Review of Mathematical Statistics

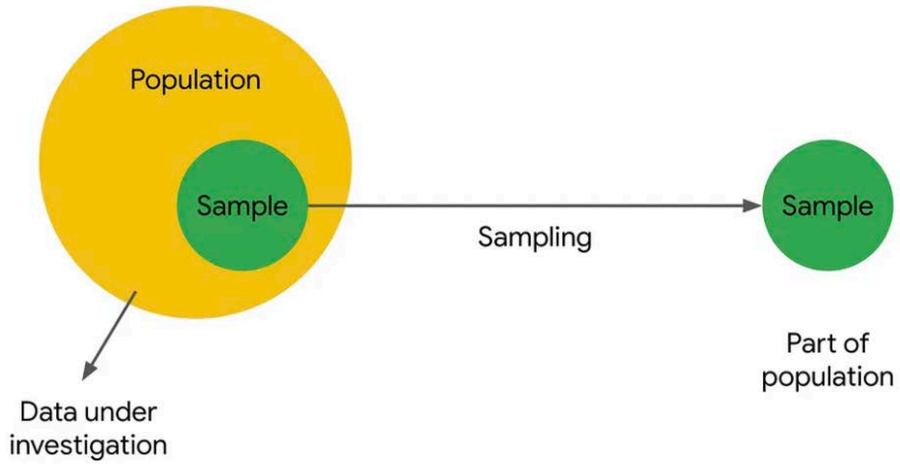
数理统计知识点回顾

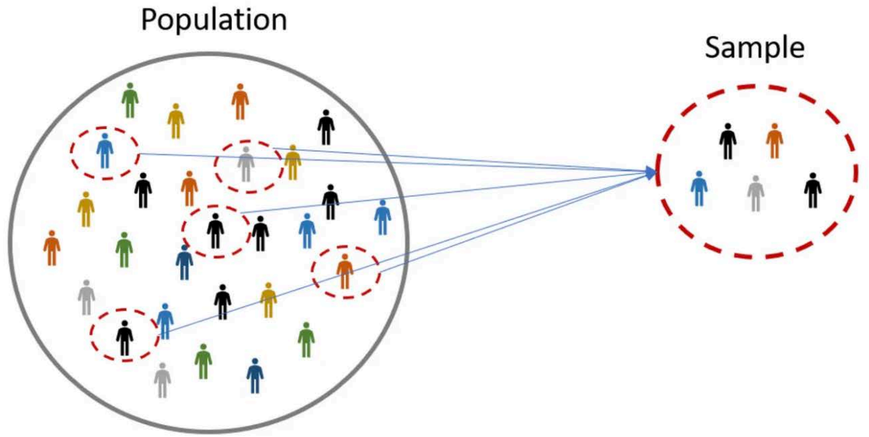
大纲

- 总体与样本
- 概率论基础
- 随机变量及其分布
- 点估计
- 置信区间
- 假设检验

总体与样本

- 总体：所有研究对象的全体
- 样本：所能观测到的总体的一部分
- 总体与样本的关系

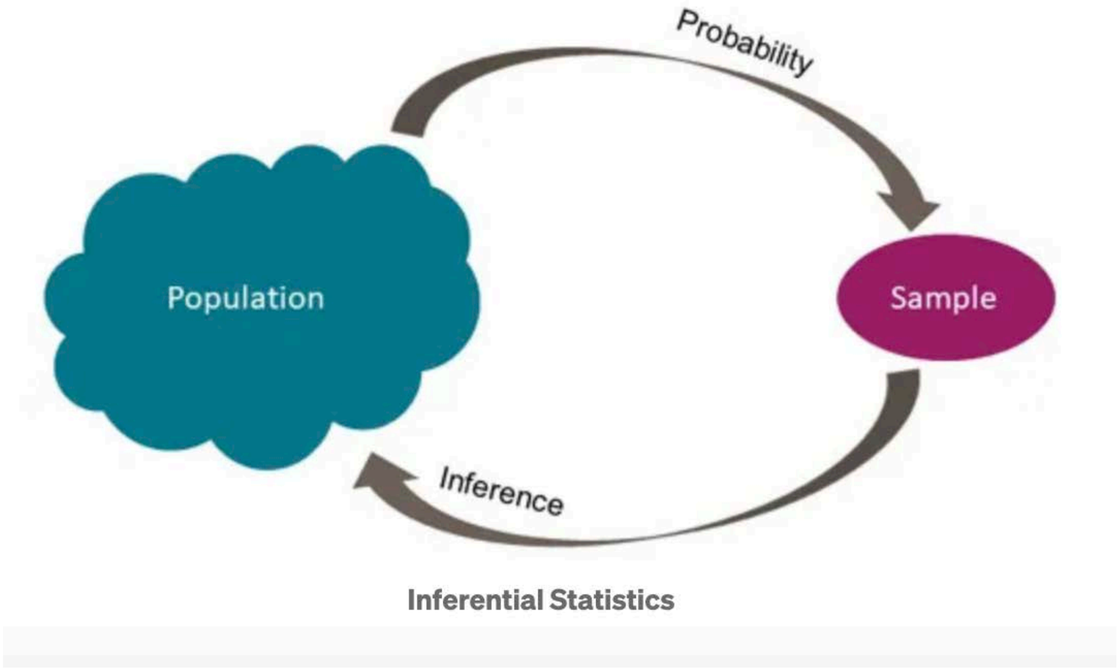




总体与样本

- 如果我们考虑抛掷一枚硬币1 万次，那么我们得到6 千次正面的概率是多少？

- 如果我们观测到在1 万次抛掷中共有6 千次是正面，那么这门硬币是一枚标准的硬币吗？



概率论基础

- 试验：任何能够产生观测的行为
