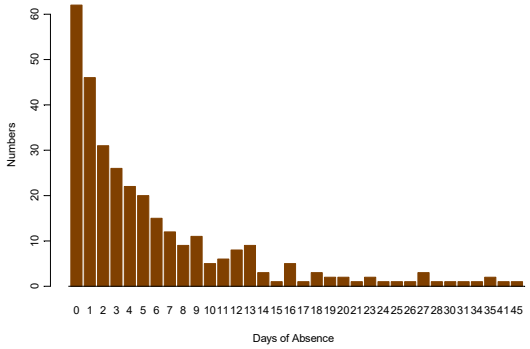
- 什么是统计学?

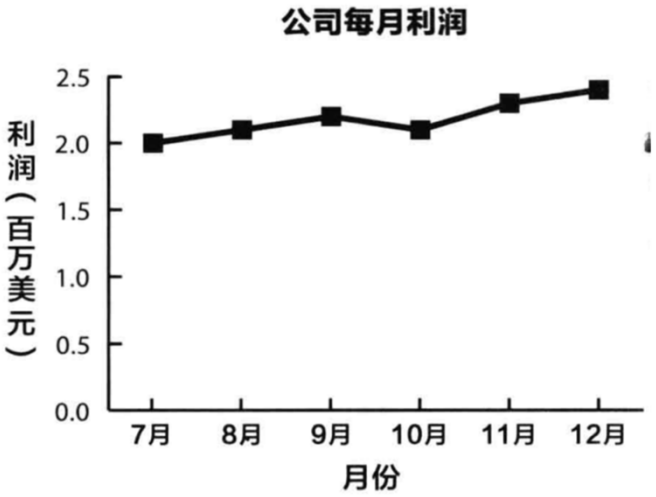
统计学是一门收集、整理和分析数据,然后运用概率知识从特定的数据中得出统计推论的科学。其目的是探索隐藏在数据里面的规律性，以达到对客观事物的科学认识

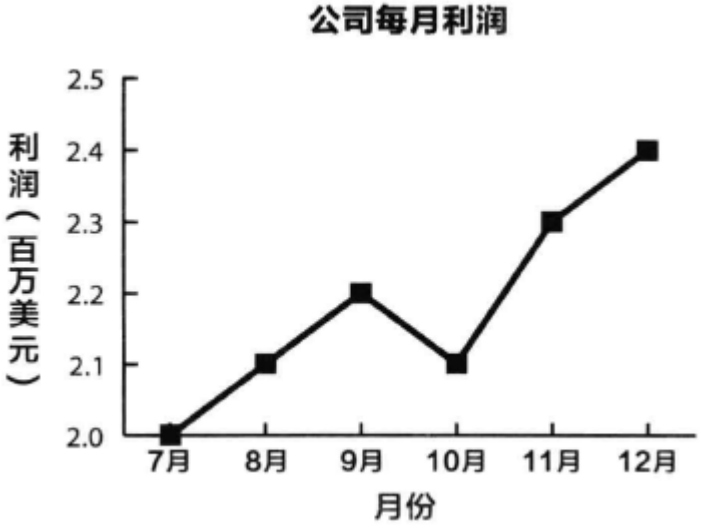
血糖

性别（1男2女）	平均值	数字	中位数	最大值(X)	最小值	方差	标准偏差
男性	6.5924	62	5.9600	13.38	4.21	4.232	2.05718
女性	5.9652	93	5.6200	17.94	4.35	3.935	1.98360
总计	6.2161	155	5.6900	17.94	4.21	4.122	2.03025




量化投资?
统计学
物联网?
互联网?
旅游交通?

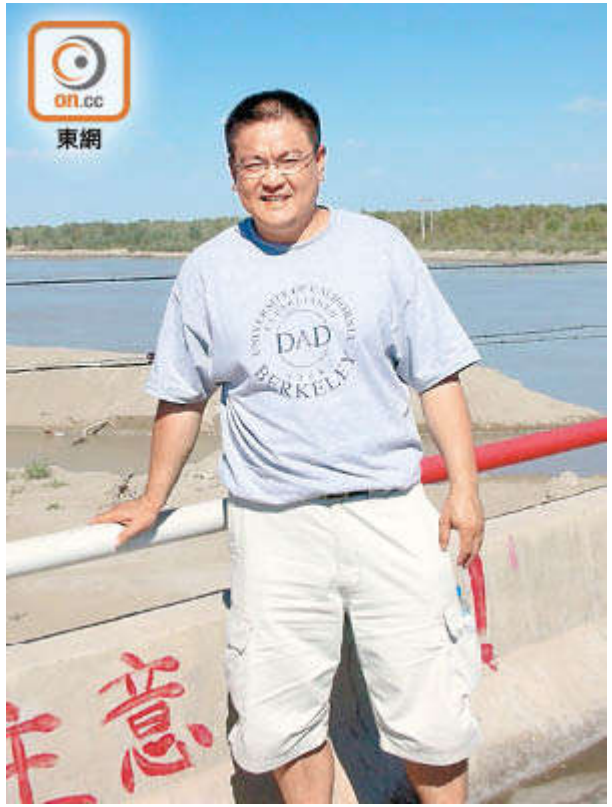

公司每月利润
利润（百万美元）
2.5
2.0
1.5
1.0
0.5
0.0
7月
8月
9月
10月
11月
12月
月份


公司每月利润
利润（百万美元）
2.5
2.4
2.3
2.2
2.1
2.0
7月
8月
9月
10月
11月
12月
月份

中大教授 必勝方程式 賭馬贏半億

東網Money18 即秒報價掌握大市走勢

將賭馬當作學術研究課題的中文大學統計學系教授顧鳴高，與一名外籍「金主」合夥，透過自己研發的方程式在港賭馬，該方程式近年愈贏愈多之際，顧鳴高忽然要求拆夥，觸發一場訴訟。顧的「金主」民事控告顧私下用該方程式賭馬，要求顧交代因此贏得的彩金，但顧只肯交出三個馬季的賭馬紀錄。高院昨下令馬會向顧的「金主」交出顧在另外四個馬季的賭馬紀錄，高院的判決書更透露，顧在三個馬季內已贏得五千六百萬元，其「得意弟子」雖不屬高薪一族，近年卻有錢與家人購入了七個物業。



