfrac{S_1^2}{S_2^2}\dfrac{1}{F_{\frac{\alpha}{2}}(n_1-1,n_2-1)}<\dfrac{\sigma_1^2}{\sigma_2^2}<\dfrac{S_1^2}{S_2^2}\dfrac{1}{F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)}\right\}=1-\alpha$

置信区间为：$\left[\dfrac{S_1^2}{S_2^2}\dfrac{1}{F_{\frac{\alpha}{2}}(n_1-1,n_2-1)},\dfrac{S_1^2}{S_2^2}\dfrac{1}{F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)}\right]$

例3.6：两台机床生产同一个型号的滚珠，从甲机床生产的滚珠中抽取8个，从乙机床生产的滚珠中抽取9个，测得这些滚珠得直径(毫米)如下：

甲机床 15.0　14.8　15.2　15.4　14.9　15.1　15.2　14.8

乙机床 15.2　15.0　14.8　15.1　14.6　14.8　15.1　14.5　15.0

设两机床生产的滚珠直径分别为X,Y,且 $X\sim N(\mu_1,\sigma_1^2), Y\sim N(\mu_2,\sigma_2^2)$

(1) $\sigma_1=0.18,\sigma_2=0.24$,求$\mu_1-\mu_2$的置信度为0.90的置信区间；

(2) 若$\sigma_1=\sigma_2=\sigma$未知，求$\mu_1-\mu_2$的置信度为0.90的置信区间；

(3) 若μ_1,μ_2未知，求$\dfrac{\sigma_1^2}{\sigma_2^2}$的置信度为0.90的置信区间；

解：$n_1=8$, $\bar{x}=15.05$, $S_1^2=0.0457$; $n_2=9$, $\bar{y}=14.9$, $S_2^2=0.0575$

(1) 当$\sigma_1=0.18,\sigma_2=0.24$时,求$\mu_1-\mu_2$的置信度为0.90的置信区间为：

$$\left(\bar{X}-\bar{Y}\pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$$

查表得：$Z_{0.05}=1.645$,从而所求区间为$(-0.018,0.318)$

(2) 当$\sigma_1=\sigma_2=\sigma$未知时，$\mu_1-\mu_2$的置信度为0.90的置信区间为：

$$\left(\bar{X}-\bar{Y}\pm t_{\alpha/2}(n_1+n_2-2)S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

$t_{0.05}(15)=1.7531, S_W=0.228, \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}=0.486$

从而所求区间为$(-0.044,0.344)$

(3) 当μ_1,μ_2未知时，$\frac{\sigma_1^2}{\sigma_2^2}$的置信度为0.90的置信区间为：

$$\left\{\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2}(n_1-1,n_2-1)},\frac{S_1^2}{S_2^2}\frac{1}{F_{1-\alpha/2}(n_1-1,n_2-1)}\right\}$$

由$F_{0.05}(7,8)=3.50, F_{0.95}(7,8)=\frac{1}{F_{0.05}(8,7)}=\frac{1}{3.73}=0.268$

得$\frac{\sigma_1^2}{\sigma_2^2}$的置信度为0.90的置信区间为$(0.227,2.965)$

第四讲　抽样调查

常见的抽样方法

（1）简单随机抽样

对北航学生的研究能力进行抽样测试。在北航全校学生中随机抽取 ***n*** 名学生。

（2）分层抽样

分层次抽样：专科、本科、硕士生、博士、博士后。

（3）整群抽样

在本科生中，随机抽取若干个班，观察每个班的全部学生。

（4）分段抽样

全国调查，随机抽取若干省，再随机抽取若干市，再随机抽取若干区，...

（5）非随机抽样

在临沂小商品市场抽样，询问进货地点。编制抽样框很困难。

4.1 简单随机抽样方法

简单随机抽样：

每一个容量为 ***n*** 的可能样本被抽到的概率都是一样的。

原则：调查者不能根据主观意图挑选调查单位。而是在总体中，按照随机原则和纯粹偶然性的方法抽取样本。

方法：（1）抽签法

（2）随机数字表，随机数发生器

抽签法：先将调查总体的每个单位编上号码，然后将号码写在卡片上搅拌均匀，任意从中选取。抽到一个号码，就对上一个单位，直到抽足预先规定的样本数目为止。

$$\text{简单随机抽样}\begin{cases}\text{无放回抽样}\begin{cases}\text{有限总体}N\\\text{无限总体}N\to+\infty\end{cases}\\\text{放回抽样}\end{cases}$$

优点：　可以获得一个无偏倚的样本

使用限制：实施操作并不简单

（1）保证样本点分布均匀；

（2）有时,调查人员要了解所有样本中的个体有时是很困难的。

（3）样本容量较小时，一些比例少但是很重要的个体不能入样，使样本的代表性受到影响。

例如：在人民银行随机抽取100名职员，可能会抽不到高层管理人员。

TBT调查在全国抽1000家企业，可能会有许多大型企业不能入样。

(1) 总体均值的估计

● **放回抽样**

总体均值的点估计

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

$$D(\bar{x})=\frac{\sigma^2}{n}$$

总体均值的区间估计（抽样误差）

$$\left(\bar{x}-t_{\alpha/2}\frac{s}{\sqrt{n}},\bar{x}+t_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

● **不放回抽样**

总体均值的点估计 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

N—总体中的个体数量

n—样本容量

$$D(\bar{x})=\boxed{\frac{N-n}{N-1}}\cdot\frac{\sigma^2}{n}$$

$\frac{N-n}{N-1}$称为有限总体的“修正系数”。

当$N\to\infty,\frac{N-n}{N-1}\to1$。

注意：$\frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}<\frac{\sigma^2}{n}$　同样样本容量下，不放回抽样的误差更小！