- 样本空间：由试验所有可能结果构成的结合Ɛ
- 事件：样本空间Ɛ的子集
- 集合运算：交、并、补
- 事件A发生的概率

$$P(A) = \frac{\#A\text{中结果个数}}{\#\text{Ɛ中结果个数}}$$

- 乘法原理
- 排列
- 组合

概率运算

- $P(A^c) = P(A)$
- 如果A与B互斥

$$P(A \cap B) = p(\emptyset) = 0$$

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- 三组事件时

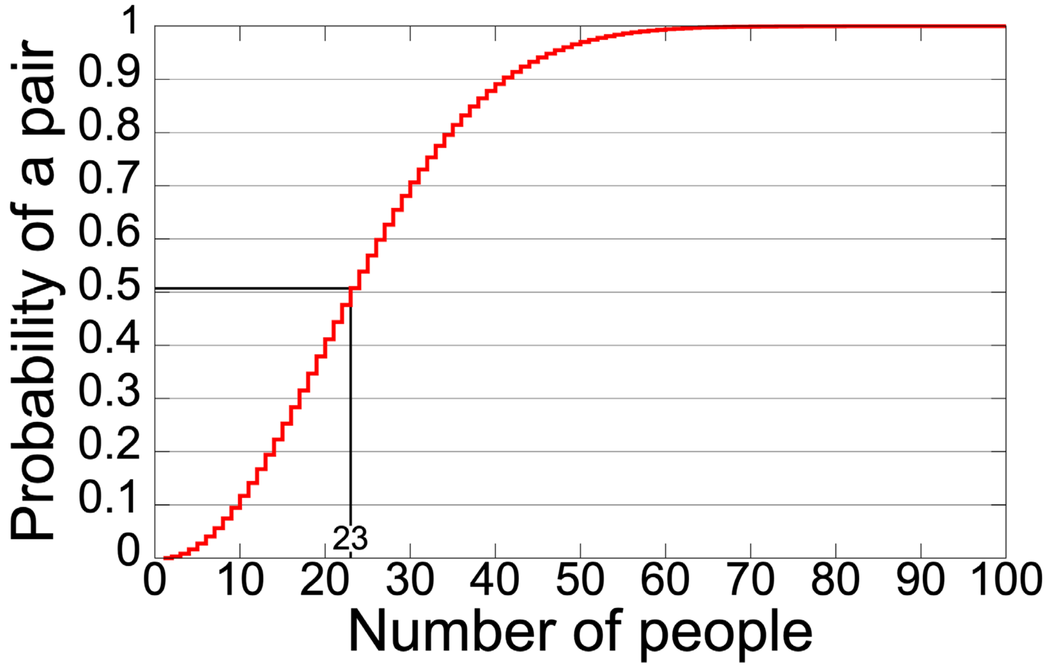
$$P(A \cup B \cup C)$$
$$= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