- 本福特定律（Benford's Law）

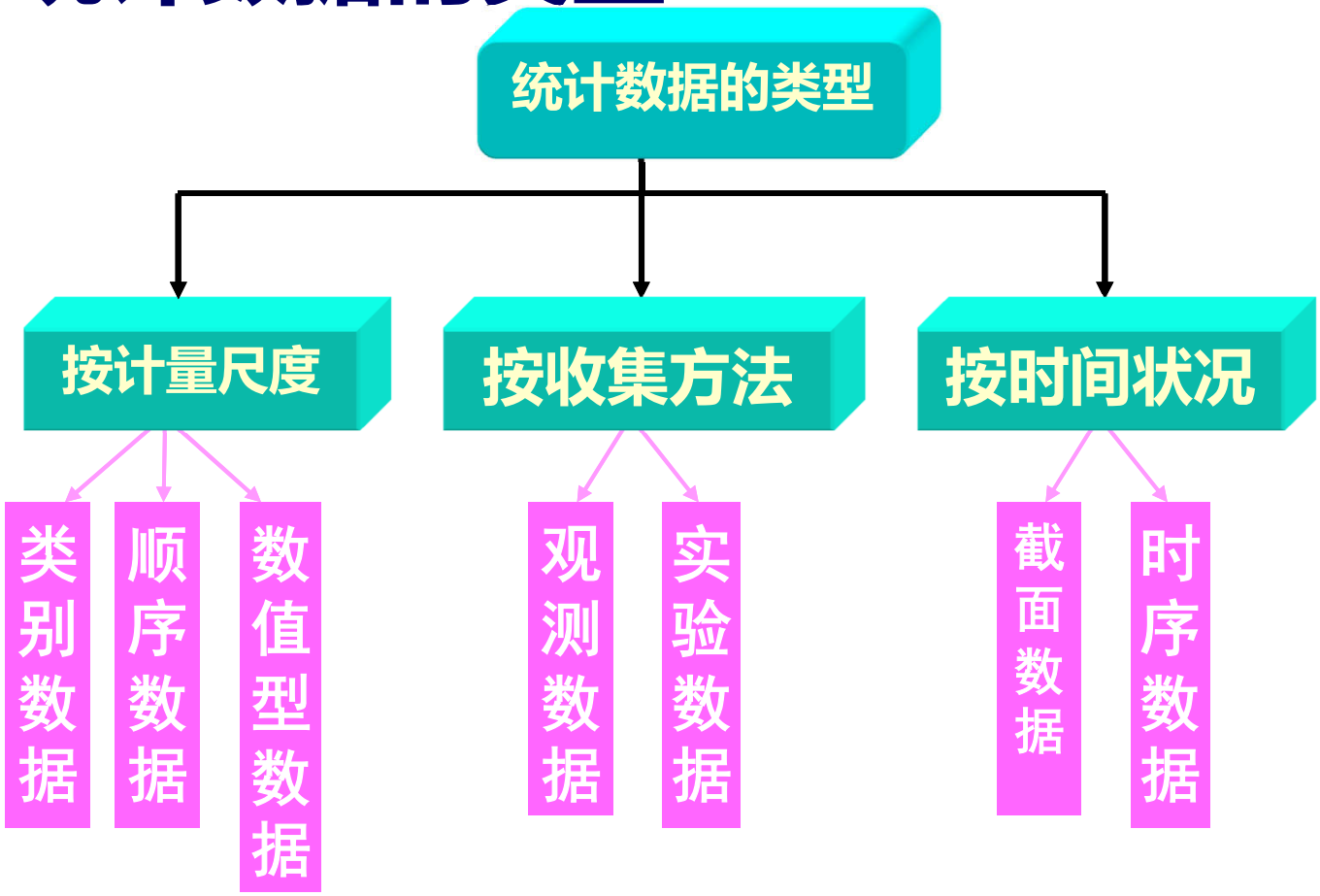
1	2	3	4	5	6	7	8	9
30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%

- 对于财务数据，数学家们还发 现，在那些假账中，数字5和6居然是最常见的打头数字。

- **什么是数据?**
  - **为表述和解释现实问题所收集 分析和汇总的事实依据与图表.**

例: 股票数据(数据文件shadow);客户资料;生产记录;库存记录;市场调查数据等等

统计数据的类型


统计数据的类型
按计量尺度
按收集方法
按时间状况
类别数据
顺序数据
数值型数据
观测数据
实验数据
截面数据
时序数据

数据的筛选

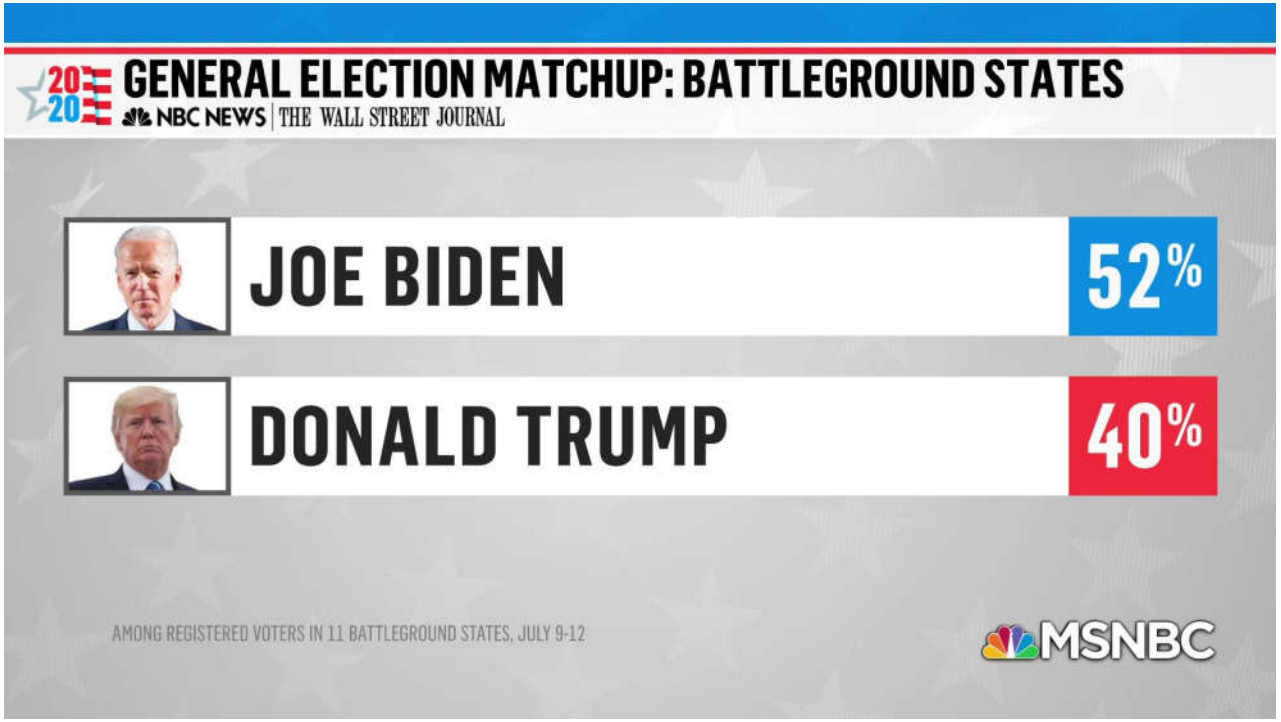
1. 对审核过程中发现的错误应尽可能予以纠正
2. 当发现数据中的错误不能予以纠正，或者有些数据不符合调查的要求而又无法弥补时，需要对数据进行筛选
3. 数据筛选的内容包括：
   - 将某些不符合要求的数据或有明显错误的数据予以剔除
   - 将符合某种特定条件的数据筛选出来