总体均值的区间估计

[自由度 $df=(n-1)$]

$$\bar{x}\pm t_{\alpha/2}\sqrt{\frac{N-n}{N-1}}\frac{s}{\sqrt{n}}$$

例：

某居民区共有 $N=200$户居民，随机抽取 $n=20$位居民，他们每日收看电视的时间如下：

60　90　100　30　90　60　180　80　70　90

180　120　30　60　90　120　80　80　100　90

求该居民区居民平均每日收看电视时间的点估计和区间估计；

求该居民区居民平均每日收看电视时间的点估计和区间估计；

$$\bar{x}=\frac{1}{20}[60+90+100+\cdots+90]=90\text{（分钟）}$$

$$s^2=\frac{1}{20-1}\left[(60-90)^2+(90-90)^2+\cdots+(90-90)^2\right]=1515.7895$$

$$s=\sqrt{1515.7895}=38.93$$

取$\alpha=0.05\Rightarrow t_{\alpha/2}(19)=2.093$

区间估计：$90\pm2.093\sqrt{\frac{200-20}{200-1}}\cdot\frac{38.93}{\sqrt{20}}\approx90\pm17$

相对误差为：　17 / 90 = 19%　（显然，样本容量不够大）

(2) 总体比例的估计 (大样本）

● **放回抽样**

总体比例：$p=\frac{A}{N}$

样本比例：$\hat{p}=\frac{a}{n}$　(点估计）

$$E(\hat{p})=p,\qquad D(\hat{p})=\frac{1}{n}p(1-p)$$

区间估计：$\hat{p}\pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

● **不放回抽样**

(1)点估计：$\hat{p}=\frac{a}{n}$

$$E(\hat{p})=p,\qquad D(\hat{p})=\boxed{\frac{N-n}{N-1}}\cdot\frac{p(1-p)}{n}$$

其中：$\frac{N-n}{N-1}$称为“修正系数”

(2)区间估计：$\hat{p}\pm z_{\alpha/2}\sqrt{\frac{N-n}{N-1}}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

例题：某城市想要估计下岗职工中女性所占的比例，随机抽取了100名下岗职工，其中65人为女性。试估计该城市下岗职工中女性比例，并指出估计误差。置信水平要求为95%。

已知 $n=100$，$\alpha=0.05$，$z_{\alpha/2}=1.96$ $\hat{p}=65\%$

放回抽样的置信区间为：$0.65\pm1.96\sqrt{\frac{0.65\times0.35}{100}}=65\%\pm9.35\%$

不放回抽样的置信区间半长：$N\to\infty$

$$z_{\alpha/2}\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\approx1.96\sqrt{\frac{0.65\times0.35}{100}}=9.35\%$$

4.2 样本容量的确定

问题：估计某地区的平均收入 $D=z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

假若已知： $\sigma=4000$ ¥

希望抽样误差 D= $|\bar{x}-\mu|\le500$

并且要求置信度为（1- α）= 0.95

问： 样本容量应该多大 ？

95% C. I.是 $\_(\underline{\qquad\quad}\underset{\bar{x}}{.}\leftarrow D\rightarrow)\_$ $\underset{\mu}{\uparrow}$

$$(\bar{x}-1.96\frac{4000}{\sqrt{n}},\bar{x}+1.96\frac{4000}{\sqrt{n}})$$

要求 $D=|\bar{x}-\mu|\le500$

则: $D=1.96\frac{4000}{\sqrt{n}}\le500$

$$n=\frac{1.96^2(4000)^2}{500^2}=245.86$$

样本容量应不少于 246 人。

1、估计总体均值时需要的样本容量

放回抽样

在构造总体均值 μ 的置信度为 100(1-α)的置信区间时（总体方差已知）

$$(\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

置信区间的半长 D 等于

$$D=\frac{z_{\alpha/2}\sigma}{\sqrt{n}}\Rightarrow\sqrt{n}=\frac{z_{\alpha/2}\sigma}{D}\qquad n=\frac{z_{\alpha/2}^2\sigma^2}{D^2}$$

例题： $n=\frac{z_{\alpha/2}^2\sigma^2}{D^2}$

某厨具代理商欲了解其长期用户每月平均购买支出额。问至少要抽取多大容量的样本，才能使样本均值与总体均值的绝对误差在置信度不低于95%的条件下小于1？

问题1.总体标准差 σ 在抽样之前未知！

问题2. 在未确定样本容量 n 之前，无法计算样本标准差！

预抽样：

先在该公司固定用户中随机抽取 n=30 的样本， 经计算得到：$\bar{x}=110$, $s=13.12$

95% C.I.

$$(110-1.96\frac{13.12}{\sqrt{30}},\ 110+1.96\frac{13.2}{\sqrt{30}})$$

$$=(110-4.7,\ 110+4.7)$$

精度不够（要求误差为110±1）：$D=1$

$$n=\left(\frac{1.96\times13.12}{1}\right)^2=661$$

如何确定调查所需要的精度 *D*

$$n=\frac{4s^2}{D^2}$$

$\bar{x}=100,\quad D=10$
$\bar{x}=1000,\quad D=10$

应用时，由于存在量纲问题，可以采用相对误差：

$$\frac{D}{\bar{x}}=r \quad \Rightarrow \quad D=r\cdot\bar{x}$$

$\bar{x}=100,\quad r=5\%,\quad D=r\times\bar{x}=5$
$\bar{x}=1000,\quad r=5\%,\quad D=r\times\bar{x}=50$

所以常用的方法是：$n=\dfrac{4s^2}{(r\cdot\bar{x})^2}$

不放回抽样

置信区间：$\left(\bar{x}-t_{\alpha/2}\sqrt{\frac{N-n}{N-1}}\cdot\frac{s}{\sqrt{n}},\bar{x}+t_{\alpha/2}\sqrt{\frac{N-n}{N-1}}\cdot\frac{s}{\sqrt{n}}\right)$