概率运算-生日悖论

- 例：一个有n 位同学的班级里，至少有两位同学有着相同生日的概率是多少？

解：令A表示至少有两位同学有相同身体，则

$$P(A) = 1 - P(A^c) = 1 - \frac{365 \times 364 \times \cdots \times (365 - n + 1)}{365^n}$$



条件概率

- 条件概率：给定B发生的情况下，A发生的概率为

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$$

- 如果$P(A \cap B) = P(A)P(B)$，则A与B是独立的
- 贝叶斯公式

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B|A\_i)}{\sum_k^K P(A_k)\,P(B|A_k)}$$

条件概率-吸毒者测试

- 例：假设一个毒品的检测结果的敏感度与可靠度均为99%，也就是说，当被检者吸毒时，每次检测呈阳性(positive, +) 的概率为99%。而被检者不吸毒时，每次检测呈阴性(negative, -) 的概率为99%。假设某地区吸毒人群比重为0.5% .如果该地区一个随机个体被测试出阳性，其真的是吸毒者的概率是多少？

解：令User表示个体为吸毒者，+表示测试为阳性的时间，则

$$
\begin{aligned}
P(\text{User} \mid +) &= \frac{P(+ \mid \text{User})P(\text{User})}{P(+)} \\
&= \frac{P(+ \mid \text{User})P(\text{User})}{P(+ \mid \text{User})P(\text{User}) + P(+ \mid \text{Non-user})P(\text{Non-user})} \\
&= \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995} \\
&\approx 33.2\%
\end{aligned}
$$

随机变量

- 随机变量：表示随机现象结果的变量
- 描述统计量（图形）
    - 直方图：频数，相对频数，密度
    - 饼图：比率
    - 箱线图：中位数，四分之一分位数，四分之三分位数，异常点
- 描述统计量（数值）
    - 位置：均值、中位数、切尾中位数
    - 变异性：方差，标准差

离散型随机变量

- 离散型随机变量：一个随机变量X取有限个或可列个值

  - 概率密度函数（pdf）：
    $$P(X = x) = P(e \in \mathcal{E}: X(e) = x)$$
  - 累积密度函数（cdf）：
    $$F(x) = P(X \leq x) = P(e \in \mathcal{E}: X(e) \leq x)$$
  - 期望：$\mathrm{E}(X) = \sum_x xP(X = x)$
  - 方差：$\mathrm{v}ar(X) = E(X - EX)^2 = E(X^2) - (EX)^2$

离散型随机变量

- 常见离散型随机变量

分布	数学标记	参数	分布律或概率密度	数学期望	方差
单点分布（退化分布）	$b_0(a,1)$	a	$P(x=a)=1$	a	0
(0-1)分布（两点分布或伯努利分布）	$b(1,p)$	$0<p<1$	$P\{X=k\}=p^k(1-p)^{1-k}, k=0,1$	p	1-p
二项分布	$B(n,p)$	$0<p<1$ $n\geq 1$	$P\{X=k\}=C_n^k p^k(1-p)^{n-k}$ K=0,1,2⋯	np	np(1-p)
负二项分布（帕斯卡分布）	$B_0(r,p)$	$0<p<1$ $r\geq 1$	$P\{X=k\}=C_{k-1}^{r-1}p^r(1-p)^{k-r}$ K=r,r+1,⋯	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$
几何分布	$G(p)$	$0<p<1$	$P\{X=k\}=(1-p)^{k-1}p$ K=1,2,⋯	$\frac{1}{p}$	$\frac{1-p}{p^2}$
超几何分布	$H(N,M,n)$	N,M,n (M≤N,n≤N)	$P\{X=k\}=\frac{C_M^k C_{N-M}^{n-k}}{C_N^k}$ $k\in Z, \max\{0,n-N+M\}\leq k\leq \min\{n,M\}$	$\frac{nM}{N}$	$\frac{nM}{N}\left(1-\frac{M}{N}\right)\left(\frac{N-n}{N-1}\right)$
泊松分布	$\pi(\lambda)$	$\lambda>0$	$P\{X=k\}=\frac{\lambda^k e^{-\lambda}}{k!}$ K=0,1,2,⋯	λ	λ