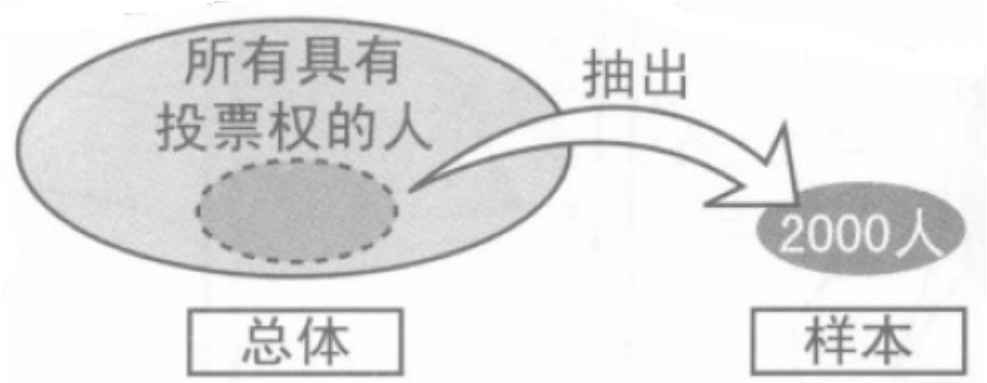
数理统计学是一门以数据为基础的学科.

数理统计学的任务就是如何获得样本和利用样本，从而对事物的某些未知方面进行分析、推断并作出一定的决策。

样本和总体

- 研究随机现象要先知道其概率分布
  - 然而现实中，随机变量服从的分布并不总是确定的
- 对被研究的随机变量进行大量的观察或者实验，其概率特征能够得到
  - 利用有限信息做出可靠的推断


2020
GENERAL ELECTION MATCHUP: BATTLEGROUND STATES
NBC NEWS | THE WALL STREET JOURNAL
JOE BIDEN
52%
DONALD TRUMP
40%
AMONG REGISTERED VOTERS IN 11 BATTLEGROUND STATES, JULY 9-12
MSNBC


所有具有
投票权的人
抽出
2000人
总体
样本

样本和总体

- 统计问题中，研究对象的全体所构成的集合称为总体
  - 西安交大所有学生
  - 某工厂每天生产的所有产品
- 构成总体的每一个元素称为个体
  - 个体取值的随机性
  - 聚合起来具有一定规律性
- 随机变量X代表总体，研究其分布F(x)

样本和总体

- 总体分布未知或是其某些参数未知，对总体中每一个个体进行观察，就可以完全了解总体的分布特征

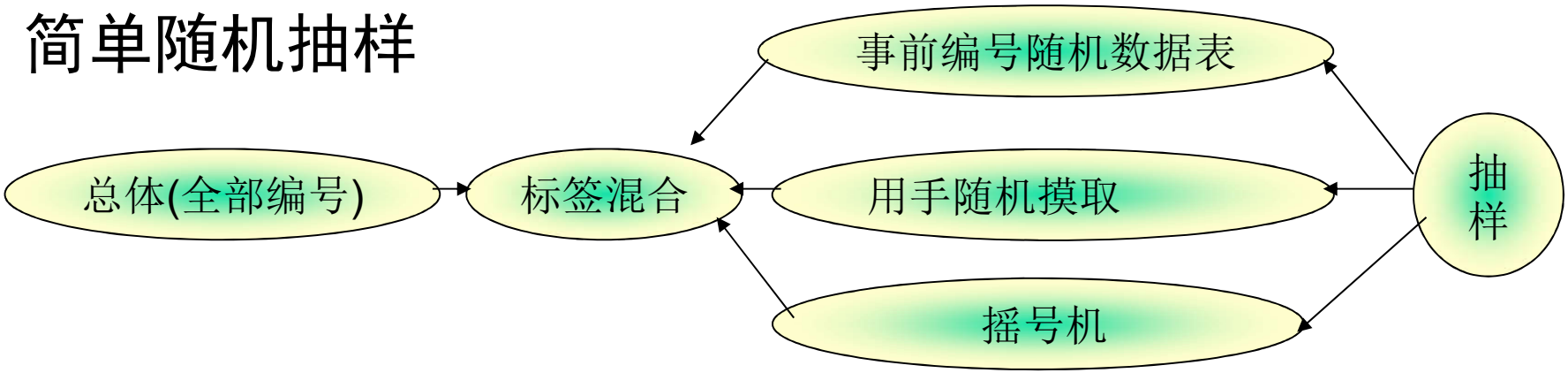
- 从总体中抽出的部分个体称为样本，样本中包含的个体的个数称为样本容量，按照一定规则取得样本的过程称为抽样，把观察或实验得到的数据称为样本观测值（观察值）

样本和总体

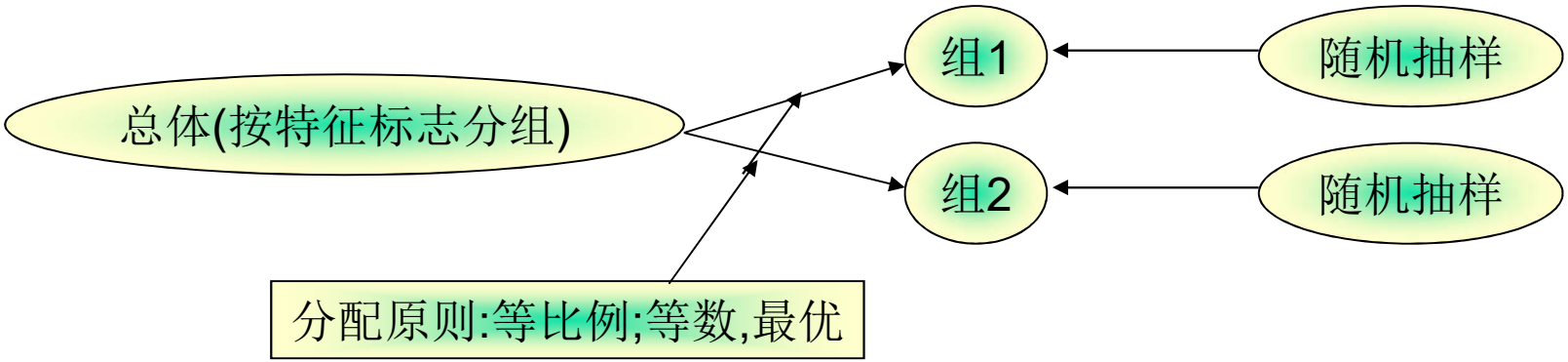
- 抽样的特点：
  - 随机原则：每个元素（或个体）有同等抽中的机会
  - 推断总体特征：样本的属性能推断出总体特征
  - 推断的精确性：把推断的误差控制在一定的精确度内