抽样误差范围：$D=\sqrt{\frac{N-n}{N-1}}\cdot\frac{t_{\alpha/2}s}{\sqrt{n}}$

要求样本容量为：

$$n\approx\frac{n_0}{1+n_0/N}<n_0$$

例：　假如固定用户：$N=2000$

$$n_0=\left(\frac{1.96\times13.12}{1}\right)^2=661$$

$$n\approx\frac{n_0}{1+n_0/N}=\frac{661}{1+661/2000}=496.8\approx497$$

注：有时为计算方便起见，常取简单随机抽样所需要的样本容量代替n。这是一种保守的做法，但计算简单，在实际调查中经常使用。

2、估计总体比率时需要的样本容量

放回抽样

置信度为 (1-α)，总体比率 p 的置信区间为

$$\left(\hat{p}-z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n},\hat{p}+z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}\right)$$

置信区间的宽度为

$$D=z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}\Rightarrow\sqrt{n}=\frac{z_{\alpha/2}\sqrt{\hat{p}\hat{q}}}{D}$$

样本容量为 $n=\dfrac{z_{\alpha/2}^2\hat{p}\hat{q}}{D^2}$

问题：在调查之前 $\hat{p}$ 是未知的

解决的办法：

$\hat{p}$	1	0.1	0.2	0.3	0.4	0.5	0.6 ……
$1-\hat{p}$	0	0.9	0.8	0.7	0.6	0.5	0.4 ……

取　$\hat{p}=0.5$　　$1-\hat{p}=0.5$

所以样本容量 n 的最大值是：

$$n=\frac{0.25z_{\alpha/2}^2}{D^2}$$

注意教材P179：记置信区间长度 d=2D

$$n=\frac{\frac{1}{4}z_{\alpha/2}^2}{\left(\frac{1}{2}d\right)^2}=\left(\frac{z_{\alpha/2}}{d}\right)^2$$

例题：

北京地区观众调查网的置信度要求90%,误差要求不超过3%。求所需要的样本容量。

解：(1-α) = 0.90，$z_{\alpha/2}=1.65$，D=0.03

$$n=\frac{0.25\times1.65^2}{0.03^2}=756(\text{人})$$

不放回抽样：

$$n\approx\frac{n_0}{1+n_0/N}<n_0$$

例：2009年3月，有政协委员提出恢复繁体字的提案。为了广泛了解民意，需要对该提案的支持率进行估计。

（1）要求置信度为0.95，置信区间长度$\boxed{\text{不超过}0.01}$，应抽取多少人？

抽样误差为0.5%

（2）如果随机抽取了4万人，其中有5600人支持该提案，计算支持率的置信区间，置信度为0.95。

解：

（1） $n=\left(\dfrac{1.96}{0.01}\right)^2=38416$　　放回抽样，至少抽取38416人

(2) $\hat{p}=\dfrac{5600}{40000}=0.14$　　14%的人表示赞同

$$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}=1.96\times\sqrt{\frac{0.14\times0.86}{40000}}=0.0034$$

置信区间： $[0.1366,\ 0.1434]$　　抽样误差为0.34%

4.3　系统抽样

又称“等距抽样”或“机械抽样”

特点：组织形式简单：不需要在抽样前对每一个单位进行编号。只要确定抽样起点和间隔，就可以确定整个样本单位。

（1）按照无关标志排队，按间隔抽取

例如：调查某企业职工收入时，按照姓氏比画排列职工名单，进行抽样。显然，职工工资与姓氏比画之间没有必然联系；

（2）按照有关标志排队，按间隔抽取

例如：进行农产品量调查时，将总体单位按照上一年度的产量高低排序。这样，可以使标志值高低不同的单位均进入样本，样本单位在总体中分布均匀，抽样误差较小。

（3）按照自然位置顺序排列，按间隔抽取

例如：工业产品检验时，按照生产时间顺序，每间隔一定时间抽取一定数量的样本；检验一打发票时，可以按照顺序，每隔10张抽取1张；在估计果园的产量时，每隔7株抽取1株。