连续型随机变量

- 连续型随机变量：一个随机变量X取值充满数轴上的一个区间

  - 概率密度函数（pdf）：

$$f(x) = \lim_{h \to 0} \frac{1}{h} P(x \le X \le x + h)$$

  - 累积密度函数（cdf）：

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt$$

  - 期望：$\mathrm{E}(X) = \int_{-\infty}^{+\infty} xf(x)dx$
  - 方差：$var(X) = E(X - EX)^2 = E(X^2) - (EX)^2$

连续型随机变量

- 连续型随机变量：一个随机变量X取值充满数轴上的一个区间

均匀分布	U(a,b)	a < b	$f(x)=\begin{cases}\frac{1}{b-a}, a<x<b\\ 0, \quad 其它\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
正态分布（高斯分布）	$N(\mu,\sigma^2)$	μ $\sigma>0$	$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}$	μ	σ^2
对数正态分布	若X~N(μ,σ 且Y = e^X则 Y 服从该分布	μ $\sigma>0$	$f(x)=\begin{cases}\frac{1}{\sqrt{2\pi}\sigma x}e^{-(\ln x-\mu)^2/(2\sigma^2)}, x>0\\ 0\ ，其它\end{cases}$	$e^{\mu+\frac{\sigma^2}{2}}$	$e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$
逆高斯分布	$N^{-1}(\mu,\lambda)$	$\lambda,\mu>0$	$f(x)=\begin{cases}\sqrt{\frac{\lambda}{2\pi x^3}}e^{-\lambda(x-\mu)^2/(2\mu^2 x)}, x>0\\ 0\ ，其它\end{cases}$	μ	$\frac{\mu^3}{\lambda}$
Γ分布 [伽玛分布]	$\Gamma(\alpha,\beta)$	$\alpha,\beta>0$	$f(x)=\begin{cases}\frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta}, x>0\\ 0\quad ,\quad 其它\end{cases}$	$\alpha\beta$	$\alpha\beta^2$
指数分布（负指数分布）	$\Gamma(1,\theta)$ 注：指数分布是Γ分布的特殊情况	$\theta>0$	$f(x)=\begin{cases}\frac{1}{\theta}e^{-\frac{x}{\theta}}, x>0\\ 0, \quad 其它\end{cases}$	θ	θ^2
χ^2分布	$\chi^2(n)$	$n\geq 1$	$f(x)=\begin{cases}\frac{1}{2^{n/2}\Gamma(n/2)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}, x>0\\ 0, \quad 其它\end{cases}$	n	2n
非中心χ^2分布	$\chi^2(n,\lambda)$	$n\geq 1$ $\lambda>0$	$f(x)=\begin{cases}\frac{e^{-\left(\frac{x+\lambda}{2}\right)}}{2^{n/2}}\sum_{i=0}^{\infty}\frac{x^{\frac{n}{2}+i-1}\lambda^i}{\Gamma\left(\frac{n}{2}+i\right)2^{2i}i!}, (x>0)\\ 0\ ，其它\end{cases}$	$n+\lambda$	$2(n+2\lambda)$
韦布尔分布	$W(\eta,\beta)$	$\eta,\beta>0$	$f(x)=\begin{cases}\frac{\beta}{\eta}(\frac{x}{\eta})^{\beta-1}e^{-(\frac{x}{\eta})^\beta}, x>0\\ 0, \quad 其它\end{cases}$	$\eta\Gamma(\frac{1}{\beta}+1)$	$\eta^2\left\{\Gamma\left(\frac{2}{\beta}+1\right)-\left[\Gamma\left(\frac{1}{\beta}+1\right)\right]^2\right\}$
拉普拉斯分布		μ $\lambda>0$	$f(x)=\frac{1}{2\lambda}e^{-\frac{\lvert x-\mu\rvert}{\lambda}}$	μ	$2\lambda^2$