抽样设计

简单随机抽样

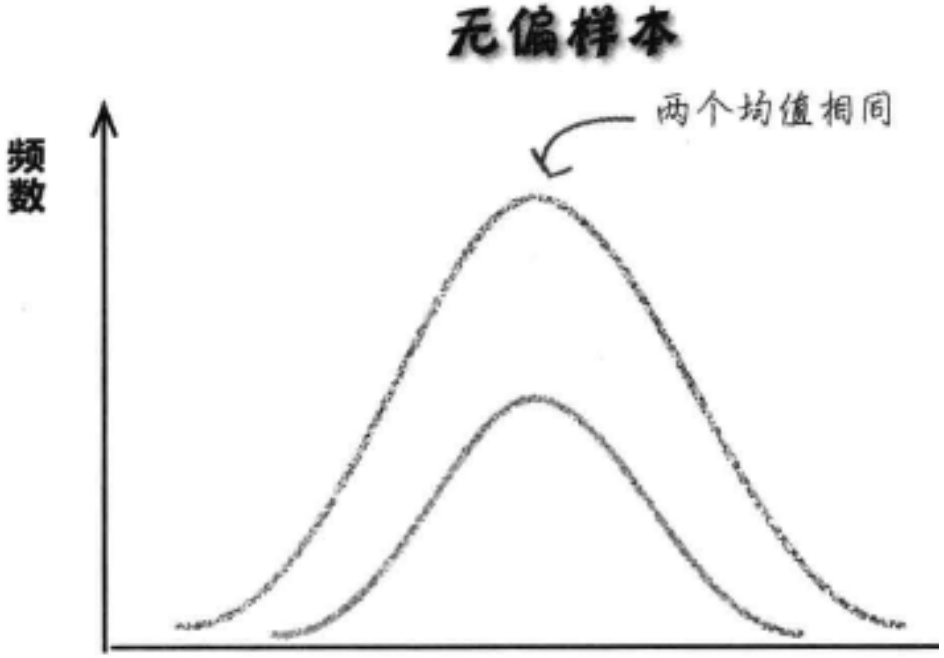


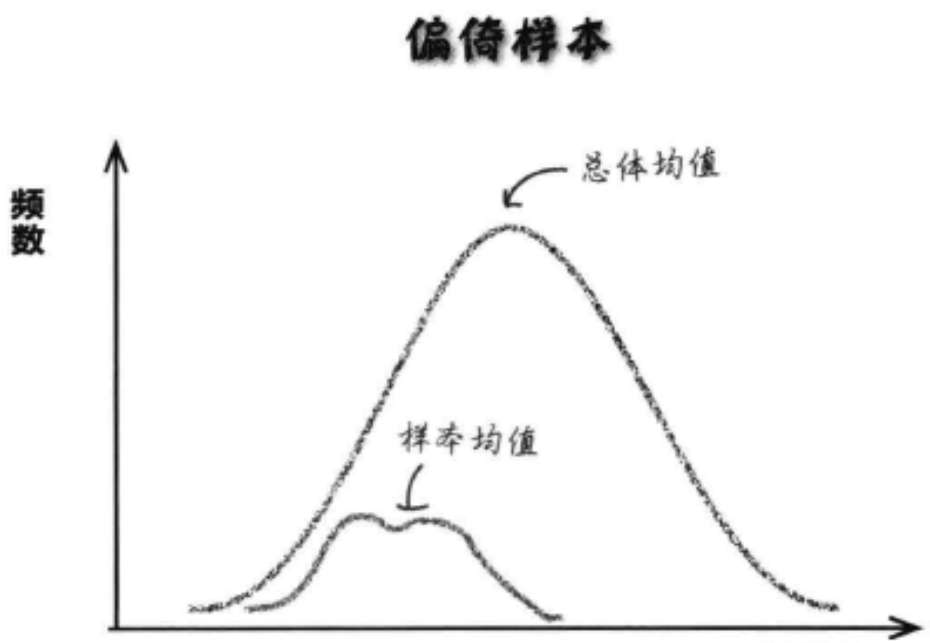
类型抽样(分层抽样或分类抽样)



注：等距抽样；多阶段抽样；双相抽样；穿插抽样(略)

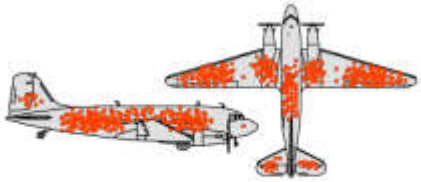
代表性抽样





幸存者偏差

- 二战时，军方为了提高飞机的防御，想在之前的机型上加强装甲。问题哪些部位需要加强装甲才最有效提高飞机的生存率？



Where to add armor?

- 弹孔稀疏的地方才最有可能是要害处，因为没怎么被击中才有机会返航

样本和总体

- 如果总体X的分布函数为F(x),则来自总体X的样本(X_1,X_2,…,X_n)的联合分布函数为

$$P(X_1 \le x_1, X_2 \le x_2, \dots, X_n \le x_n) = \prod_{i=1}^{n} P(X_i \le x_i) = \prod_{i=1}^{n} F(x_i)$$

- 离散变量用其点概率

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^{n} p(X = x_i)$$

- 连续随机变量用其概率密度函数

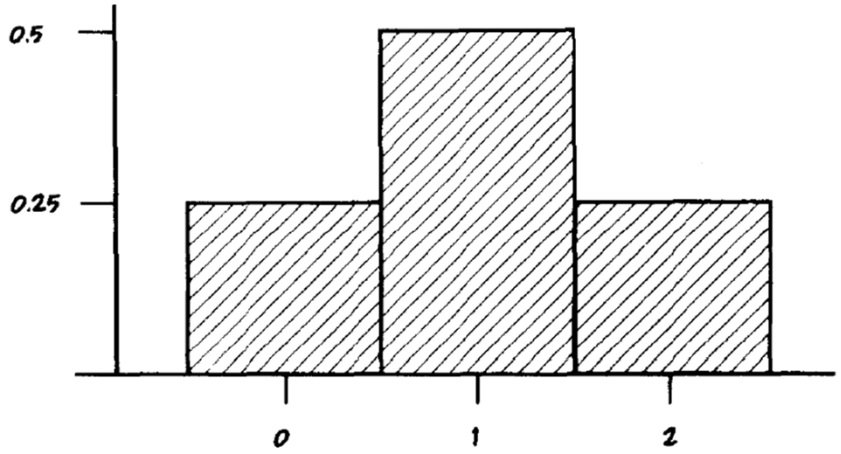
$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} f(x_i)$$