方法：随机起点，等距抽取。

（1）按照某种顺序给总体中的N个单元排列编号；

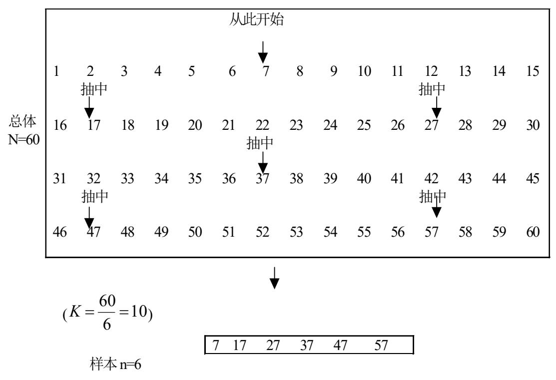
（2）按照随机数表，随机抽取一个编号 i 作为样本的第一个单元；

（3）计算间距：

$$k=\left[\frac{N}{n}\right]$$

(4)起始的样本点编号选取1~k之间的随机数。然后依次抽取编号如下的 n 个单元作为样本点。

$$i,i+k,i+2k,\cdots,i+(n-1)k$$



例如：中央电视台在建立收视率调查网时，要在某居委会拥有电视的512户中抽取5个样本户。

$$N=512,\quad n=5,\quad k=\frac{512}{5}\approx102$$

在[0,512]中任意确定一个三位数，例如是071。则被抽中的5户为：

71，173，275，377，479

抽样误差的大小与总体单位的排列顺序有关：

（1）如果总体中所有单元的排列编号是随机的，并且n 比N小得多的话，那么等距抽样的精度和简单随机抽样的精度是十分相近的。

（例如，按照姓氏比画或按照行政单位编号排序。）

（2）如果总体单元是按照某个与调查项目有关的变量的大小排序，由于等距抽样的样本点分布更加均匀，则等距抽样的精度高于简单随机抽样。

（例如，调查机械加工企业的工业增加值时，以用电量排序。）

（3）如果总体各单位的标志值存在周期变化趋势，而循环周期恰好等于抽样间隔，则等距抽样的精度低于简单随机抽样。

1,2,3,4,5,6; 1,2,3,4,5,6; 1,2,3,4,5,6; 1,2,3,4,5,6; 1,2,3,4,5,6

4.4　分层随机抽样

一、分层抽样方法

例如：

（1）对北航学生的研究能力进行抽样测试。学生层次有：专科、本科、研究生、博士、博士后。

（2）对央行的某项政策意见进行调查。可以根据调查内容分层：不同的职务层次，或者不同的部门、不同地区。

分层的原则： 在所调查的指标上，各层的相似程度高，而且层间差异大

例如 TBT影响调查：按照36个地区进行分层？（行政管理力度大）

按照22类出口地区分层？（受损情况类似）

分层抽样的特点：

采用分层抽样，使每一层内的差异大大缩小，而每一个样本单位对各层均有较高的代表性。

- 利用已知信息，提高抽样调查的精度；
- 便于组织实施；
- 在调查中，除了得到总体的有关信息外，还可以得到一些子总体的信息.

同样的样本容量下，分层抽样的抽样误差更小。

应用. TBT影响调查的分层方法： — 按照产品分层

— 按照地区管理

二. 总体均值的估计

例：

对某市600个个体商户的月零售额进行抽样调查，现申报资金分为大、中、小三类，根据调查结果的数据整理如下表。试估计该市个体户的平均月零售额，并以95%的可靠性作出区间估计。

层次	N_i	n_i	$\bar{y}_i$	s_i^2
大	60	30	20	16
中	240	40	8	4
小	300	40	1	0.5
总和	600	110		

$\hat{\bar{Y}} = \frac{1}{N}$ 

计算方法：

(1) 第 i 层样本均值： $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ ， $i = 1,2,\cdots,r$

(2)第 i 层总值估计： $\hat{Y}_i = N_i \bar{y}_i$

(3)总体平均值的估计： $\hat{\bar{Y}} = \frac{1}{N}\sum_{i=1}^{r} N_i \bar{y}_i = \sum_{i=1}^{r} W_i \bar{y}_i$

其中 $W_i = N_i / N$

总体均值 = 各层均值的加权和

方差估计： $\hat{\bar{Y}} = \sum_{i=1}^{r} W_i \bar{y}_i$

（1）放回抽样

$$D(\hat{\bar{Y}}) = \sum_{i=1}^{r} W_i^2 D(\bar{y}_i) = \sum_{i=1}^{r} W_i^2 \left(\sigma_i^2 / n_i\right)$$

（2）不放回抽样

$$D(\hat{\bar{Y}}) = \sum_{i=1}^{r} W_i^2 \frac{\sigma_i^2}{n_i}\cdot\frac{N_i - n_i}{N_i - 1} \approx \sum_{i=1}^{r} W_i^2 \cdot \frac{s_i^2}{n_i}(1 - f_i)$$

抽样比

其中： $f_i = \frac{n_i}{N_i}$ ， $\sigma_i^2 \approx s_i^2$

例题： 某市个体商户的月零售额的抽样调查

N=600
N_1=60
n_1=30
N_2=240
n_2 = 40
N_3=300
n_3=4 0

$f_1 = \frac{n_1}{N_1} = \frac{30}{60} = 0.5$ ， $\bar{y}_1 = 20$ 　 $s_1^2 = 16$

$f_2 = \frac{n_2}{N_2} = \frac{40}{240} = 0.17$ ， $\bar{y}_1 = 8$ 　 $s_1^2 = 4$

$f_3 = \frac{n_3}{N_3} = \frac{40}{300} = 0.13$ ， $\bar{y}_3 = 1$ 　 $s_3^2 = 0.5$

$$\hat{\bar{Y}} = \frac{1}{N}\sum_{i=1}^{3} N_i \bar{y}_i = \frac{1}{600}[60\times 20 + 240\times 8 + 300\times 1] = 5.7$$

$$D(\hat{\bar{Y}}) = \sum_{i=1}^{r} W_i^2 \cdot \frac{\sigma_i^2}{n_i}(1 - f_i)$$

$$\left[\left(\frac{60}{600}\right)^2 \frac{16}{30}(1-0.5) + \left(\frac{240}{600}\right)^2 \frac{4}{40}(1-0.17) + \left(\frac{300}{600}\right)^2 \frac{0.5}{40}(1-0.13)\right]$$

$= 0.0187$

区间估计： $5.7 \pm 1.96\times\sqrt{0.0187} \Rightarrow (5.43, 5.97)$ 　 $0.268 / 5.7 = 0.047$

三. 样本数目在层间的分配

问题：总的样本容量为 n , 总体分为 r 层。

每一层的样本容量应为多大?

（一）等比例分层抽样

1. 分配方案计算方法 I

在任意一层中，样本容量所占的比例都相同。

总体中的单位数：$N = N_1 + N_2 + \cdots N_r$

样本容量为： n

记： $f = \dfrac{n}{N}$

则第i层中的样本数目为： $\boxed{n_i = f \cdot N_i}$

例：N=1000, N_1=600, N_2=200, N_3=200

要抽取容量为 n=200 的样本，问每一层应抽取多少个体?