联合分布函数

- 随机变量(X,Y)的联合分布密度函数为$f(x,y)$
- 分布函数：

$$P\big((X,Y)\in A\big)=\int\int_A f(x,y)\,dxdy$$

- 若(X,Y)相互独立，则

$$f(x,y)=f_X(x)f_Y(y)$$

- 条件概率密度

$$f_{Y|X}(y|x)=\frac{f(x,y)}{f_{X(x)}}$$

- 期望

$$E(h(X,Y))=\int\int h(x,y)f(x,y)\,dxdy$$

- 协方差：$\mathrm{Cov}(X,Y)=E(XY)-EXEY$
- 相关系数：$\mathrm{Corr}(X,Y)=\frac{Cov(X,Y)}{\sqrt{var(X)var(Y)}}$

大数定律

- 一般形式：随机变量算术平均和依概率收敛于其期望

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \frac{1}{n}\sum_{i=1}^{n} EX_i$$

  - 切比雪夫大数定律：两两不相关+有限方差
  - 马尔科夫大数定律：$\frac{1}{n^2}\mathrm{var}(\sum_{i=1}^{n} X_i) \to 0 \; as \; n \to \infty$
  - 辛钦大数定律：独立同分布，具有有限数学期望

中心极限定理

- 一般形式：随机变量算术平均和标准化后依分布收敛于正态分布

$$\frac{\sum_{i=1}^{n} X_i - \sum_i EX_i}{\sqrt{\sum_i Var(X_i)}} \xrightarrow{d} N(0,1)$$

- 利用中心极限定理计算概率、分位数、样本量

  - 林德伯格-莱维中心极限定理：$X_1, \dots, X_n$独立同分布，$EX_i = \mu, VarX_i = \sigma^2$

$$\frac{\sum_{i=1}^{n} X_i - \sum_i EX_i}{\sqrt{\sum_i Var(X_i)}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} N(0,1)$$

点估计

- 简单随机抽样： $X_1,\dots,X_n$为独立同分布（i.i.d.）的（简单）随机样本
- 统计量：样本的函数$T(X_1,\dots,X_n)$，例如，样本均值、样本方差
- 参数θ点估计：基于给定样本所得到的的一个恰当的统计量
  - 若θ为总体均值，则样本均值、样本中位数均可视为θ的估计
  - 针对同一参数，往往有多种估计方法：如矩估计、极大似然估计
  - 无偏估计：$E(\hat{\theta}) = \theta$
  - 最小方差无偏估计（MVUE）：在所有无偏估计中方差最小的估计量

置信区间

- 置信区间：基于样本统计量所构建的一个有关参数的区间估计
- 正态分布均值区间估计：若$X_1, \ldots, X_n$是来自$N(\mu, \sigma^2)$的随机样本，关于μ的区间估计为

  - σ^2已知，

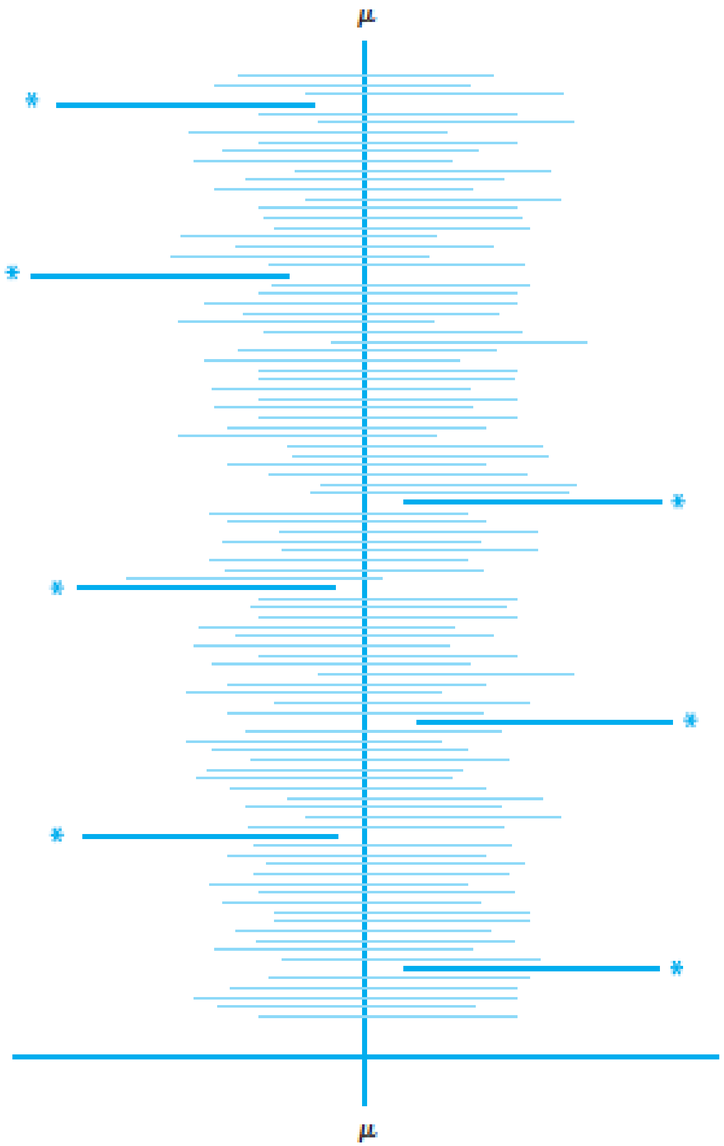
$$\left[\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma^2}{\sqrt{\mathrm{n}}}, \qquad \bar{X} + -Z_{1-\frac{\alpha}{2}}\frac{\sigma^2}{\sqrt{\mathrm{n}}}\right]$$