样本分布

- 刻画样本分布的三种形式：频数和频率分布、直方图、经验分布函数
- 对于样本值$(x_1,x_2,\ldots,x_n)$，样本频数分布是指样本值中不同数值在样本值中出现的次数，样本频率分布是指样本值中不同数值在样本值中出现的频率(频数除以样本容量)
- 对于频数$m_1,m_2,\ldots,m_l$，有

$$\sum_{i=1}^{l} m_i = n$$

- 想象一个随机试验，例如投掷两枚硬币并记录正面出现的次数，用X表示，则X就是随机变量

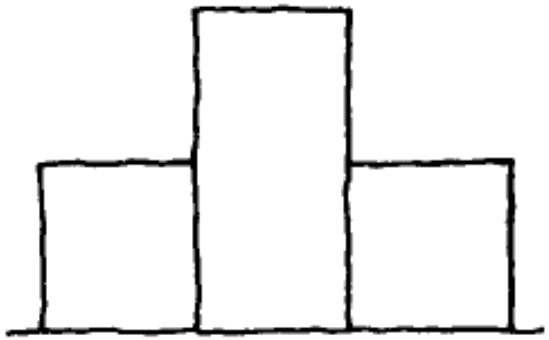


X	0	1	2
Pr(X=x)	1/4	1/2	1/4

- 上述表格称为随机变量X的概率分布

- 将两枚硬币试验重复1000次，得到

p(x)	x*	频数(m_i)	频率(m_i/n)
0.25	0	260	0.260
0.5	1	517	0.517
0.25	2	223	0.223



样本分布

- 对于离散型随机变量$P(X = x_i^*) = p_i$，由大数定律（Kolmogorov large number law）

  当n很大时，事件$X = x_i^*$的频率m_i/n趋近于概率p_i

- 对于连续型随机变量，事件$X = x_i^*$的概率为0，how?

样本分布

- 某轧钢厂生产了一批钢材，为了研究这批钢材的抗拉强度，从中随机抽取了76个样本进行试验

抗拉强度观测值											
41.0	37.0	33.0	44.2	30.5	27.0	45.0	28.5	31.2	33.5	38.5	41.5
43.0	45.5	42.5	39.0	38.8	35.5	32.5	29.6	32.6	34.5	37.5	39.5
42.8	45.1	42.8	45.8	39.8	37.2	33.8	31.2	29.0	35.2	37.8	41.2
43.8	48.0	43.6	41.8	36.6	34.8	31.0	32.0	33.5	37.4	40.8	44.7
40.2	41.3	38.8	34.1	31.8	34.6	38.3	41.3	30.0	35.2	37.5	40.5
38.1	37.3	37.1	41.5	29.5	29.1	27.5	34.8	36.5	44.2	40.0	44.5
40.6	36.2	35.8	31.5								

样本分布

- 对于观测数据经过以下四步处理：
  1. 整理数据，把样本值$x_1,x_2,\ldots,x_n$按从小到大排列

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

  2. 分组：把区间$[x_{(1)}, x_{(n)}]$分成若干个小区间$[x_{(1)}, t_1],(t_1, t_2],\ldots,(t_{l-1}, x_{(n)}]$,每个小区间的长度$d_i=t_i-t_{i-1}$称为组距，区间的中点称为组中值，一般采用等距分组，最好使每个区间内至少有一个观测值

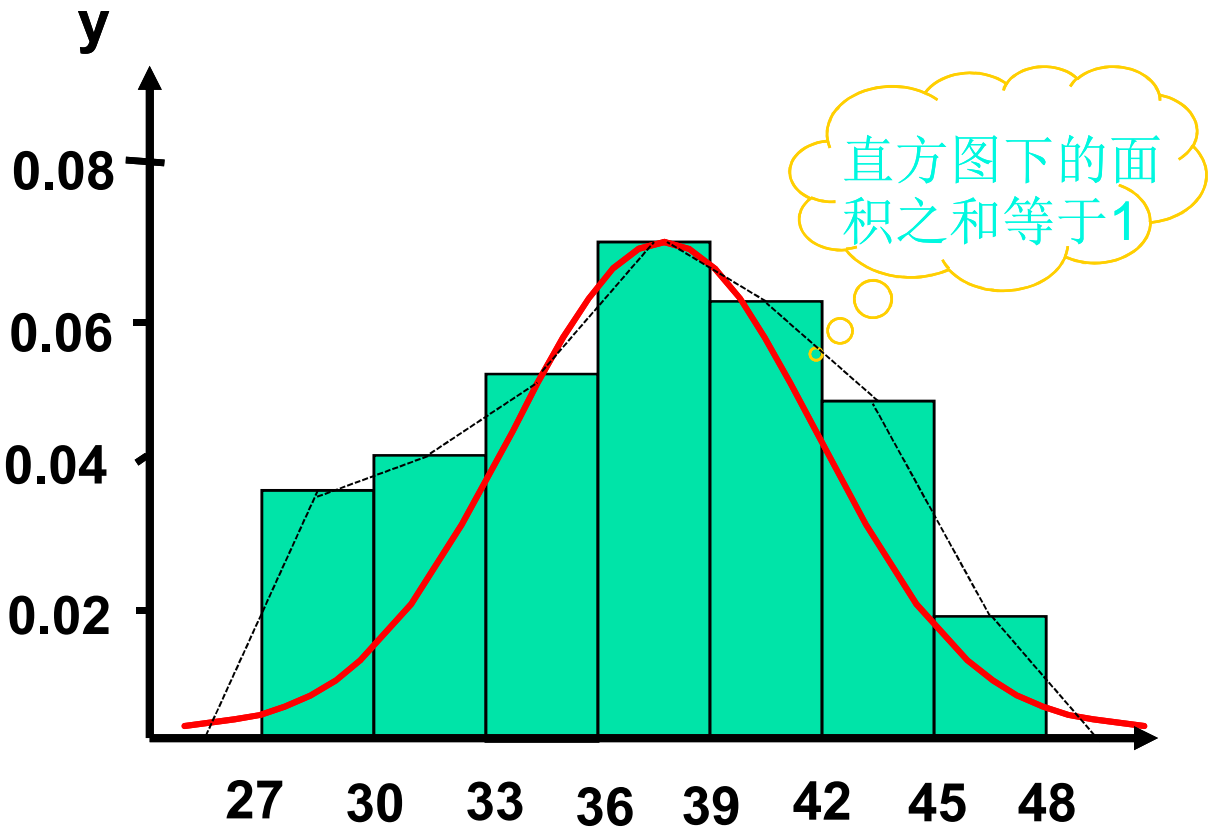
样本分布

3. 列分组频率分布表

分组	组中值	频数m_i	频率f_i	$y_i = f_i/d_i$
[27,30]	28.5	8	0.105	0.035
(30,33]	31.5	10	0.132	0.044
(33,36]	34.5	12	0.158	0.053
(36,39]	37.5	17	0.224	0.074
(39,42]	40.5	14	0.184	0.061
(42,45]	43.5	11	0.145	0.048
(45,48]	46.5	4	0.053	0.018

4. 作频率直方图：在x轴上以各区间$(t_{i-1}, t_i]$为底，以y_i为高画一排竖立的矩形，即为频率直方图，根据大数定律，当n很大时，$f_i \approx p_i = \int_{t_{i-1}}^{t_i} f(x)dx = f(\omega_i)d_i, \omega_i \in (t_{i-1}, t_i]$，于是$f_i/d_i$可以近似为密度函数