解：

$$f = \frac{n}{N} = \frac{200}{1000} = 0.2 \qquad \boxed{f_i = f \ , \ i = 1,2,3}$$

因此

$$n_1 = 0.2 \times 600 = 120$$
$$n_2 = 0.2 \times 200 = 40$$
$$n_1 = 0.2 \times 200 = 40$$

2.分配方案计算方法 II

记： $W_i = \dfrac{N_i}{N} \quad , \ i = 1,2,\cdots,r$

则： $\boxed{n_i = W_i \cdot n}$

$$\frac{N_i}{N} = W_i = \frac{n_i}{n}$$
$$n_i = \frac{n}{N} \cdot N_i = f \cdot N_i$$

例：N=1000, N_1=600, N_2=200, N_3=200

$W_1 = 0.6, \quad W_2 = 0.2, \quad W_3 = 0.2$

$n = 200:$

$$n_1 = 200 \times 0.6 = 120$$
$$n_2 = 200 \times 0.2 = 40$$
$$n_3 = 200 \times 0.2 = 40$$

---

3. 等比例分层抽样，总体均值的估计量

点估计：

$$\hat{\bar{Y}} = \sum_{i=1}^{r} W_i \cdot \bar{y}_i$$

区间估计：

（1）放回抽样的方差

$$D(\hat{\bar{Y}}) = \sum_{i=1}^{r} W_i^2 \frac{\sigma_i^2}{n_i} = \sum_{i=1}^{r} W_i \cdot \frac{\cancel{n_i}}{n} \cdot \frac{\sigma_i^2}{\cancel{n_i}}$$

$$= \frac{1}{n} \boxed{\sum_{i=1}^{r} W_i \ \sigma_i^2} = \bar{\sigma}^2 / n$$

其中， $\bar{\sigma}^2$ 表示平均层内方差。

（2）不放回抽样的方差

$$D(\hat{\bar{Y}}) = \sum_{i=1}^{r} W_i^2 \frac{\sigma_i^2}{n_i} (1 - f_i) = (1 - f) \sum_{i=1}^{r} W_i \cdot \frac{\cancel{n_i}}{n} \cdot \frac{\sigma_i^2}{\cancel{n_i}}$$

$$= \frac{1}{n} (1 - f) \boxed{\sum_{i=1}^{r} W_i \ \sigma_i^2} = (1 - f) \bar{\sigma}^2 / n$$

由于各层内的单元变化程度比较小，分层后有

$$\boxed{\sigma_i^2 < \sigma^2}$$

$$\bar{\sigma}^2 = \sum_{i=1}^{r} W_i \sigma_i^2 < \sum_{i=1}^{r} W_i \sigma^2 = \sigma^2 \sum_{i=1}^{r} W_i = \sigma^2$$

因此，同样的样本容量下，分层抽样的抽样误差更小。

四. 总体比例的估计

$$\hat{p} = \sum_{i=1}^{r} W_i \hat{p}_i$$

不放回抽样

$$D(\hat{p}) = \sum_{i=1}^{r} W_i^2 D(\hat{p}_i) = \sum_{i=1}^{r} W_i^2 \frac{N_i - n_i}{N_i - 1} \cdot \frac{p_i(1 - p_i)}{n_i}$$

$$= \sum_{i=1}^{r} W_i^2 (1 - f_i) \cdot \frac{p_i(1 - p_i)}{n_i}$$

例题： $\hat{p}=\sum_{i=1}^{r}W_i\hat{p}_i$; $D(\hat{p})=\sum_{i=1}^{r}W_i^2(1-f_i)\cdot\frac{p_i(1-p_i)}{n_i}$

某广告公司要了解电视广告的作用，拟在有关对象中调查看电视广告的比例。设对象分为三层：

N_1=155, N_2=62, N_3=93,

样本容量为40。采用等比例分层抽样，调查结果为：第一层看电视广告的比例为0.8，第二层的比例为0.25，第三层的比例为0.5。试以95%的可靠性，估计调查对象中收看电视广告比例的置信区间。

N=155 + 62 + 93 =310 $f=40/310=0.129$

n =40

$$W_1=\frac{155}{310}=0.5,\quad n_1=0.5\times 40=20$$

$$W_2=\frac{62}{310}=0.2,\quad n_2=0.2\times 40=8$$

$$W_3=\frac{93}{310}=0.3,\quad n_3=0.3\times 40=12$$

由调查结果：

$\hat{p}_i=0.8$ $\hat{p}_i=0.25$ $\hat{p}_i=0.5$

则有 $\hat{p}=0.5\times0.8+0.2\times0.25+0.3\times0.5=0.6$

方差估计：

$$D(\hat{p})=\sum_{i=1}^{3}W_i^2(1-f)\frac{\hat{p}_i(1-\hat{p}_i)}{n_i}$$

$$=0.5^2(1-0.129)\frac{0.8\times0.2}{20}+0.2^2(1-0.129)\frac{0.25\times0.75}{8}$$

$$+0.2^2\times(1-0.129)\frac{0.5\times0.5}{12}=0.0042$$

$$s=\sqrt{0.0042}=0.065$$

所以,95%置信区间为

$$0.6\pm1.96\times0.065=0.6\pm0.1274$$

从总体看，观看广告的比例约为60%，估计误差约为 $\pm13\%$，估计的可靠性为95%。

4.5 抽样调查的误差来源

调查误差 = 抽样误差 + 非抽样误差

抽样误差：由于抽选样本的随机性而产生的误差

（由于概率抽样方式不同所造成，是可以估计的）

非抽样误差：除抽样误差外，由其他各种原因而引起的误差。

产生非抽样误差的主要原因：

（1）抽样框误差：目标总体不等于抽样总体，如遗漏了有关单位，或包含了非目标单位；观测之间的复合连接；分层方案设计不当等。

（2）无应答误差：受调查人有意识不合作；无意识（由于客观原因无法接受调查，填写问卷时粗心）；

（3）计量误差：问卷设计不合理、调查指标含义不清、计量单位不标准，选择的统计量和推算方法不适当等。

案例1：《文学摘要》民意测验 抽样框≠目标总体

1936年美国总统选举

F.D. Roosevelt (罗斯福）任美国总统的第一任期届满(民主党)