  - σ^2未知

$$\left[\bar{X} - t_{1-\frac{\alpha}{2}}\frac{\sigma^2}{\sqrt{\mathrm{n}}}, \qquad \bar{X} + -t_{1-\frac{\alpha}{2}}\frac{\sigma^2}{\sqrt{\mathrm{n}}}\right]$$

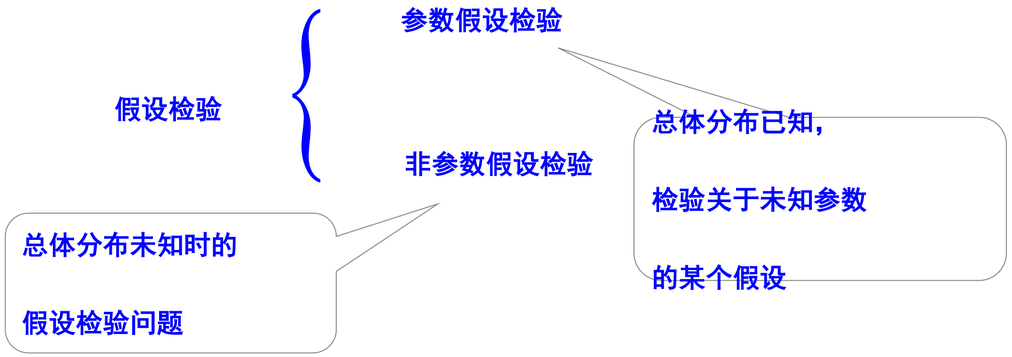
置信区间解释

- 置信区间可以被视为是一个随机的区间，即其端点是随机的
- 若从总体中随机抽取100组随机样本，并对每一组样本构造置信区间
- 在这构造的100个置信区间中，会有大约$100(1-\alpha)$个置信区间会包含μ



假设检验

- 在自然科学和社会科学等中，常常要对某些重要问题做出回答：是或否
- 如月球比地球早形成吗？一种新药对某种病有效吗？某种股票会涨吗？等等
- 为了回答这些问题，我们需要对感兴趣的问题进行试验或观察获得相关数据，根据这些数据决定是或否的过程称为假设检验
- 在总体X的分布完全未知，或只知其分布但不知其参数的情况下，我们对X的分布或分布中的参数作出某种假设，然后根据样本，用统计分析方法检验这一假设是否合理，从而作出接受或拒绝这一假设的决定.