5. 作概率密度曲线：将各矩形中点联结得到一条折线，样本容量越大，分组越细，得到的概率密度曲线越准确



样本分布

- 对于样本值$(x_1,x_2,\ldots,x_n)$，将其从小到大排列为$x_{(1)}\le x_{(2)}\le \ldots\le x_{(n)}$，对于任意的$x(-\infty<x<+\infty)$，定义函数

$$\begin{cases} 0, & x < x_{(1)} \\ \dfrac{m_{1+}m_2 + \cdots + m_i}{n}, & x_{(i)} \le x < x_{(i+1)}, i = 1,2,\ldots,n-1 \\ 1, & x \ge x_{(n)} \end{cases}$$

- 称$F_n(x)$为总体X的经验分布函数
  - 单调、非降、右连续，$0 \le F_n(x) \le 1$
  - $F_n(-\infty) = 0$，$F_n(+\infty) = 1$

样本分布

- $F_n(x)$表示$x_1, x_2, \ldots, x_n$落入区间$[-\infty, x]$内的频率；
- 对于不同样本观察值，将得到不同的经验分布函数，经验分布函数不仅与样本容量有关，还与样本值有关，因为每次实验中的样本值是随机的；
- 当n很大时，$F_n(x)$可以作为$F(x)$的估计。

统计量

- 对于来自总体X的样本$(X_1, X_2, \ldots, X_n)$，若$T = g(X_1, X_2, \ldots, X_n)$不包含未知参数，则称T为一个统计量

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

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \qquad B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$$

统计量

- 定理　对于来自总体X的样本$(X_1, X_2, \ldots, X_n)$，且$E(X) = \mu$和$D(X) = \sigma^2$均存在，则

1. 样本均值$\bar{X}$的数学期望和方差分别为

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

2. 样本方差的数学期望为

$$E(S^2) = \sigma^2 \qquad S^2 \xrightarrow{P} \sigma^2$$

统计量

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

$$D(\bar{X}) = D\left(\frac{(X_1 + X_2 + \cdots + X_n)}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{\sigma^2}{n}$$

由大数定律得到$\bar{X} \xrightarrow{P} \mu$

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

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

统计量

- （3）由大数定律知

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} E(X^2) = (\sigma^2 + \mu^2)$$

由（1）知

$$\bar{X} \xrightarrow{p} \mu$$

于是有

$$S^2 = \frac{n}{n-1}\left[1/n\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right] \xrightarrow{p} \sigma^2 + \mu^2 - \mu^2$$

统计量

- 对于来自总体X的样本$(X_1,..X_n)$将其观测值按照从小到大顺序排列，得到

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

称$X_{(1)}, X_{(2)}, \ldots, X_{(n)}$为样本$(X_1,..X_n)$的次序统计量

- 统计量$X_{(1)} = \min(X_1,..X_n)$和$X_{(n)} = max(X_1,..X_n)$称为最小次序统计量和最大次序统计量

统计量

- 统计量

$$R = X_{(n)} - X_{(1)}$$

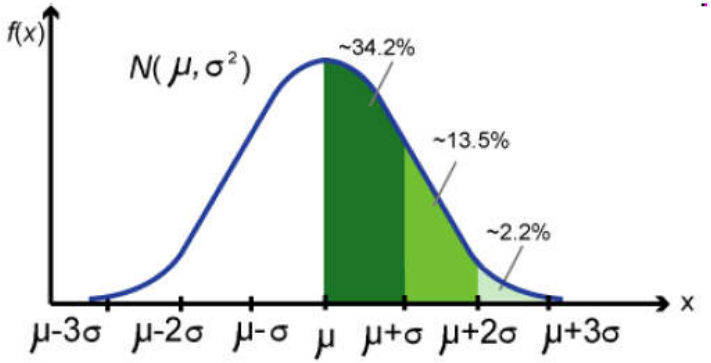
称为样本极差

- 进一步还可以定义四分位差和样本中位数

抽样分布

- 正态分布

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



68%的个案是落在离均值 1 个标准偏差(1σ)的范围内
95%的个案是落在离均值 2 个标准偏差(2σ)的范围内
99%的个案是落在离均值 3 个标准偏差(3σ)的范围内

抽样分布

- Γ函数

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

性质

$\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$ $\Gamma(n+1)=n!$ $\Gamma(1/2)=\sqrt{\pi}$

$\Gamma(1)=1$ $\Gamma(2)=1$ $\Gamma(3)=2$

抽样分布

- χ^2 分布

定义：设随机变量$X_1,\cdots,X_n$相互独立，都服从$N(0,1)$，

则称 $\chi^2=\sum_{i=1}^{n}X_i^2$

服从自由度为n的χ^2分布，记为$\chi^2\sim\chi^2(n)$

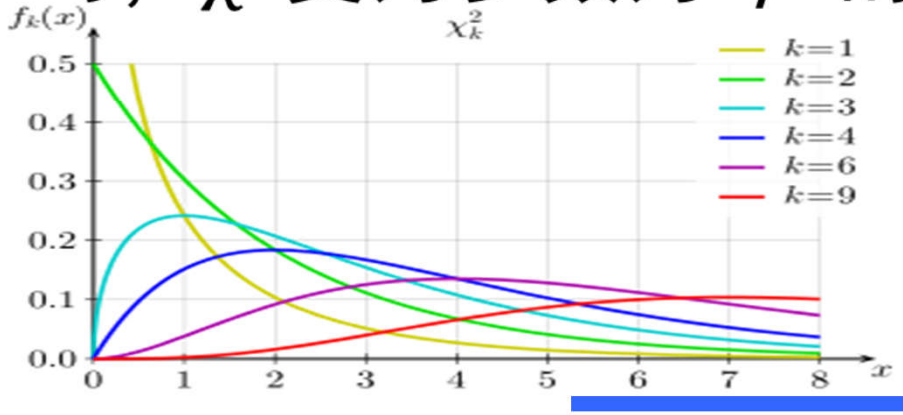
自由度指(1)式右端包含的独立变量的个数.