A. Landon (兰登）Kansas州州长(共和党)

经济背景：国家正努力从大萧条中恢复，失业人数高达九百万人。

The literary Digest《文学摘要》进行民意测验，将问卷邮寄给一千万人，他们的名字和地址摘自电话簿或俱乐部会员名册。其中240万人寄回答案（回收率24%）。

预测结果：Roosevelt 43%，Landon 57%

竞选结果：Roosevelt 62%，Landon 38%

主要原因：选择偏倚——将一类人排除在样本框之外（当时四个家庭中，只有一家安装电话）

不回答偏倚——低收入和高收入的人倾向不回答

1936年美国总统竞选（Gallup的预测）

✪样本容量3000人，在《摘要》公布其预测结果之前，仅以一个百分位数的误差预言了《摘要》的预测结果。

✪利用一个约5万人的样本，正确地预测了Roosevelt的胜利。

	Roosevelt的百分数
盖洛普预言《摘要》的预测结果	44
《摘要》预测的选举结果	43
盖洛普预测的选举结果	56
选举结果	62

方法： 从《摘要》要用的名单中随机选取3000人，并给他们每人寄去一张明信片，询问他们打算怎样投票。

大样本并不能防止偏倚：当抽样框不正确时，抽取一个大的样本并无帮助，它只不过是在较大的规模下，去重复基本错误。

Gallup1936~1948年采用定额抽样

定额抽样：样本被精心挑选，以使在某些关键特征上与总体相似。在规定定额内，访问人员可以自由选取任何人。

例如：在 St. Louis 的访问人员访问13个对象，并规定其中

- 6人住在近郊，7人住在市中心；
- 男人7名，女人6名；
- 在男人中，3人40岁以下，4人40岁以上；1名黑人，6名白人。
- 6名白人支付的月租：1人支付的金额不少于44.01$
  3人支付的金额为18.01~44.00 $
  2人支付的金额不超过18.00 $

年份	预测共和党得票	共和党实际得票	有利于共和党的偏差
1936	44	38	6
1940	48	45	3
1944	48	46	2
1948	50	45	5

Gallup民意测验在1948年后总统选举中的记录
（随机抽样：访问员无任何自主处理的权利）

年份	样本容量	获胜候选人	预测值	选举结果	误差
1952	5385	艾森豪威尔	51.0%	55.4%	+4.4%
1956	8144	艾森豪威尔	59.5%	57.8%	-1.7%
1960	8015	肯尼迪	51.0%	50.1%	-0.9%
1964	6625	约翰逊	64.0%	61.3%	-2.7%
1968	4414	尼克松	43.0%	43.5%	-0.5%
1972	3689	尼克松	62.0%	61.8%	-0.2%
1976	3439	卡特	49.5%	51.1%	+1.6%
1980	3500	里根	55.3%	51.6%	-3.7%
1984	3456	里根	59.0%	59.2%	-0.2%
1988	4089	布什	56.0%	53.9%	-2.1%

案例2　可口可乐问卷设计失败

问题与思考：

20世纪80年代，美国可口可乐公司耗资500万美元，进行了历时2年的市场调查，调查了近20万名消费者。决定放弃传统配方，推出一代新的可口可乐。却几乎产生灾难性的后果。

可口可乐发展将近百年。但在20世纪80年代，它的市场销售增长率从平均每年13%猛降到2%。市场占有率从曾是百事可乐的2倍，变成只领先2.9个百分点。

市场调查与决策：

（1）出动2000名调查员，在10个主要城市调查消费者的口味。**问卷的主要问题是：“如果在可口可乐配方中增加一种新的成分，使它喝起来更柔和，您愿意吗？**结果有一多半的人表示接受，只有11%的人表示不安。

（2）公司投资400万美元进行大规模的口味尝试活动。13个大城市的19.1万消费者参与口味尝试活动。在众多口味饮料中，消费者对新口味可乐青睐有加。55%的品尝者认为新口味超过传统配方。**结论：立即生产新可乐。**

（3）经过与全世界瓶装厂商量，并进行财务预算，公司决定：**用新可乐代替传统可乐，停止传统可乐的生产与销售。**

结果：

新饮料上市4个小时，可口可乐公司接到650个抗议电话。10天后，每天接到5000多个抗议电话。更有雪片似的抗议信件。有人甚至说要改喝茶水来代替可乐。公司不得不开辟83个热线，雇佣大量的公关人员来处理这些抱怨和抗议。

3个月以后，市场调研表明，只有不到30%的消费者说新可乐的好话了。愤怒的情绪在美国蔓延。社会学家认为，可口可乐公司把一个神圣的象征毁掉了。

罗伯特.戈伊朱埃塔不得不率领公司全体高层管理者站在可口可乐的标志下，向公众道歉，并宣布立即恢复传统配方生产。全国一片沸腾。有议员在参议会回上发表演说：“这是美国历史上一个非常有意义的时刻，它表明有些民族精神使不可更改的。”

问题的根源是什么？ **耗资巨大、范围广泛、被调查者反映良好**

（决策是：放弃老饮料）

其他案例：调查中的非抽样误差

1、分层抽样方案设计不当，造成选择偏倚：按产品分层（样本分配原则是出口额高的产品多抽；对于一个产品，根据其出口额在全国各地分布分配样本。）

问题：一些出口总额小的地区会不能入样。

2、样本点之间的复合连接，造成重复统计

例如：企业类型（生产型企业、流通型企业）

3、抽样框中包含非目标单位：若以上年企业出口额作为抽样依据；但该企业的受调查产品当年没有出口。减少有效样本数量

4、避免调查表中内容的歧异：“所调查的产品”→“本问卷所调查的产品”