假设检验步骤

- 建立假设 $H_0 \quad vs \quad H_1$;
- 选取检验统计量 $T(X_1, \dots, X_n)$，使得当$H0$成立时，T的分布完全已知，并确定拒绝域W的形状;
- 在给定的显著性水平α下，确定拒绝域 W;
- 根据样本观测值计算检验统计量$T(X_1, \dots, X_n)$，判断其是否属于拒绝域 W，做出最终判断。

表 检验的两类错误

观测数据情况	总体情况	
	H_0为真	H_1为真
$(x_1, \dots, x_n) \in W$	犯第一类错误	正确
$(x_1, \dots, x_n) \in \overline{W}$	正确	犯第二类错误

- 控制第一类错误概率（显著性水平），使犯第二类错误概率尽可能小（或功效尽可能大）

检验P值

- 在一个假设检验问题中，利用观测值能够做出拒绝原假设的最小显著性水平称为检验的$\boldsymbol{p}$ 值
- $\boldsymbol{p}$ 值是在原假设为真的情况下，所得到的的样本观察结果或更极端结果出现的概率
- 引进检验的$\boldsymbol{p}$ 值的概念有明显的好处

  - 第一，它比较客观，避免了事先确定显著性水平；
  - 其次，由检验的p 值与人们心目中的显著性水平α 进行比较可以很容易作出检验的结论：

    如果$\alpha \geq p$，则在显著性水平α 下拒绝 H_0；

    如果$\alpha < p$，则在显著性水平α 下接受 H_0.
  - p 值在应用中很方便，如今的统计软件中对检验问题一般都会给出检验的p 值。

单个正态总体均值检验

表 单个正态总体均值的假设检验

检验法	条件	原假设H_0	备择假设H_1	检验统计量	拒绝域	p值
U检验	σ已知	$\mu \le \mu_0$ $\mu \ge \mu_0$ $\mu = \mu_0$	$\mu > \mu_0$ $\mu < \mu_0$ $\mu \ne \mu_0$	$U = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$\{U \ge u_{1-\alpha}\}$ $\{U \le u_{\alpha}\}$ $\{\lvert U\rvert \ge u_{1-\alpha/2}\}$	$1-\Phi(u_0)$ $\Phi(u_0)$ $2(1-\Phi(\lvert u_0\rvert))$
T检验	σ未知	$\mu \le \mu_0$ $\mu \ge \mu_0$ $\mu = \mu_0$	$\mu > \mu_0$ $\mu < \mu_0$ $\mu \ne \mu_0$	$T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$	$\{T \ge t_{1-\alpha}(n-1)\}$ $\{T \le t_{\alpha}(n-1)\}$ $\{\lvert T\rvert \ge t_{1-\alpha/2}(n-1)\}$	$P(T \ge t_0)$ $p(T \le t_0)$ $P(\lvert T\rvert \ge \lvert t_0\rvert)$

单个正态总体均值检验

表　两个正态总体均值差的检验

检验法	条件	原假设 H_0	备择假设 H_1	检验统计量	拒绝域	p值
U检验	σ_1,σ_2 已知	$\mu_1 \le \mu_2$ $\mu_1 \ge \mu_2$ $\mu_1 = \mu_2$	$\mu_1 > \mu_2$ $\mu_1 < \mu_2$ $\mu_1 \ne \mu_2$	$U_1 = \dfrac{(\bar{X}-\bar{Y})}{\sqrt{\dfrac{\sigma_1^2}{m}+\dfrac{\sigma_2^2}{n}}}$	$\{U_1 \ge u_{1-\alpha}\}$ $\{U_1 \le u_{\alpha}\}$ $\{\lvert U_1\rvert \ge u_{1-\alpha/2}\}$	$1-\Phi(u_1)$ $\Phi(u_1)$ $2(1-\Phi(\lvert u_1\rvert))$
T检验	σ_1,σ_2 未知$\sigma_1=\sigma_2$	$\mu_1 \le \mu_2$ $\mu_1 \ge \mu_2$ $\mu_1 = \mu_2$	$\mu_1 > \mu_2$ $\mu_1 < \mu_2$ $\mu_1 \ne \mu_2$	$T_1 = \dfrac{(\bar{X}-\bar{Y})}{S_w\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}}$	$\{T \ge t_{1-\alpha}(m+n-2)\}$ $\{T \le t_{\alpha}(m+n-2)\}$ $\{\lvert T\rvert \ge t_{1-\alpha/2}(m+n-2)\}$	$P(T_1 \ge t_1)$ $p(T_1 \le t_1)$ $P(\lvert T_1\rvert \ge \lvert t_1\rvert)$

注：$S_w^2 = ((m-1)S_X^2+(n-1)S_Y^2)/(m+n-2)$