抽样分布

- $Z \sim \chi^2(n)$，卡方（Chi-square）分布的密度函数

$$\chi^2(x;n) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\Gamma(n/2)} x^{\frac{n}{2}-1} e^{-x/2}, x > 0 \\ 0, x \le 0 \end{cases}$$

- 当n=2时，χ^2变为参数为1/2的指数分布



抽样分布

- χ^2分布具有：

性质5.4.1　　$Z\sim\chi^2(n)$, 则 $E(Z)=n, D(Z)=2n$

证明1　　$$E(Z)=\int_0^{+\infty} x\chi^2(x;n)dx=\int_0^{+\infty}\frac{1}{2^{\frac{n}{2}}\Gamma(n/2)}x^{\frac{n}{2}}e^{-x/2}dx$$

令$\frac{x}{2}=t$, 有

$$E(Z)=\int_0^{+\infty}\frac{2}{\Gamma(n/2)}t^{\frac{n}{2}}e^{-t}dt=\frac{2\Gamma(\frac{n}{2}+1)}{\Gamma(n/2)}=2*\frac{n}{2}=n$$

抽样分布

性质5.4.1　　$Z\sim\chi^2(n)$, 则 $E(Z)=n, D(Z)=2n$

证明2　$$E(Z)=\mathrm{E}(\sum_{i=1}^{n}\mathrm{X}_i^2)=\sum_{i=1}^{n}\mathrm{E}(\mathrm{X}_i^2)=\sum_{i=1}^{n}(\mathrm{D}(\mathrm{X}_i)+\mathrm{E}^2(\mathrm{X}_i))$$

由$\mathrm{D}(\mathrm{X}_i)=1,\ \mathrm{E}(\mathrm{X}_i)=0,$ 得

$$E(Z)=\mathrm{n}$$

抽样分布

性质5.4.1　　$Z\sim\chi^2(n)$, 则 $E(Z)=n, D(Z)=2n$

证明1

$$E(Z^2)=\int_0^{+\infty}\frac{1}{2^{\frac{n}{2}}\Gamma(n/2)}x^{\frac{n}{2}+1}e^{-x/2}dx$$

令$\frac{x}{2}=t$, 有

$$E(Z^2)=\int_0^{+\infty}\frac{2*2}{\Gamma(n/2)}t^{\frac{n}{2}+1}e^{-t}dt$$

$$=\frac{4\Gamma(\frac{n}{2}+2)}{\Gamma(n/2)}=4*\left(\frac{n}{2}+1\right)*\frac{n}{2}=n(n+2)$$

$$D(Z)\ =E(Z^2)-E^2(Z)=2n$$

抽样分布

证明2

$$D(Z)=D(\sum_{i=1}^{n}X_i^2)=\sum_{i=1}^{n}D(X_i^2)=\sum_{i=1}^{n}(E(X_i^4)-E^2(X_i^2))$$

由上面的结论，知$E(X_i^2)=1$。

接下来计算$E(X_i^4)$

$$E(X_i^4)=\int_{-\infty}^{+\infty}x^4\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=2\int_0^{+\infty}x^4\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$$

令 $\frac{x}{\sqrt{2}}=t$, $dx=\sqrt{2}dt$

$$E(X_i^4)=2\int_0^{+\infty}x^4\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=8\int_0^{+\infty}t^4\cdot\frac{1}{\sqrt{\pi}}e^{-t^2}dt=\frac{4}{\sqrt{\pi}}\Gamma(\frac{5}{2})$$

伽马函数具有如下递归性质:

$$E(X_i^4)=\frac{4}{\sqrt{\pi}}\Gamma(\frac{5}{2})=\frac{4}{\sqrt{\pi}}\cdot\frac{3}{2}\cdot\frac{1}{2}\cdot\sqrt{\pi}=3$$

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

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

抽样分布

性质5.4.2　设$Z_1, Z_2, \dots Z_m$相互独立，且 $Z_i \sim \chi^2(n_i)$，

则$\sum_{i=1}^{m} Z_i \sim \chi^2(n_1 + \cdots + n_m)$

- 设$X_1, X_2, \dots X_n$为来自正态总体$N(\mu, \sigma^2)$的简单随机样本，问

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

服从什么分布？

抽样分布

t 分布

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

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

服从自由度为n的t分布.
(也称为学生氏分布)
记为 $T \sim t(n)$.

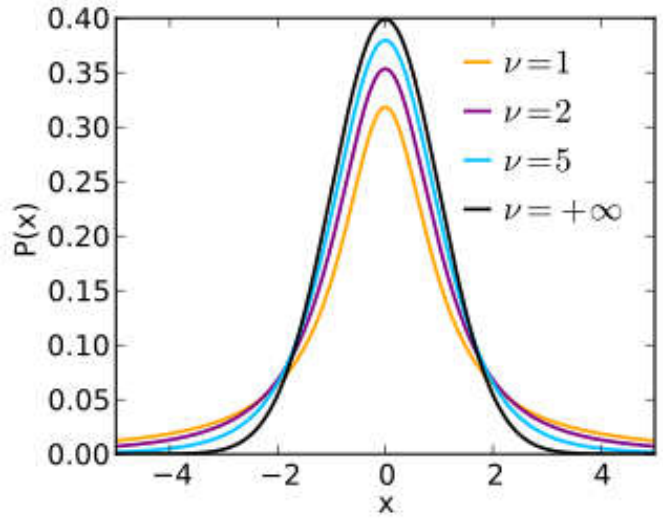
William Gosset
(1876–1937)
1908年提出t–分布

抽样分布

- t 分布具有概率密度

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

- t 分布曲线



$n=1$ 的 t 分布就是柯西分布

$$f(x;1)=\frac{1}{\pi\left(1+x^2\right)},-\infty<x<+\infty$$

当$n\to\infty, t(n)\to N(0,1)$

$$n\to\infty, f(x;n)\to\frac{1}{\sqrt{2\pi}}e^{-x^2/2},-\infty<x<+\infty$$

抽样分布

- 设随机变量$X\sim N(0,1)$, $Y\sim\chi^2(n)$且X与Y 独立，则

$$T=\frac{X}{\sqrt{Y/n}}\sim t(n)$$

证　令$Z=\sqrt{Y/n}$, 则Z的密度为

$$f(z)=\frac{n^{\frac{n}{2}}z^{n-1}e^{-\frac{nz^2}{2}}}{2^{\frac{n}{2}-1}\Gamma\left(\frac{n}{2}\right)},z>0$$

抽样分布

$Z=\frac{Y}{X}$，X Y 相互独立

⇩

$$f_Z(z)=\int_{-\infty}^{+\infty}|x|\,f_Y(xz)f_X(x)\mathrm{d}x$$

于是$T=\frac{X}{Z}$的密度为

$$g(t)=\int_{-\infty}^{+\infty}|z|\frac{e^{-\frac{(tz)^2}{2}}}{\sqrt{2\pi}}f(z)dz$$

$$=\frac{n^{\frac{n}{2}}}{\sqrt{\pi}2^{\frac{n-1}{2}}\Gamma\left(\frac{n}{2}\right)}\int_0^{+\infty}z^n e^{-\frac{(n+t^2)z^2}{2}}dz$$

令$\frac{(n+t^2)z^2}{2}=s$,有

$$g(t)=\frac{1}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}(1+\frac{t^2}{n})^{-(n+1)/2}\int_0^{+\infty}s^{(n-1)/2}e^{-s}ds$$

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

抽样分布

- 设$X_1, \ldots, X_n, X_{n+1}$是来自正态总体$N(\mu, \sigma^2)$的简单随机样本，问

$$Y = \frac{\sqrt{n}(X_{n+1} - \mu)}{\sqrt{\sum_{i=1}^{n}(X_i - \mu)^2}}$$

服从什么分布？

抽样分布

- F 分布

定义：设$X \sim \chi^2(n_1), Y \sim \chi^2(n_2)$，且$X, Y$独立，则称随机变量$F = \dfrac{X/n_1}{Y/n_2}$