“进口国”→“贸易对象国”；

5、加强调查人员的责任意识：采取登记制度和汇总结果的报告制度。

抽样调查作业

采用抽样调查方法，估计全班同学的平均身高

1、首先：计算总体均值和方差（留做参考）

2、预抽样（n=30）：估计样本均值与方差

3、选取抽样的相对误差 5%或10%

4、计算样本容量 $n_0=\dfrac{4s^2}{(r\cdot\bar{x})^2}$ $n\approx\dfrac{n_0}{1+n_0/N}<n_0$

5、等距抽样：随机起点，等距抽取

6、给出点估计和区间估计 $\bar{x}\pm 2\sqrt{\dfrac{N-n}{N-1}}\dfrac{s}{\sqrt{n}}$

7、对比总体参数，对于你的分析过程和结论进行评价与思考

阅读与练习

简单随机抽样总体总值的估计

分层抽样总体总值的估计

Excel 软件应用

一、简单随机抽样总体总值的估计

1. 例题：

某工厂欲了解工人由于停工待料及机器故障所造成的每周工时损失。全厂共有750人。从中抽取50个工人进行调查，得到每个工人平均每周的工时损失数为 $\bar{y}=10.31$ 小时，且 $s^2=2.25$ 。估计全厂由于停工待料及机器故障造成的工时损失数。（α=0.05）

已知：$N=750,\quad n=50\quad ,\bar{y}=10.31,\quad s^2=2.25$

求： 总体Y的点估计和区间估计。

2. 点估计方法

计算公式： $\hat{Y}=N\bar{y}=\frac{N}{n}\sum_{i=1}^{n}y_i$

问题：为什么要先求样本均值 $\bar{y}$，再求 $\hat{Y}=N\bar{y}$？
为什么不直接用公式： $\hat{Y}=\sum_{i=1}^{n}y_i$

答案：（1） $\sum_{i=1}^{n}y_i\neq\sum_{i=1}^{N}y_i$

（2） 样本均值的波动小于个别观测值y_i的波动。

$$D(\bar{y})=\sigma^2/n$$

例如，我们很可能从总体中抽取一个身高1.80的个体，但却不可能抽取一个身高平均值为 $\bar{y}=1.80$ 的10个人的样本。在样本中，高、中、矮个子互相平均后，对总体的概括性更强。

点估计：

$$\hat{Y}=N\bar{y}$$

区间估计：

$$\hat{Y}\pm t_{\alpha/2}\sqrt{D(\hat{Y})}$$

$$D(\hat{Y})=D(N\bar{y})=N^2D(\bar{y})\begin{cases}N^2\cdot\dfrac{\sigma^2}{n} & (\text{放回抽样})\\ N^2\cdot\dfrac{N-n}{N-1}\cdot\dfrac{\sigma^2}{n} & (\text{不放回})\end{cases}$$

由此可见，总值估计的抽样误差要比均值估计的抽样误差扩大 N 倍。但是相对误差不变。

例题：

已知：$N=750,\quad n=50\quad ,\bar{y}=10.31,\quad s^2=2.25$

求： 总体Y的点估计和区间估计。

$\hat{Y}=N\bar{y}=750\times10.31=7732.5$ （小时）

$$D(\hat{Y})=N^2\frac{N-n}{N-1}\cdot\frac{s^2}{n}=750^2\times\frac{750-50}{750-1}\times\frac{2.25}{50}=24470.52$$

（$1-\alpha$）置信区间：

$$7732.5\pm1.96\times\sqrt{24470.52}=7732.5\pm1.96\times156.43$$
$$=（7426.20,\quad 8039.40）$$

$$\frac{1.96\times156.43}{7732.5}=0.04$$

二、分层抽样总体总值的估计

1. 点估计 $\hat{Y}=N\hat{\bar{Y}}=N\sum_{i=1}^{r}\frac{N_i}{N}\bar{y}_i=\sum_{i=1}^{r}N_i\bar{y}_i$

2. 区间估计 $D(\hat{Y})=N^2D(\hat{\bar{Y}})$

(1)放回抽样：$D(\hat{Y})=N^2D(\hat{\bar{Y}})=N^2\sum_{i=1}^{r}W_i^2\frac{\sigma_i^2}{n_i}$

(2)不放回抽样：$D(\hat{Y})=N^2\sum_{i=1}^{r}W_i^2\frac{\sigma_i^2}{n_i}\cdot\frac{N_i-n_i}{N_i-1}$

应用时，取：$s_i^2\approx\sigma_i^2$

例题：某市个体户的月零售额的抽样调查，估计全市个体户**总的月销售额**。

根据前面计算：

$N=600\quad \hat{\bar{Y}}=5.7\quad D(\hat{\bar{Y}})=0.0187$

所以有：

$\hat{Y}=N\hat{\bar{Y}}=600\times 5.7=3420$ （千元）

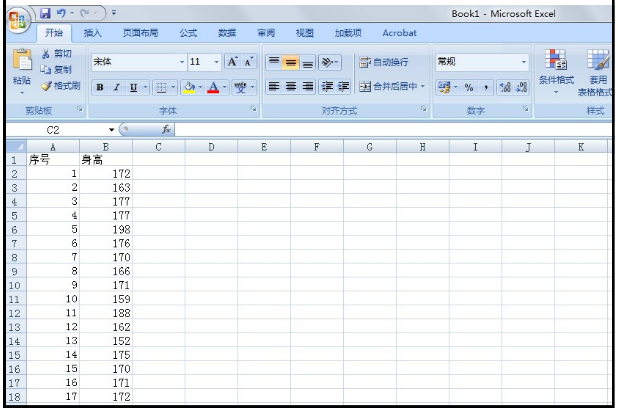
$D(\hat{Y})=N^2D(\hat{\bar{Y}})\quad =600^2\times 0.0187=6732$

置信区间： $3420\pm 1.96\times\sqrt{6732}=3420\pm 160.8156$

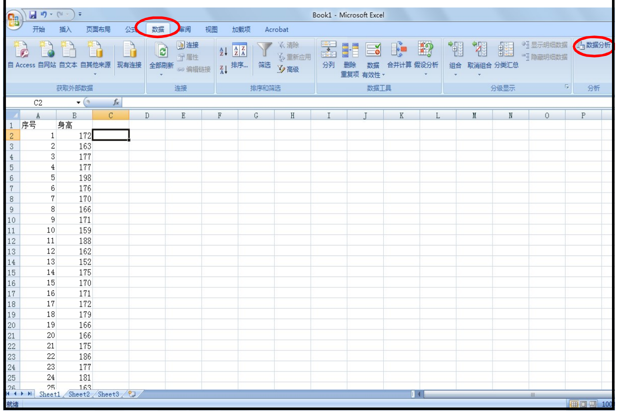
$(3259.184, 3580.816)$ 160.8156/3420=0.047

总值估计的与总体均值的相对误差不变。

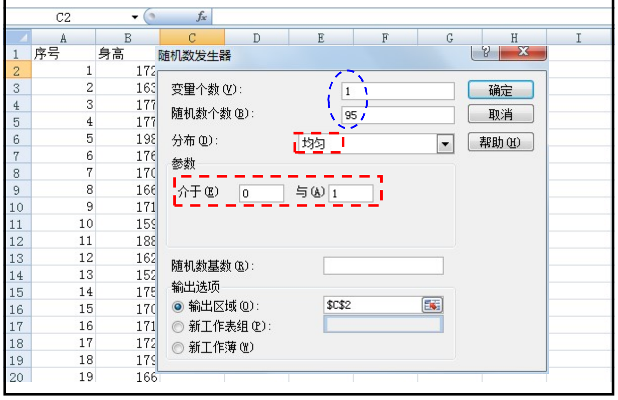
三、随机抽取30个样本点的方法

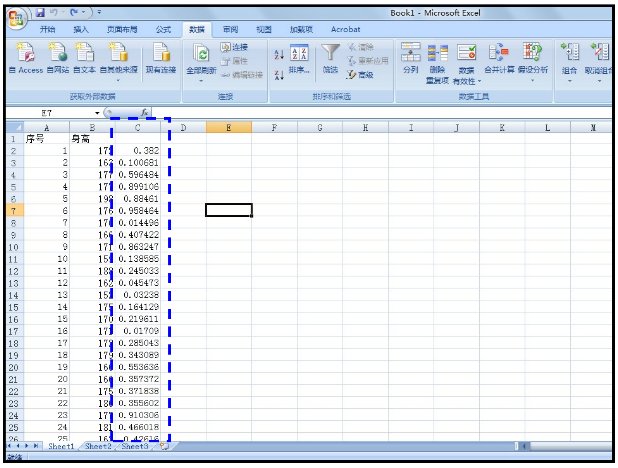


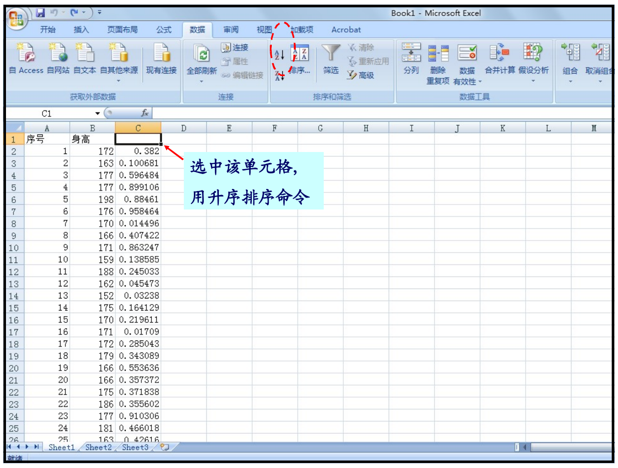
数据→数据分析

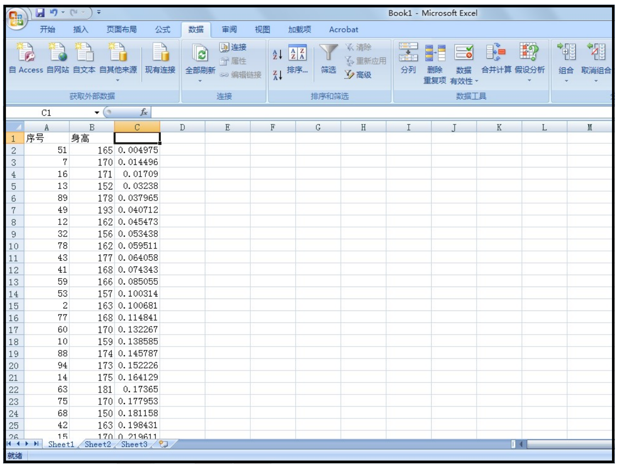


数据→数据分析→随机数发生器









序号	身高	
51	165	0.004975
7	170	0.014496
16	171	0.01709
13	152	0.03238
89	178	0.037965
49	193	0.040712
12	162	0.045473
32	156	0.053438
78	162	0.059511
43	177	0.064058
41	168	0.074343
59	166	0.085055
53	157	0.100314
2	163	0.100681
77	168	0.114841
60	170	0.132267
10	159	0.138585
88	174	0.145787
94	173	0.152226
14	175	0.164129
63	181	0.17365
75	170	0.177953
68	150	0.181158
42	163	0.198431
15	170	0.219611