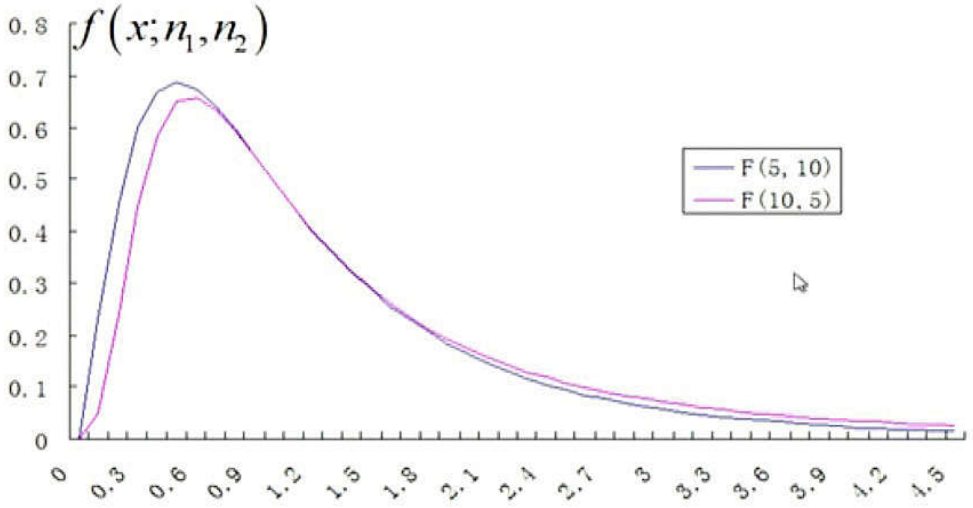
服从自由度为(n_1, n_2)的F分布，记为$F \sim F(n_1, n_2)$，

其中n_1称为第一自由度，n_2称为第二自由度.

抽样分布

- F 分布具有概率密度

$$f(x;n_1;n_2)=\frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)\left(\frac{n_1}{n_2}x\right)^{\frac{n_1}{2}-1}\left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}},x>0$$



$F\sim F(n_1,n_2)$,则$\frac{1}{F}\sim F(n_2,n_1)$

抽样分布

设$X_1, \ldots, X_{2n}$是来自正态总体$N(\mu, \sigma^2)$的简单随机样本，问

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

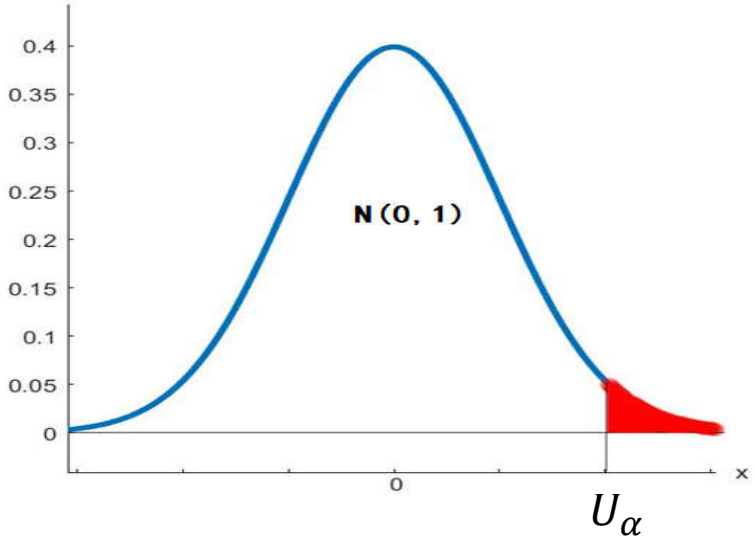
服从什么分布？

分位数

- 定义　　设随机变量X的分布函数为$F(x)=P(X\leq x)$,对于给定的$\alpha(0<\alpha\leq 1)$,若存在实数x_α使得

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

- 则称x_α为随机变量X的上侧α分位数



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

正态总体的抽样分布

- 定理　　设$X_1, \ldots, X_n$是来自正态总体$N(\mu, \sigma^2)$的样本，$\bar{X}$为样本均值，S^2为样本方差，则

1. $\bar{X} \sim N(\mu, \sigma^2/n)$
2. $\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2(n-1)$
3. $\bar{X}$与S^2独立

1. 证明：

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

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

$X_1, X_2, \ldots X_n$独立且都服从正态分布，

而且$\bar{X}$是$X_1, X_2, \ldots X_n$的线性组合

$\Rightarrow \bar{X}$服从正态分布，即$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

2. 思考

设总体$X \sim N(\mu, \sigma^2), X_1, X_2, \cdots, X_n$是样本，

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

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

2. 思考

设总体$X \sim N(\mu, \sigma^2), X_1, X_2, \cdots, X_n$是样本，

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

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

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

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

$X_1-\bar{X}, \cdots, X_n-\bar{X}$

有一个约束条件

$$\sum_{i=1}^{n}(X_i-\bar{X})=0$$

- 定理　　设$X_1,\dots,X_n$是来自正态总体$N(\mu,\sigma^2)$的样本，则

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

证　　$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}\sim N(0,1)$, $\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$　　且相互独立

于是　　$$T=\frac{\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}}=\frac{\sqrt{n}(\bar{X}-\mu)}{S}\sim t(n-1)$$

- 定理　　设$(X_1,\dots,X_{n_1})$和$(Y_1,\dots,Y_{n_2})$是分别来自正态总体$N(\mu_1,\sigma^2)$和$N(\mu_2,\sigma^2)$的样本，且两组样本独立，$\bar{X}=\frac{1}{n_1}\sum_{i=1}^{n_1}X_i$，$\bar{Y}=\frac{1}{n_2}\sum_{i=1}^{n_2}Y_i$，$S_{1n_1}^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\bar{X})^2$，$S_{2n_2}^2=\frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i-\bar{Y})^2$,则有

$$T=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)$$

其中

$$S_W^2=\frac{(n_1-1)S_{1n_1}^2+(n_2-1)S_{2n_2}^2}{n_1+n_2-2}$$

- 定理　　设$(X_1,\dots,X_{n_1})$和$(Y_1,\dots,Y_{n_2})$是分别来自正态总体$N(\mu_1,\sigma_1^2)$和$N(\mu_2,\sigma_2^2)$的样本，且两组样本独立，则

$$F=\frac{\sigma_2^2 S_{1n_1}^2}{\sigma_1^2 S_{2n_2}^2}\sim F(n_1-1,n_2-1)$$

或者

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

这是因为

$$\frac{\dfrac{(n_1-1)S_{1n_1}^2}{\sigma_1^2}/(n_1-1)}{\dfrac{(n_2-1)S_{2n_1}^2}{\sigma_2^2}/(n_2-1)}\longrightarrow\chi^2(n_1-1)\longrightarrow\chi^2(n_2-1)$$